

The OECD Program to Validate the Rat Uterotrophic Bioassay. Phase 2: Coded Single-Dose Studies

Jun Kanno,¹ Lesley Onyon,² Shyamal Peddada,³ John Ashby,⁴ Elard Jacob,⁵ and William Owens⁶

¹National Institute of Health Sciences, Tokyo, Japan; ²Environmental Health and Safety Division, Organisation for Economic Co-operation and Development, Paris, France; ³National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA; ⁴Syngenta Central Toxicology Laboratory, Macclesfield, Cheshire, United Kingdom; ⁵BASF Aktiengesellschaft, Ludwigshafen, Germany; ⁶The Procter & Gamble Company, Cincinnati, Ohio, USA

The Organisation for Economic Co-operation and Development has completed phase 2 of an international validation program for the rodent uterotrophic bioassay. This portion of phase 2 assessed the reproducibility of the assay with a battery of positive and negative test substances. Positive agonists of the estrogen receptor included the potent reference estrogen 17 α -ethinyl estradiol (EE), and the weak estrogen agonists bisphenol A, genistein, methoxychlor, nonylphenol, and *o,p'*-DDT. The negative test substance or nonagonist was *n*-dibutylphthalate. The test substances were coded, and prescribed doses of each test substance were administered in 16 laboratories. Two versions of the uterotrophic assay, the intact immature and the adult ovariectomized female rat, were tested and compared using four standardized protocols covering both *sc* and *po* administration. Assay reproducibility was compared using *a*) EE doses identical to those used in phase 1 and in parallel dose–response studies, *b*) single doses of the weak agonists identical to one of five doses from the dose–response studies, and *c*) a single dose of the negative test substance. The results were reproducible and in agreement both within individual laboratories and across the participating laboratories for the same test substance and protocol. The few exceptions are examined in detail. The reproducibility was achieved despite a variety of different experimental conditions (e.g., variations in animal strain, diet, housing protocol, bedding, vehicle, animal age). In conclusion, both versions of the uterotrophic bioassay and all protocols appear robust, reproducible, and transferable across laboratories and able to detect weak estrogen agonists. These results will be submitted along with other data for independent peer review to provide support for the validation of the uterotrophic bioassay. **Key words:** endocrine disruption, estrogen, rat uterus, uterotrophic. *Environ Health Perspect* 111:1550–1558 (2003). doi:10.1289/ehp.5870 available via <http://dx.doi.org/> [Online 23 January 2003]

The Organisation for Economic Co-operation and Development (OECD) has undertaken the validation of the uterotrophic bioassay. The management of the validation program and the results of other portions of the validation program have been described in other reports (Kanno et al. 2001, 2003). A central objective of the OECD validation program is to establish the reliability of standardized protocols for the uterotrophic bioassay. A demonstration of reliability is based on the transferability of a protocol among laboratories, where the protocol results are reproducible among laboratories (ICCVAM 1997; OECD 1998). Two aspects of reliability require demonstration in a validation program: *a*) the assay's sensitivity, or ability to respond to and detect positive substances, and *b*) the assay's specificity, or absence of response to negative substances (ICCVAM 1997; OECD 1998). Additionally, sensitivity and specificity should be assessed over time and should include data gathered using coded or blinded test substances (ICCVAM 1997; OECD 1998).

The studies in this paper are intended to demonstrate the reliability of the uterotrophic bioassay, including its sensitivity and specificity with coded samples. The test substances

were a potent reference agonist, five weak estrogen agonists, and a negative test substance. Four protocols are included in the validation studies to address the two primary versions of the uterotrophic assay, the intact, immature, and the adult ovariectomized (OVX) female rat as well as the primary routes of administration, oral gavage and *sc* injection. A previous article demonstrated the reproducibility of the dose response of the reference agonist, 17 α -ethinyl estradiol (EE), with both versions and all protocols (Kanno et al. 2001). An accompanying article demonstrates the reproducibility of both versions and all protocols using dose responses of the five weak agonist test substances (Kanno et al. 2003). Because all laboratories performed the EE dose response separate from these data, and almost all laboratories performed the weak agonist dose–response and coded single-dose studies at separate times, a comparison of the data provides for an assessment of bioassay reproducibility over time.

Materials and Methods

Test substances. Test substances were obtained and distributed through a centralized chemical repository at TNO, Zeist, the Netherlands. This repository is described in

the accompanying paper, including a full description of the chemical identities, purities, and sources (Kanno et al. 2001), with the exception of the negative test substance, *n*-dibutylphthalate (DBP) (CAS no. 84-74-2, purity 99.9%) which was obtained from Sigma Aldrich (St. Louis, MO, USA). Because of the coded nature of this study, the amounts of test substance needed by each laboratory were calculated for each protocol. These amounts were preweighed into individually coded, opaque vials at the central repository prior to their shipment.

Animal supply, husbandry, and preparation. The details of how participating laboratories obtained animals, the housing and husbandry conditions, the age of the animals, compliance with the OECD guidelines on animal care (OECD 2000) and appropriate national regulations, and the animal preparation and observation prior to test substance administration have been described previously (Kanno et al. 2001, 2003).

Protocols. The details of the individual protocols have also been described previously (Kanno et al. 2001, 2003). Briefly, protocol A uses intact, immature female rats with dosing by oral gavage for 3 consecutive days. Protocol B uses intact, immature female rats with dosing by *sc* injection for 3 consecutive days. Protocol C uses young adult OVX rats as described above with dosing by *sc* injection for 3 consecutive days. Protocol D [previously called protocol C' (Kanno et al. 2001)] also uses young adult OVX rats and extends the *sc* injection dosing to a total of 7 days.

This article is part of the mini-monograph "The OECD Validation of the Uterotrophic Bioassay."

Address correspondence to J.W. Owens, Central Product Safety, The Procter & Gamble Company, 11810 East Miami River Rd., Cincinnati, OH 45252 USA. Telephone: (513) 627-1385. Fax: (513) 627-1208. E-mail: owens.jw@pg.com

Address any correspondence or questions about the OECD Programme and OECD documents to Herman Koeter, OECD Environmental Health and Safety Division, OECD, 2 rue André Pascal, 75775 Paris Cedex 16 France. Telephone: 33 1 45 24 9849. Fax: 33 1 45 24 1675. E-mail: herman.koeter@oecd.org

This paper represents the opinions of the authors and may not reflect the official positions and recommendations of the OECD.

The authors declare they have no conflict of interest. Received 1 July 2002; accepted 7 October 2002.

Coded samples, vehicle, test substance preparation, and dosing. For each test substance, individualized instructions, depending on the amount to be shipped, were given to each laboratory. The instructions specifically stated the volume of test vehicle to be added to the coded vials to provide a reference dose solution for each test substance. Further instructions were provided to adjust the administered test volume based on the recorded body weight (bw) of the animals to provide the prescribed experimental doses.

Participating laboratories were asked to have one set of personnel prepare the test substance dose solutions and administer the preparations and a second set perform the necropsy and record the uterine weights. This was intended to minimize the chances of working out the code for each test substance. Material safety data sheets were provided in a sealed envelope to a nominated person at each laboratory, who agreed to keep this envelope sealed except in cases of emergency. A generic material safety data sheet was prepared and supplied to cover all test substances so that the health and safety of personnel at the laboratory would not be compromised. The other details of the vehicle, test substance preparation, and animal-dosing procedures have been previously described (Kanno et al. 2001, 2003).

Necropsy, dissection, and uterine weight. As described previously, the animals were killed humanely 24 hr after the last test substance administration in the same sequence as the test substance was administered. The dissection of the uterus and the measurement of wet and blotted uterine weights to the nearest 0.1 mg were performed as described previously (Kanno et al. 2001, 2003).

Study management and quality control. The study management and quality control have been previously described. The VMG requested that the studies be performed under OECD Good Laboratory Practice (GLP) guidelines (OECD 2002). However, full GLP compliance was not a requirement for a laboratory's participation in the validation program, and several of the laboratories did not perform their studies under GLP. Data were received, and after an initial statistical analysis was performed, all laboratories were requested to audit these raw data and respond to specific queries on outliers and questionable data. A small number of data corrections were made, and reporting errors on dilutions, samples, and identity of control groups were either corrected or clarified.

Statistics. The recording and statistical procedures, data evaluation by an analysis of covariance, logarithmic transformation of uterine data, use of the Dunnett and Hsu pairwise comparison test, studentized residual plots, and use of the ratio of the geometric

means of the uterine weights (relative to the vehicle control) after adjusting for the body weight of the animal at necropsy with upper and lower 95% confidence levels have all been previously described (Kanno et al. 2003).

To draw inferences across laboratories about the reproducibility of results at a given dose for each protocol, mixed-effects linear models were used, where the laboratories were treated as the random effects. Such an analysis takes into consideration both between-lab variability and within-lab variability, and provides an overall summary of the results. Thus, the analysis enables the computation of a mean response to a chemical across labs, and the lower and the upper 95% confidence limits under each protocol. This use of mixed-effects linear models is termed the "global analysis."

Design of Phase 2 Coded Single-Dose Studies

The objective of the coded single-dose studies was to produce the data to assess the reproducibility of the uterotrophic bioassay both within the same laboratory and across the multiple, participating laboratories. Further, the reproducibility was to be assessed over time and using blinded or coded samples.

Overall design rationale. Three types of test substances were included: a potent reference test substance, EE; five weak estrogen receptor agonists: genistein (GN), methoxychlor (MX), nonylphenol (NP), bisphenol A (BPA), and 1,1,1-trichloro-2,2-bis(*o,p'*-chlorophenyl)ethane or *o,p'*-DDT (DDT); and a negative test substance, DBP. A robust statistical comparison required that identical doses be selected so that the same prescribed doses for each test substance were used in every laboratory.

Two EE doses were selected from phase 1 (Kanno et al. 2001) to generate two additional sets of data to assess the reproducibility of the bioassay. The first EE dose for a given route of administration was the first minimally effective dose in the lower portion of the dose-response curve that was a statistically significant response in all laboratories in phase 1. The second EE dose was then 0.5-log higher than the first dose, and this second dose had given responses near or at the maximum uterine response in phase 1. These selected doses were used as control reference doses as part of the accompanying dose-response studies (Kanno et al. 2003) and in these studies as coded samples. This design produces three data sets of replicate doses to assess the reproducibility of the uterine response over time.

The selection of the positive weak agonists and a series of five prescribed doses for each are described in the accompanying paper (Kanno et al. 2003). The participating laboratories were required in the dose-response studies to use the three intermediate doses,

whereas the lowest and highest of the five doses were optional. Therefore, the third or fourth dose in the series was selected for this coded single-dose study. As a result, two sets of replicate doses would be available, one from the dose-response studies and one from the coded single-dose studies, and would include all five weak agonists in all four standardized protocols.

The negative test substance, DBP, was chosen based on two lines of evidence. First, DBP does not display binding affinity for the rat uterine estrogen receptor, i.e., there is no displacement of bound [³H]17 β -estradiol at concentrations up to 1 mM concentrations *in vitro* (Blair et al. 2000). Second, *in vivo* toxicological studies, with some including gene activation profiles, indicate that DBP does not elicit responses indicative of an estrogen mode of action (Ema et al. 2000; Ema and Miyawaki 2001; Mylchreest et al. 1998, 1999; Schulz et al. 2001; Zacharewski et al. 1998). A single data set that included data for all four standardized protocols was judged adequate for the negative chemical to conserve resources and animals.

Selected doses. Two reference EE doses selected were 1 and 3 μ g/kg bw/day for oral gavage and 0.3 and 1 μ g/kg bw/day for sc injection. For the weak estrogen receptor agonists, the selected doses for the oral gavage studies were 600 mg BPA/kg bw/day, 300 mg GN/kg bw/day; 300 mg MX/kg bw/day; 250 mg NP/kg bw/day; and 300 mg DDT/kg bw/day. Doses for the sc injection studies were 300 mg BPA/kg bw/day; 35 mg GN/kg bw/day; 500 mg MX/kg bw/day; 80 mg NP/kg bw/day; and 100 mg DDT/kg bw/day. For the negative test substance, DBP, a limit dose was selected for each route of administration: 1,000 mg/kg bw/day for oral gavage and 500 mg/kg bw/day for sc injection.

Results

The coded single-dose studies were performed by 16 laboratories. Laboratories 6, 7, 9, 10, and 15, which either participated in phase 1 (Kanno et al. 2001) or the dose-response studies in phase 2 (Kanno et al. 2003), did not participate in the coded single-dose studies. However, their EE results from these studies were included in the comparison of the EE results generated in the coded single-dose studies. Despite the size of this international study, the actual difficulties encountered were few. For example, laboratories 17 and 19 may lack results for MX, BPA, GN, or DDT, because some of these substances were not administered after these two laboratories experienced difficulty in solubilization during dosage preparation. A few laboratories misinterpreted the EE dilution instructions, so that a few dose concentrations were either reversed or were incorrect (e.g., the high EE dose in

laboratory 12). Except for laboratory 1, audits of the records were able to correct the data for the reversals. Finally, uterine wall punctures were reported in three animals in separate laboratories and groups during dissection. The possible losses of imbibed fluid did not affect any results.

Mortalities, decreases in body weight or body weight gain, and clinical signs. Of 1,842 animals administered test substances in the coded single-dose studies, 42 mortalities were observed in eight laboratories. All mortalities in the coded single-dose studies were in protocol A (2 in GN studies, 3 in MC studies, 3 in DBP studies, 6 in BPA studies, 8 in DDT studies, and 19 in NP studies). As with the dose–response experiments, a dose-related pattern of modest reductions in body weights and diminished body weight gains was often observed in the immature animal studies and in the OVX studies where the dosing was extended to 7 days. Decreases in body weights at terminal sacrifice approaching or greater than 10% were observed with NP in most protocol A studies, DDT in protocol A, BPA in protocol D, MX in protocol D, and the high EE dose in some protocol D studies (data not shown), indicating that a maximum tolerated dose had been exceeded. Clinical signs were reported in conjunction with the mortalities and body weight losses, including piloerection, crouched positions, and labored breathing.

Ethinyl estradiol studies. Within each protocol, the mean increases in the body weight–adjusted blotted uterine weights of both the low and high EE doses were reproducible. The low and high EE dose results for the dose–response and coded single-dose protocols are shown in Table 1 and Table 2, respectively, and the phase 1 results have been previously reported (Kanno et al. 2001). In protocol A, the results of the three sets of EE data were reproducible. The blotted uterine weight increases were statistically significant at both EE doses, and the weights increased in a dose-related manner. There were two exceptions. Laboratory 1 did not achieve statistical significance at the lower 1 µg EE/kg/day dose in the dose–response studies, but had achieved statistical significance at this dose in phase 1. In laboratory 13, the ratio of mean uterine weight of the test substance group relative to the vehicle control group was nearly five at the lower EE dose in the coded single-dose studies. The ratio was a more modest value of 1.5 to 2 in phase 1 and the dose–response studies. In protocol B, the results of the three sets of EE data were reproducible with two exceptions. First, the ratio of the uterine weight increases in laboratories 9, 15, 18 at the lower EE dose was 3.5 to 5 in phase 2, compared with approximately 2 in phase 1. Second, laboratory 19 failed to achieve statistical significance at either EE dose. In protocol C, the results of the three sets of EE data were reproducible with one exception. Laboratory 19 achieved statistical

significance with the low EE dose, but not the high EE dose, in the coded single-dose studies. This same laboratory had shown a low responsiveness to EE in protocol C in phase 1 (Table 5 and Figure 1C in Kanno et al. 2001). In protocol D, the results of the three sets of EE data were reproducible. As noted in phase 1 (Kanno et al. 2001), the extended dosing in protocol D again typically led to a further increase in the blotted uterine weights over protocol C at both the low and high EE doses, but the increased number of EE doses also led to decreases in body weight gains.

Weak agonist studies. The results for the same BPA dose in the dose–response and coded single-dose studies are shown in Table 3. In protocol A, even at a dose of 600 mg BPA/kg/day, the relative uterine response was very weak and did not exceed a value of 2 in any laboratory. In the response distribution from this modest response, five laboratories failed to achieve statistical significance. Although all five had increased absolute uterine weights, the 95% lower confidence level did not exceed 1 as necessary for statistical significance. In three of these laboratories, animal mortalities occurred, decreasing the power. In protocol B, at a dose of 300 mg BPA/kg/day, the mean ratio values of the relative increase in uterine weight were between 1.5 and 2.8. In this response distribution, 3 of 23 experiments did not achieve statistical significance. In laboratory 12, the mean uterine weight was increased over 30%, but did not

Table 1. Relative increase in uterine weights versus vehicle controls with replicate low EE doses.

Lab	Protocol							
	A 1 µg/kg/day		B 0.3 µg/kg/day		C 0.3 µg/kg/day		D 0.3 µg/kg/day	
	Coded studies ^a	Dose response ^b	Coded studies	Dose response	Coded studies	Dose response	Coded studies	Dose response
1	—	1.10 (0.92, 1.32) ^{c,d}	—	2.17 (1.80, 2.62)*	—	1.98 (1.70, 2.30)*	—	2.93 (2.38, 3.60)*
	—	0.99 (0.77, 1.26) ^{a,d}	—	2.49 (2.12, 2.93)*	—	2.14 (1.88, 2.43)*	—	2.75 (2.30, 3.30)*
2	2.78 (2.19, 3.53)*	3.17 (2.52, 3.99)*	2.25 (1.80, 2.82)*	2.11 (1.70, 2.62)*	2.45 (1.87, 3.19)*	2.41 (2.06, 2.81)*	3.31 (2.74, 4.01)*	3.71 (2.91, 4.74)*
3	1.52 (1.24, 1.88)*	2.25 (1.77, 2.86)*	2.42 (1.96, 3.00)*	2.00 (1.68, 2.38)*	2.35 (1.84, 3.02)*	2.43 (2.14, 2.77)*	3.05 (2.40, 3.88)*	2.77 (2.44, 3.14)*
	—	1.53 (1.20, 1.96)*	—	2.31 (1.89, 2.81)*	—	2.96 (2.36, 3.71)*	—	2.85 (2.50, 3.26)*
4	2.81 (2.18, 3.61)*	3.20 (2.36, 4.35)*	1.79 (1.34, 2.39)*	2.75 (1.86, 4.06)*	—	—	—	—
5	1.36 (1.07, 1.74)*	1.40 (1.04, 1.90)*	2.07 (1.68, 2.55)*	See note ^e	—	—	—	—
6	—	—	—	1.59 (1.15, 2.18)*	—	2.43 (1.55, 3.82)*	—	—
7	—	2.16 (1.73, 2.69) ^f	—	1.73 (1.48, 2.01)*	—	1.78 (1.47, 2.16)*	—	2.45 (1.97, 3.05)*
	—	3.15 (2.57, 3.85) ^f	—	1.79 (1.52, 2.10)*	—	2.71 (2.27, 3.24)*	—	3.28 (2.48, 4.35)*
8	3.31 (2.73, 4.00)*	3.09 (2.55, 3.73)*	2.99 (2.40, 3.72)*	2.65 (2.37, 2.97)*	2.72 (2.03, 3.63)*	2.16 (1.91, 2.43)*	—	—
9	—	2.19 (1.72, 2.79)*	—	4.22 (3.63, 4.91)*	—	—	—	4.14 (3.34, 5.13)*
11	2.88 (2.26, 3.68)*	3.04 (2.42, 3.83)*	3.24 (2.48, 4.24)*	3.50 (2.83, 4.34)*	2.92 (2.47, 3.46)*	3.04 (2.63, 3.51)*	4.82 (3.61, 6.42)*	5.16 (3.65, 7.28)*
12	2.98 (1.47, 6.03)*	2.85 (2.21, 3.67)*	1.32 (0.91, 1.90) ^d	1.68 (1.11, 2.53)*	2.00 (1.44, 2.77)*	1.95 (1.52, 2.49)*	—	—
13	4.74 (3.88, 5.80)*	1.44 (1.06, 1.95)*	5.04 (3.67, 6.90)*	1.61 (1.08, 2.42)	—	—	—	—
14	2.44 (1.61, 3.70)*	3.11 (2.44, 3.98)*	2.69 (1.97, 3.68)*	2.61 (2.05, 3.32)*	—	—	—	—
15	—	—	—	4.45 (3.46, 5.71)*	—	—	—	—
16	—	—	1.58 (1.02, 2.46)*	—	—	—	—	—
17	—	—	1.83 (1.29, 2.61)*	—	—	—	—	—
18	—	—	3.51 (2.86, 4.32)*	3.81 (3.23, 4.50)*	—	—	—	—
19	—	—	0.95 (0.60, 1.50) ^d	—	1.97 (1.62, 2.40)*	—	—	—
20	—	—	1.78 (1.36, 2.33)*	1.83 (1.45, 2.31)*	—	—	—	—
	9/9	13/15	12/14	18/18	6/6	11/11	3/3	9/9

—, laboratory did not perform this particular study. ^aResults from studies with coded or blinded doses for each substance. ^bResults from the dose–response studies reported in the accompanying paper (Kanno et al. 2003). ^cRatio of geometric means of treated blotted uterine weights to the vehicle control blotted uterine weights after adjusting for the body weights at necropsy as a covariable (lower 95% confidence limit, upper 95% confidence limit). ^dThis study did not achieve statistical significance. ^eThis laboratory used po dilution instructions to use doses of 1 and 3 µg/kg/day. Therefore, no 0.3-µg/kg/day dose was available. ^fThis laboratory used sc dilution instructions to use doses of 0.3 and 1 µg/kg/day. The 1-µg/kg/day dose was the actual low EE dose and is reported here. *Level of significance, $p < 0.05$.

achieve significance. In laboratory 20, little or no evidence of a response was seen in either the dose–response or the coded single-dose studies. In protocol C, the ratio of the mean treated uterine weight relative to the vehicle controls was 2.3 to 3.4, and all laboratories in this response distribution were able to achieve statistical significance. This ratio value for the

adult OVX animals was consistently greater than for the immature animals in protocol B. In protocol D, the mean blotted uterine weights appeared to be increased by the extended dosing period, and all six laboratories achieved statistical significance.

The results for the same GN dose in the dose–response and coded single-dose studies

are shown in Table 4. The mean uterine responses at the selected GN doses relative to controls were 2 or greater for most laboratories. All laboratories in their respective response distributions achieved statistical significance in each protocol. In the case of GN, the immature animals in protocol B appeared to have a somewhat higher mean response than the

Table 2. Relative increase in uterine weights versus vehicle controls with replicate high EE doses.

Lab	Protocol							
	A 3 µg/kg/day		B 1 µg/kg/day		C 1 µg/kg/day		D 1 µg/kg/day	
	Coded studies	Dose response	Coded studies	Dose response	Coded studies	Dose response	Coded studies	Dose response
1	—	1.50 (1.25, 1.79)* ^a	—	4.19 (3.47, 5.05)*	—	3.13 (2.66, 3.67)*	—	4.40 (3.45, 5.61)*
	—	1.64 (1.29, 2.10)*	—	4.07 (3.46, 4.80)*	—	3.77 (3.27, 4.35)*	—	4.11 (3.36, 5.04)*
2	4.41 (3.47, 5.60)*	4.13 (3.27, 5.22)*	4.42 (3.52, 5.54)*	4.44 (3.60, 5.48)*	3.68 (2.77, 4.88)*	3.19 (2.69, 3.78)*	4.74 (3.88, 5.81)*	4.86 (3.55, 6.64)*
3	2.79 (2.27, 3.44)*	2.55 (2.00, 3.25)*	4.63 (3.74, 5.73)*	3.82 (3.17, 4.60)*	3.54 (2.74, 4.57)*	3.57 (3.06, 4.18)*	4.41 (3.32, 5.87)*	3.67 (3.13, 4.29)*
	—	2.80 (2.20, 3.58)*	—	3.79 (3.11, 4.63)*	—	3.61 (2.84, 4.59)*	—	4.04 (3.49, 4.69)*
4	3.70 (2.88, 4.76)*	4.04 (2.97, 5.48)*	3.41 (2.55, 4.56)*	4.52 (3.06, 6.67)*	—	—	—	—
5	1.80 (1.40, 2.31)*	1.91 (1.41, 2.59)*	4.15 (3.37, 5.11)*	3.61 (2.91, 4.46)* ^b	—	—	—	—
6	—	—	—	2.30 (1.71, 3.10)*	—	3.89 (2.45, 6.17)*	—	—
7	—	See note ^c	—	4.06 (3.49, 4.72)*	—	3.29 (2.69, 4.01)*	—	4.50 (3.53, 5.73)*
	—	—	—	4.16 (3.53, 4.90)*	—	4.32 (3.55, 5.25)*	—	5.67 (4.15, 7.74)*
8	5.02 (4.15, 6.08)*	4.69 (3.88, 5.66)*	4.76 (3.81, 5.95)*	4.96 (4.43, 5.55)*	3.31 (2.47, 4.42)*	2.70 (2.39, 3.05)*	—	—
9	—	5.19 (4.10, 6.58)	—	4.26 (3.64, 4.98)*	—	—	—	4.68 (3.66, 5.99)*
11	4.29 (3.36, 5.48)*	4.52 (3.54, 5.78)*	4.26 (3.17, 5.71)*	4.58 (3.70, 5.69)*	3.92 (3.28, 4.68)*	3.97 (3.36, 4.69)*	5.47 (4.14, 7.21)*	5.85 (4.26, 8.05)*
12	See note ^d	4.68 (3.63, 6.02)*	See note ^d	3.64 (2.43, 5.45)*	See note ^d	3.08 (2.41, 3.94)*	—	—
13	4.66 (3.83, 5.66)*	2.55 (2.05, 3.17)*	5.21 (3.81, 7.12)*	3.44 (2.25, 5.27)	—	—	—	—
14	4.20 (2.73, 6.47)*	4.69 (3.74, 5.89)*	4.83 (3.52, 6.63)*	4.55 (3.59, 5.76)*	—	—	—	—
15	—	—	—	4.95 (3.66, 6.69)*	—	—	—	—
16	—	—	3.29 (2.13, 5.09)*	—	—	—	—	—
17	—	—	3.50 (2.45, 5.00)*	—	—	—	—	—
18	—	—	5.12 (4.18, 6.29)*	5.62 (4.89, 6.47)*	—	—	—	—
19	—	—	0.80 (0.50, 1.27) ^e	—	0.96 (0.79, 1.16) ^e	—	—	—
20	—	—	2.41 (1.87, 3.11)*	2.38 (1.90, 2.99)*	—	—	—	—
	8/8	13/13	12/13	19/19	4/5	11/11	3/3	9/9

—, laboratory did not perform this particular study. ^aRatio of geometric means of treated blotted uterine weights to the vehicle control blotted uterine weights after adjusting for the body weights at necropsy as a covariable (lower 95% confidence limit, upper 95% confidence limit). ^bThis laboratory used po dilution instructions to use doses of 1 and 3 µg/kg/day. The 1 µg/kg/day dose was the actual high EE dose and is reported here. ^cThis laboratory used sc dilution instructions for doses of 0.3 and 1 µg/kg/day. Therefore, no 3 µg/kg/day high EE dose was performed. ^dThis laboratory incorrectly diluted the high EE dose in all studies. ^eThis study did not achieve statistical significance. *Level of significance, *p* < 0.05.

Table 3. Relative increase in uterine weights versus vehicle controls with replicate BPA doses.

Lab	Protocol							
	A 600 mg/kg/day		B 300 mg/kg/day		C 300 mg/kg/day		D 300 mg/kg/day	
	Coded studies	Dose response	Coded studies	Dose response	Coded studies	Dose response	Coded studies	Dose response
1	1.11 (0.90, 1.37) ^{a,b}	—	1.58 (1.21, 2.08)*	—	2.61 (2.23, 3.06)*	—	3.65 (2.84, 4.68)*	—
2	1.45 (1.14, 1.84)*	1.49 (1.25, 1.77)* ^c	1.77 (1.40, 2.24)*	2.30 (1.88, 2.81)*	2.61 (1.99, 3.42)*	2.79 (2.28, 3.41)*	3.91 (3.21, 4.76)*	3.74 (2.89, 4.84)*
3	1.40 (1.14, 1.73)*	—	2.00 (1.62, 2.48)*	—	2.89 (2.24, 3.72)*	—	3.26 (2.51, 4.24)*	—
4	1.36 (1.05, 1.74)*	—	1.45 (1.08, 1.94)*	—	—	—	—	—
5	1.23 (0.95, 1.57) ^b	—	2.02 (1.64, 2.49)*	—	—	—	—	—
6	—	—	—	1.37 (1.05, 1.79)*	—	2.41 (1.79, 3.23)*	—	—
7	—	1.31 (1.03, 1.66)* ^c	—	1.95 (1.64, 2.32)*	—	3.44 (2.76, 4.30)*	—	3.90 (3.18, 4.78)*
8	1.91 (1.58, 2.31)*	—	1.91 (1.53, 2.39)*	1.91 (1.50, 2.43)*	2.89 (2.16, 3.88)*	2.65 (2.16, 3.24)*	—	—
11	1.41 (1.11, 1.80)*	—	1.82 (1.36, 2.44)*	—	3.39 (2.85, 4.05)*	—	4.05 (3.08, 5.33)*	—
12	1.08 (0.43, 2.71) ^{b,d}	1.63 (1.29, 2.06)* ^c	1.60 (1.12, 2.31)*	1.33 (0.88, 1.99) ^b	2.30 (1.67, 3.18)*	2.72 (2.05, 3.61)*	—	—
13	1.25 (1.01, 1.56)* ^d	1.17 (0.79, 1.72) ^{b,c}	1.52 (1.11, 2.08)*	1.72 (1.08, 2.76)*	—	—	—	—
14	1.50 (0.95, 2.37) ^{b,d}	—	2.82 (2.05, 3.87)*	—	—	—	—	—
15	—	—	—	1.37 (1.03, 1.81)*	—	—	—	—
16	—	—	2.11 (1.37, 3.22)*	—	—	—	—	—
17	—	—	2.35 (1.64, 3.37)*	—	—	—	—	—
18	—	—	2.32 (1.88, 2.86)*	2.12 (1.81, 2.50)*	—	—	—	—
19	—	—	—	—	2.42 (1.99, 2.95)*	—	—	—
20	—	—	1.11 (0.86, 1.44) ^b	0.95 (0.69, 1.32) ^b	—	—	—	—
	6/10	3/4	13/14	7/9	7/7	5/5	4/4	2/2

—, laboratory did not perform this particular study. ^aRatio of geometric means of treated blotted uterine weights to the vehicle control blotted uterine weights after adjusting for the body weights at necropsy as a covariable (lower 95% confidence limit, upper 95% confidence limit). ^bThis study did not achieve statistical significance. ^cIn the dose–response studies at this dose, one animal died in laboratory 2, one in laboratory 7, one in laboratory 12, and one in laboratory 13. ^dIn the coded single-dose studies at this dose, three animals died in laboratory 12, one in laboratory 13, and two in laboratory 14. *Level of significance, *p* < 0.05.

OVX animals in protocol C and even in protocol D with the extended dosing period.

The results for the same MX dose in the dose–response and coded single-dose studies are shown in Table 5. The mean uterine responses at the selected MX doses relative to controls were 2 or greater for most laboratories, and often exceed 3 in protocols A and B. All laboratories in their respective response distributions achieved statistical significance in each protocol. In the case of MX, the immature animals in protocol B appeared to have a somewhat higher mean response than the OVX animals in protocol C and even protocol D with the extended dosing period.

The results for the same NP dose in the dose–response and coded single-dose studies are shown in Table 6. In protocol A, 13 of 14 studies achieved statistical significance at a

dose of 250 mg NP/kg/day. This is at first surprising, given that 11 of these laboratories experienced animal mortalities that reduced their power of the already small group size of six. However, the mean relative increase in uterine weights was no lower than 1.71 in any study, and the only laboratory that did not reach statistical significance had only two surviving animals and a mean relative increase of 1.97. In the sc protocols, the mean relative increases in uterine weight at the selected dose of 80 mg NP/kg/day were more modest, and greater than 2 in only 6 of 42 studies. In protocol B, 17 of 24 studies combined from the coded single-dose and the dose–response sets achieved statistical significance. In protocol C, 8 of 12 studies achieved statistical significance. In protocol D, all NP coded samples achieved statistical significance with the extended dosing period.

The results for the same DDT dose in the dose–response and coded single-dose studies are shown in Table 7. In protocol A, all 13 studies achieved statistical significance at a dose of 300 mg DDT/kg/day, as the minimum mean relative increase in uterine weight was 2.67. In the sc protocols at a dose of 100 mg DDT/kg/day, the relative increase in uterine weights was considerably lower, with only 4 of 36 studies greater than 1.5. As a result, only 6 of 19 studies achieved statistical significance in protocol B, 5 of 11 in protocol C, and 4 of 6 studies in protocol D with the extended dosing period.

Dibutylphthalate studies. The results for the DBP studies are shown in Table 8. In protocols A and D, none of the 15 DBP-treated groups were statistically significant versus the vehicle controls. However, in protocol B, the results of 4 of 14 studies, and in protocol C,

Table 4. Relative increase in uterine weights versus vehicle controls with replicate GN doses.

Lab	Protocol							
	A 300 mg/kg/day		B 35 mg/kg/day		C 35 mg/kg/day		D 35 mg/kg/day	
	Coded studies	Dose response	Coded studies	Dose response	Coded studies	Dose response	Coded studies	Dose response
1	2.39 (1.94, 2.95) ^a	2.22 (1.67, 2.95) [*]	2.16 (1.64, 2.84) [*]	2.33 (1.75, 3.10) [*]	1.67 (1.42, 1.97) [*]	1.78 (1.46, 2.18) [*]	2.11 (1.68, 2.65) [*]	2.20 (1.78, 2.73) [*]
2	2.47 (1.93, 3.17) ^a	—	2.95 (2.35, 3.70) [*]	—	2.07 (1.59, 2.70) [*]	—	2.33 (1.93, 2.80) [*]	—
3	2.73 (2.21, 3.36) [*]	—	2.69 (2.17, 3.33) [*]	—	1.66 (1.30, 2.13) [*]	—	1.85 (1.48, 2.32) [*]	—
4	2.58 (2.01, 3.31) [*]	—	2.26 (1.69, 3.03) [*]	—	—	—	—	—
5	1.60 (1.25, 2.05) [*]	—	2.31 (1.88, 2.85) [*]	—	—	—	—	—
8	3.20 (2.65, 3.88) [*]	2.96 (2.42, 3.61) [*]	2.53 (2.03, 3.15) [*]	2.69 (2.19, 3.30) [*]	1.95 (1.46, 2.61) [*]	—	—	—
9	—	2.57 (2.03, 3.25) [*]	—	2.57 (2.19, 3.02) [*]	—	1.87 (1.56, 2.23) [*]	—	2.54 (2.15, 2.99) [*]
11	2.86 (2.25, 3.65) [*]	—	2.38 (1.78, 3.17) [*]	—	1.73 (1.47, 2.05) [*]	—	2.46 (1.93, 3.14) [*]	—
12	3.74 (1.83, 7.62) [*]	3.47 (2.71, 4.45) [*]	2.20 (1.53, 3.17) [*]	2.28 (1.70, 3.05) [*]	1.57 (1.12, 2.20) [*]	1.56 (1.09, 2.21) [*]	—	—
13	2.64 (2.17, 3.22) [*]	—	2.25 (1.64, 3.07) [*]	—	—	—	—	—
14	2.98 (1.93, 4.61) ^a	—	3.44 (2.50, 4.72) [*]	—	—	—	—	—
16	—	—	3.21 (2.10, 4.89) [*]	—	—	—	—	—
18	—	—	2.53 (2.06, 3.11) [*]	—	—	—	—	—
20	—	—	1.32 (1.02, 1.70) [*]	—	—	—	—	—
	10/10	4/4	13/13	4/4	6/6	3/3	4/4	2/2

—, laboratory did not perform this particular study. ^aRatio of geometric means of treated blotted uterine weights to the vehicle control blotted uterine weights after adjusting for the body weights at necropsy as a covariable (lower 95% confidence limit, upper 95% confidence limit). ^bIn the coded single-dose studies at this dose, one animal died in laboratory 2 and one in laboratory 14. ^cLevel of significance, $p < 0.05$.

Table 5. Relative increase in uterine weights versus vehicle controls with replicate MX doses.

Lab	Protocol							
	A 300 mg/kg/day		B 500 mg/kg/day		C 500 mg/kg/day		D 500 mg/kg/day	
	Coded studies	Dose response	Coded studies	Dose response	Coded studies	Dose response	Coded studies	Dose response
1	2.97 (2.39, 3.71) ^a	2.59 (2.13, 3.15) [*]	2.73 (2.08, 3.58) [*]	2.47 (1.91, 3.21) [*]	1.96 (1.67, 2.30) [*]	2.32 (1.86, 2.89) [*]	2.23 (1.77, 2.80) [*]	2.38 (1.76, 3.22) [*]
2	3.14 (2.47, 4.00) [*]	—	3.01 (2.39, 3.80) [*]	—	2.08 (1.58, 2.73) [*]	—	2.71 (2.21, 3.31) [*]	—
3	2.77 (2.24, 3.41) [*]	2.94 (2.34, 3.69) ^a	2.66 (2.15, 3.29) [*]	2.98 (2.42, 3.65) [*]	2.11 (1.64, 2.71) [*]	2.42 (1.96, 2.98) [*]	2.67 (2.03, 3.51) [*]	2.46 (1.98, 3.06) [*]
4	3.01 (2.34, 3.86) [*]	—	3.33 (2.49, 4.45) [*]	—	—	—	—	—
5	3.10 (2.41, 3.99) [*]	—	3.61 (2.93, 4.45) [*]	—	—	—	—	—
8	3.71 (3.07, 4.49) [*]	—	2.91 (2.33, 3.63) [*]	—	2.08 (1.54, 2.80) [*]	—	—	—
11	3.46 (2.67, 4.48) [*]	—	2.39 (1.81, 3.16) [*]	—	3.14 (2.63, 3.75) [*]	—	3.34 (2.55, 4.36) [*]	—
12	3.20 (1.34, 7.61) [*]	3.98 (3.07, 5.15) [*]	3.14 (2.18, 4.51) [*]	3.34 (2.53, 4.40) [*]	1.49 (1.09, 2.03) [*]	1.95 (1.45, 2.62) [*]	—	—
13	3.31 (2.72, 4.02) [*]	—	2.89 (2.11, 3.96) [*]	—	—	—	—	—
14	2.95 (1.94, 4.48) ^a	3.46 (2.51, 4.77) [*]	4.07 (2.97, 5.56) [*]	3.76 (2.78, 5.09) [*]	—	—	—	—
16	—	—	4.29 (2.81, 6.55) [*]	—	—	—	—	—
17	—	—	3.25 (2.28, 4.63) [*]	—	—	—	—	—
18	—	—	3.18 (2.59, 3.90) [*]	—	—	—	—	—
20	—	—	1.76 (1.37, 2.28) [*]	—	—	—	—	—
	10/10	4/4	14/14	4/4	6/6	3/3	4/4	2/2

—, laboratory did not perform this particular study. ^aRatio of geometric means of treated blotted uterine weights to the vehicle control blotted uterine weights after adjusting for the body weights at necropsy as a covariable (lower 95% confidence limit, upper 95% confidence limit). ^bIn the dose–response studies at this dose, one animal died in laboratory 3. ^cIn the coded single-dose studies at this dose, three animals died in laboratory 14. ^{*}Level of significance, $p < 0.05$.

the results of 1 of 7 studies, achieved statistical significance. Of these five studies, three had significantly increased blotted uterine weights when treated with DBP, whereas the other two had significantly decreased blotted uterine weights when treated with DBP.

Discussion and Conclusions

The OECD is composed of over 20 nations, and OECD protocols such as the uterotrophic bioassay are intended for use in all of the member nations. As such, this validation study was carried out in 21 laboratories in

nine nations. Funding for the study came primarily from national regulatory agencies and industry associations, but several laboratories freely contributed their time and effort to the study. This large, international nature of the program, however, increased the organizational and logistical workload. For example, the protocol had to be clearly understood by speakers of a variety of languages for the procedures to be performed in a similar manner in all laboratories. Data had to be recorded in the different laboratories and provided to an independent statistician in a

accurate, timely, and efficient manner. The animal husbandry supplies, vehicles, and reagents, as well as the laboratory animal themselves, also had to be widely and readily available. Finally, the central repository had to deal with international shipments with different customs regulations and laboratory safety regulations.

The Validation Management Group addressed these challenges with several efforts. Both the ovariectomy and uterine dissection procedures were videotaped, and the videotape was distributed to the technical staff of

Table 6. Relative increase in uterine weights versus vehicle controls with replicate NP doses.

Lab	Protocol							
	A 250 mg/kg/day		B 80 mg/kg/day		C 80 mg/kg/day		D 80 mg/kg/day	
	Coded studies	Dose response	Coded studies	Dose response	Coded studies	Dose response	Coded studies	Dose response
1	1.71 (1.37, 2.14) ^a	—	1.65 (1.26, 2.17) [*]	—	1.43 (1.22, 1.68) [*]	—	1.54 (1.23, 1.93) [*]	—
2	2.03 (1.48, 2.77) ^a	—	1.34 (1.06, 1.68) [*]	—	1.24 (0.95, 1.62) ^c	—	1.86 (1.55, 2.24) [*]	—
3	1.80 (1.43, 2.27) [*]	—	1.81 (1.46, 2.24) [*]	—	1.37 (1.07, 1.76) [*]	—	1.73 (1.37, 2.18) [*]	—
4	1.89 (1.24, 2.88) ^a	2.61 (1.69, 4.04) ^d	1.45 (1.08, 1.94) [*]	2.05 (1.44, 2.92) [*]	—	—	—	—
5	1.74 (1.28, 2.35) ^a	—	1.64 (1.30, 2.07) [*]	—	—	—	—	—
6	—	—	—	1.24 (0.91, 1.68) ^c	—	1.16 (0.90, 1.48) ^c	—	2.11 (1.73, 2.58) [*]
7	—	2.17 (1.72, 2.74) ^d	—	1.68 (1.36, 2.08) [*]	—	1.64 (1.32, 2.03) [*]	—	—
8	2.89 (2.33, 3.57) ^a	—	1.32 (1.06, 1.65) [*]	1.44 (1.15, 1.80) [*]	1.59 (1.19, 2.13) [*]	1.17 (0.98, 1.39) ^c	—	—
9	—	2.17 (1.62, 2.90) [*]	—	1.86 (1.47, 2.36) [*]	—	1.23 (0.99, 1.52) ^c	—	1.83 (1.51, 2.21) [*]
11	2.33 (1.65, 3.28) ^a	—	2.05 (1.54, 2.73) [*]	—	1.38 (1.16, 1.63) [*]	—	2.02 (1.57, 2.60) [*]	—
12	1.97 (0.73, 5.33) ^b	2.95 (2.02, 4.32) ^d	1.71 (1.19, 2.47) [*]	2.02 (1.49, 2.75) [*]	1.38 (1.01, 1.90) [*]	1.33 (1.02, 1.73) [*]	—	—
13	2.24 (1.81, 2.78) ^a	—	1.08 (0.79, 1.48) ^c	—	—	—	—	—
14	2.05 (1.17, 3.59) ^a	—	1.72 (1.26, 2.35) [*]	—	—	—	—	—
15	—	—	—	1.22 (0.91, 1.65) ^c	—	—	—	—
16	—	—	1.30 (0.85, 1.99)	—	—	—	—	—
17	—	—	2.49 (1.74, 3.55) [*]	—	—	—	—	—
18	—	—	1.73 (1.40, 2.14) [*]	1.93 (1.56, 2.39) [*]	—	—	—	—
19	—	—	1.22 (0.76, 1.96) ^c	—	1.40 (1.15, 1.70) [*]	—	—	—
20	—	—	1.07 (0.83, 1.38) ^c	0.75 (0.51, 1.11) ^c	—	—	—	—
	9/10	4/4	11/15	6/9	6/7	2/5	4/4	2/2

—, laboratory did not perform this particular study. ^aRatio of geometric means of treated blotted uterine weights to the vehicle control blotted uterine weights after adjusting for the body weights at necropsy as a covariable (lower 95% confidence limit, upper 95% confidence limit). ^bIn the coded single-dose studies, one animal died in laboratory 2, four animals died in laboratory 4, two animals died in laboratory 5, one animal died in laboratory 8, two animals died in laboratory 11, four animals died in laboratory 12, one animal died in laboratory 13, and four animals died in laboratory 14. ^cThis study did not achieve statistical significance. ^dIn the dose-response studies at this dose, two animals died in laboratory 4, one animal died in laboratory 7, and three animals died in laboratory 12. *Level of significance $p < 0.05$.

Table 7. Relative increase in uterine weights versus vehicle controls with replicate DDT doses.

Lab	Protocol							
	A 300 mg/kg/day		B 100 mg/kg/day		C 100 mg/kg/day		D 100 mg/kg/day	
	Coded studies	Dose response	Coded studies	Dose response	Coded studies	Dose response	Coded studies	Dose response
1	2.70 (2.15, 3.39) ^a	—	2.05 (1.57, 2.70) [*]	—	1.65 (1.41, 1.93) [*]	—	1.05 (0.83, 1.31) ^b	—
2	3.68 (2.88, 4.71) ^a	—	1.13 (0.90, 1.42) ^b	—	1.43 (1.09, 1.86) [*]	—	1.21 (1.01, 1.46) [*]	—
3	3.05 (2.45, 3.81) [*]	2.67 (1.99, 3.59) [*]	1.18 (0.95, 1.46) ^b	1.01 (0.83, 1.23) ^b	1.29 (1.0035 ^d , 1.65) [*]	1.43 (1.21, 1.69) [*]	1.14 (0.92, 1.43) ^b	1.18 (1.03, 1.36) [*]
4	3.76 (2.91, 4.87) [*]	—	1.57 (1.17, 2.10) [*]	—	—	—	—	—
5	2.92 (2.23, 3.83) [*]	2.71 (1.92, 3.24) [*]	0.95 (0.78, 1.17) ^b	1.18 (0.91, 1.54) ^b	—	—	—	—
8	3.87 (3.18, 4.71) [*]	—	1.06 (0.85, 1.32) ^b	—	1.17 (0.87, 1.57) ^b	1.25 (0.98, 1.59) ^b	—	—
11	3.58 (2.79, 4.60) [*]	3.43 (2.96, 3.98) [*]	1.03 (0.78, 1.37) ^b	1.08 (0.87, 1.34) ^b	1.24 (1.05, 1.47) [*]	—	1.31 (1.03, 1.68) [*]	1.48 (1.17, 1.87) [*]
12	See note ^b	3.45 (2.41, 4.94) ^a	1.50 (1.04, 2.16) [*]	1.47 (1.11, 1.94) [*]	0.96 (0.71, 1.31) ^b	1.31 (0.96, 1.78) ^b	—	—
13	4.12 (3.32, 5.12) [*]	—	1.79 (1.31, 2.45) [*]	—	—	—	—	—
14	4.26 (2.65, 6.83) ^a	—	1.17 (0.85, 1.60) ^b	—	—	—	—	—
16	—	—	1.29 (0.84, 1.97) ^b	—	—	—	—	—
17	—	—	1.46 (1.02, 2.08) [*]	—	—	—	—	—
18	—	—	0.98 (0.80, 1.21) ^b	—	—	—	—	—
19	—	—	1.07 (0.67, 1.69) ^b	—	1.06 (0.87, 1.29) ^b	—	—	—
20	—	—	0.79 (0.61, 1.02) ^b	—	—	—	—	—
	9/9	4/4	5/15	1/4	4/8	1/3	2/4	2/2

—, laboratory did not perform this particular study. ^aRatio of geometric means of treated blotted uterine weights to the vehicle control blotted uterine weights after adjusting for the body weights at necropsy as a covariable (lower 95% confidence limit, upper 95% confidence limit). ^bThis study did not achieve statistical significance. ^cIn the coded single-dose studies, all six animals died in laboratory 12, and two animals died in laboratory 14. ^dWith the lower confidence level number > 1.00, the result is statistically significant. ^eIn the dose-response studies at this dose, one animal died in laboratory 12. *Level of significance, $p < 0.05$.

all the participating laboratories. The draft protocols were distributed to all national authorities and participating laboratories for comments and inquiries for any ambiguities. A common electronic spreadsheet was constructed and distributed for comment so the data could be recorded and electronically transmitted to the independent statistician.

Despite these efforts and preparations, some laboratories encountered difficulty with certain dose-preparation instructions, two errors in the spreadsheet itself were later discovered, and the breakage of some vials during shipment required their rapid replacement because of the imminent delivery of immature animals whose births were timed for protocols A and B. Given the number of laboratories and individual studies, these were minor problems that did not affect the quality or the success of the results.

It should also be recognized that the protocols allowed variations in a number of experimental conditions. These variables include the choice of rat strain, the laboratory diet, housing and husbandry practices such as the use of cage bedding, the administration vehicle, and to a modest degree, the age of both immature and OVX animals. The judgment was that rigorous and detailed standardization of all of these variables would constrain

the ability to widely and easily practice the uterotrophic bioassay in many of the OECD member nations, where the intended purpose is as a rapid screening bioassay for a large number of chemicals. The laboratory specifics for most laboratories have been described previously (Table 1 in Kanno et al. 2001; Table 7 in Kanno et al. 2003) or can be found for the remaining laboratories in Table 9.

The coded nature of the study also introduced some difficulties. To avoid giving very specific information that could be used to identify the coded test substances, broad general advice was given about dose preparation. Unfortunately, estrogen receptor agonists and antagonists tend to be hydrophobic and to have limited solubility. As noted, some laboratories encountered difficulty in solubilizing the test substances, and two laboratories decided to halt administration of particular preparations rather than administer apparent suspensions. This experience also suggests another source of variation in administered doses among the participating laboratories.

As with the dose-response studies, there was a consistent association in the coded single-dose studies between the occurrence of mortalities, reduced body weight gain, and clinical signs with the weak agonists DDT and NP in protocol A, and for reduced body

weight gain with the EE high dose, BPA, and MX in protocol D. The 10% and greater differences in body weights between vehicle and treated animals occurred within just 4 days (protocols A, B, and C) or 8 days (protocol D) of treatment initiation, indicating a rapid onset of systemic toxicity at those doses. Despite the apparent magnitude of these insults, the uterine response appeared to remain undiminished, confirming the underlying robustness of this biological response for estrogen-screening programs.

Overall, for each protocol, the mean relative increase in uterine weight was reproducible within and among laboratories for both the dose-response and coded single-dose studies with each test substance. The dose-response results for each protocol and test substance are in the accompanying paper (Kanno et al. 2003). It is important to distinguish between when the results for a given test substance have been consistently reproduced within and across laboratories over time from whether statistical significance was consistently achieved in all or none of the laboratories. The objective here is the former, the reproducibility of the bioassay. The results here should be interpreted by taking into account the following considerations. First, several of selected doses were in the lower regions of a substance's dose-response curve (Kanno et al. 2001, 2003). Second, the lower region of the dose-response curve implies a distribution of statistically positive and negative responses, with the ratio between positive and negative results depending upon the precise dose employed in the dose response of that particular substance. That is, the rate of studies lacking statistical significance should rise as the doses move further down the dose-response curve for a substance, particularly in the case of weak agonists when the slope of the dose response is shallow. Several doses herein were at or near maximum uterine responses, for example, the high EE po and sc doses, the GN and MX po and sc doses, and the DDT po dose, and these doses consistently achieved statistical significance. Where the selected doses were increasingly in the lower portion of the dose-response curve, although the numerical results were reproducible within and across laboratories, an increasing number of studies

Table 8. Relative increase in uterine weights versus vehicle controls with replicate DBP doses.

Lab	Protocol			
	A 1000 mg/kg/day	B 500 mg/kg/day	C 500 mg/kg/day	D 500 mg/kg/day
1	0.91 (0.74, 1.13) ^a	0.97 (0.74, 1.28)	1.37 (1.17, 1.61) ^{*b}	0.91 (0.73, 1.15)
2	0.99 (0.78, 1.26)	1.04 (0.83, 1.31)	0.99 (0.76, 1.28)	1.03 (0.86, 1.24)
3	1.00 (0.81, 1.23)	1.01 (0.81, 1.25)	0.90 (0.70, 1.15)	1.01 (0.81, 1.26)
4	0.99 (0.77, 1.28)	0.85 (0.64, 1.14)	—	—
5	1.03 (0.81, 1.32)	1.06 (0.86, 1.31)	—	—
8	0.98 (0.81, 1.18)	1.00 (0.80, 1.24)	1.24 (0.92, 1.65)	—
11	0.95 (0.75, 1.22)	1.39 (1.06, 1.82) ^{*b}	0.99 (0.83, 1.16)	1.00 (0.78, 1.28)
12	0.91 (0.38, 2.20)	—	0.97 (0.67, 1.40)	0.93 (0.68, 1.26)
13	0.86 (0.69, 1.07)	0.98 (0.72, 1.34)	—	—
14	0.91 (0.60, 1.38)	1.38 (1.01, 1.89) ^{*b}	—	—
16	—	0.90 (0.59, 1.38)	—	—
17	—	0.88 (0.62, 1.26)	—	—
18	—	0.74 (0.6, 0.91) ^{*b}	—	—
19	—	0.75 (0.47, 1.20)	0.84 (0.69, 1.02)	—
20	—	0.73 (0.56, 0.96) ^{*b}	—	—
	0/10	4/14	1/7	0/5

—, laboratory did not perform this particular study. ^aRatio of geometric means of treated blotted uterine weights to the vehicle control blotted uterine weights after adjusting for the body weights at necropsy as a covariable (lower 95% confidence limit, upper 95% confidence limit). ^bThis study did not achieve statistical significance. *Level of significance $p < 0.05$.

Table 9. Details for animals, diet, vehicles, and bedding in laboratories participating only in coded single-dose studies.^a

Lab	Strain of rat ^b		Animal diet ^b		For oral gavage ^b		For sc injection ^b		Bedding
	Immature rats	OVX rats	Immature	OVX	Vehicle 1	Vehicle 2	Vehicle 1	Vehicle 2	
16	Wistar	NA	Altromin 1324 FORTII	NA	NA	NA	Peanut oil	NA	Wood chip - low dust
17	Wistar/Han	NA	Altromin 1324	NA	NA	NA	Peanut oil	NA	Tapvei bedding
19	CD Sprague-Dawley	CD Sprague-Dawley	RM1(E) SQC expanded pellet	RM1(E) SQC expanded pellet	NA	NA	Corn oil	Ethanol/corn oil	Lignocel grade 4/4 woodflakes (immature)/ none for OVX

^aThe details for laboratories participating in the dose-response studies and that may have participated in the coded single-dose studies herein can be found in Table 7 in Kanno et al. (2003). ^bDetailed information is available from the corresponding author of this article.

failed to achieve a statistically significant difference, for example, the BPA po dose, the NP sc dose for adult OVX animals, and certainly the DDT sc dose.

To assess reproducibility, the mean relative uterine weight increases were calculated in an overall global analysis (Table 10). The uterotrophic responses were consistent and reproducible between the dose–response and the coded single-dose studies without exception for every test substance and every protocol. The global analysis in Table 10 also shows subtle test substance–specific differences in the protocols that were consistent in both the dose–response and coded single-dose studies. Comparing the intact, immature, and adult OVX versions as protocols B and C, respectively, the adult OVX version appears to be more responsive with BPA, whereas the intact, immature version appears to be more responsive with GN and MX. More than doubling the time of treatment with extended dosing (protocol D), did increase the response with BPA, and marginally with GN, MX, and NP. The global analysis includes the results of all laboratories, regardless of mortalities in protocol A or the possible issues with laboratories 19 and 20 that are discussed below. Except for the lower means in the coded single-dose, high-dose EE studies for protocols B and C, no overall impact of their inclusion was observed.

The data were analyzed for an association between uterine weights and body weights and for the variability and power of the wet and the blotted weights. Although there was no consistent correlation between uterine weight and body weight, the data suggest that

body weight is more strongly correlated with uterine weight in the immature animals than in the adult OVX animals. As with phase 1 and the dose–response studies, wet uterine weights were more variable than blotted weights (Kanno et al. 2001, 2003). The blotted uterine weights in phase 2, again, showed slightly less interlaboratory and intragroup variability than wet weights with imbibed fluid, suggesting that blotted uterine weight will provide slightly better power for detecting uterotrophic effects than the wet weight.

In 5 of 36 studies, the uterine weights after DBP treatment were statistically different from controls, indicating a certain rate of false positives and negatives will occur. Three sets of results were statistically higher than the vehicle groups, a false positive rate of about 8%, and two were statistically lower. This nearly even division into higher and lower differences supports random chance due to variability about the baseline. In further support, the margins by which the respective upper and lower 95% confidence intervals achieved statistical significance were minimal (Table 8). In absolute terms, the mean relative increase in uterine weight in these three incidents was just under 40% and suggests a source of variability in the uterine weight from one group to another. When the raw body weight and uterine weight data were in these laboratories were examined, there were no obvious anomalies or inconsistencies such as outliers or high standard deviations when compared with other laboratories. When the overall patterns of these laboratories were assessed, one (laboratory 20) had the minimum response with five of six test chemicals and was below

average for the two EE doses, consistent with its statistically decreased result. A second (laboratory 14) had responses that were the maximum with two test substances and above average for the remainder and the EE doses, consistent with its statistically increased result. The patterns of the other laboratories were unremarkable. Four of the five incidents occurred with immature animals. Although body weights were randomized, there is the possibility of group-to-group variations based on a litter-related effect. The animals used would have been born on the same day, meaning that the animals were likely from a limited number of litters. In fact, some investigators have taken the precaution to also randomize their groups by litter (Christian et al. 1998). As the litter of origin for each individual was not recorded, this possibility cannot be assessed here. It is clear, however, that borderline false positives can occur with the present protocols, and that a weight-of-the-evidence integration of the uterotrophic results with other structural, *in vitro*, and *in vivo* data may be necessary for interpretation. Similarly, false negatives may also occur, and data to qualify the performing laboratory and criteria to accept the results may be necessary (Owens et al. 2003).

The results in three laboratories deserve comment. These laboratories displayed a trend toward lower responsiveness to both the EE and to several weak agonists when compared with other laboratories. The performance of laboratory 6 with its high vehicle control weights and limited responsiveness in some cases has been previously noted (Kanno et al. 2003). Here, we also note the lower general response in this laboratory to the EE doses in protocol B (Tables 1 and 2). Laboratory 19 observed no statistically significant uterotrophic responses for the test substances it could formulate or either of the two EE doses in protocol B (Tables 1, 2, 6, and 7). The pattern of responses in this laboratory in protocol C, however, was unremarkable when compared with other laboratories. A close examination of the data, including dietary analyses, has not revealed any apparent reasons for this lack of responsiveness. Laboratory 20 observed statistical significance with both EE doses, but the relative increases in weight were somewhat lower than other labs at the low EE dose and among the lowest at the high EE dose (Tables 1 and 2). Although statistical significance was observed with GN and MX, the increases in the uterine weights were the lowest observed in any laboratory (Tables 4 and 5). Statistical significance was not observed in either of the dose–response or the coded single-dose studies with either BPA or NP, and again, the increase in the uterine weights were the lowest observed in any laboratory (Tables 3 and 6). A review of the data and laboratory variables first indicated that the vehicle control

Table 10. Global analysis of results.

Substance/dose	Protocol			
	A	B	C	D
BPA mg/kg/day	600	300	300	300
DR	1.41 (1.07, 1.85) ^a	1.61 (1.00, 2.58)	2.73 (2.07, 3.61)	3.78 (2.98, 4.79)
CSD	1.34 (1.09, 1.66)	1.85 (1.58, 2.16)	2.68 (2.36, 3.04)	3.84 (3.39, 4.35)
DDT mg/kg/day	300	100	100	100
DR	3.13 (2.38, 4.12)	1.16 (0.94, 1.44)	1.33 (1.04, 1.69)	1.31 (1.08, 1.59)
CSD	3.60 (2.94, 4.41)	1.23 (0.97, 1.58)	1.24 (1.00, 1.52)	1.17 (1.06, 1.30)
GN mg/kg/day	300	35	35	35
DR	2.75 (1.98, 3.80)	2.47 (1.82, 3.37)	1.73 (1.45, 2.07)	2.36 (1.61, 3.46)
CSD	2.65 (2.21, 3.18)	2.42 (2.05, 2.86)	1.77 (1.58, 2.00)	2.18 (1.91, 2.49)
MX mg/kg/day	300	500	500	500
DR	3.16 (2.09, 4.79)	3.13 (1.70, 5.75)	2.25 (1.79, 2.83)	2.43 (1.55, 3.83)
CSD	3.21 (2.58, 3.99)	3.03 (2.54, 3.62)	2.07 (1.72, 2.48)	2.62 (2.28, 3.00)
NP mg/kg/day	250	80	80	80
DR	2.40 (1.90, 3.04)	1.51 (1.05, 2.16)	1.29 (1.06, 1.58)	1.96 (1.59, 2.42)
CSD	2.12 (1.72, 2.61)	1.53 (1.26, 1.88)	1.40 (1.24, 1.57)	1.77 (1.58, 1.98)
EE µg/kg/day	1	0.3	0.3	0.3
DR	2.27 (1.71, 3.02)	2.42 (1.86, 3.13)	2.33 (1.97, 2.76)	3.34 (2.79, 4.01)
CSD	2.57 (1.88, 3.51)	2.18 (1.64, 2.90)	2.30 (2.02, 2.62)	3.50 (2.80, 4.37)
EE µg/kg/day	3	1	1	1
DR	3.42 (2.56, 4.57)	5.09 (2.44, 10.62)	3.40 (2.87, 4.03)	4.51 (3.75, 5.43)
CSD	3.78 (2.83, 5.05)	3.56 (2.61, 4.85)	2.67 (1.60, 4.43)	4.87 (4.34, 5.45)
DBP mg/kg/day	1,000	500	500	500
CSD	0.95 (0.77, 1.18)	0.97 (0.80, 1.17)	1.02 (0.84, 1.24)	0.99 (0.91, 1.07)

Abbreviations: CSD, coded single-dose studies; DR, dose–response studies.

^aRatio of geometric means of treated blotted uterine weights to the vehicle control blotted uterine weights after adjusting for the body weights at necropsy as a covariable (lower 95% confidence limit, upper 95% confidence limit).

uterine weights were > 50 mg, which was well above the 20- to 40-mg range in most other laboratories. Then, an analysis of laboratory diets for phytoestrogens found that laboratory 20's diet had the highest combined total GN and daidzein levels of > 500 µg/g diet. This leads to the suspicion that the dynamic range of the bioassay in this particular case may have been impaired by the high phytoestrogen content of the diet (Owens et al. 2003).

Collectively with other observations in the dose–response studies, these data suggest the need to monitor the uterine weights of vehicle control animals, to specify that laboratory diets have low to modest phytoestrogen levels (< 350 µg/g diet) (Owens et al. 2003), and to qualify laboratories with both reference and weak agonists before performing tests of unknown substances. In addition, care should be taken not to exceed the maximum tolerated dose, to reduce animal pain, suffering, and mortalities. The results in the current coded dose study provide additional evidence that a strong uterine response occurs even in the presence of severe systemic toxicity. The robustness of the uterine response in turn supports its use in a screening assay.

In conclusion, the uterotrophic bioassay yields reproducible results within the same laboratory and across the participating laboratories over time with a range of test substances including the EE positive reference substance, the five weak agonist substances (BPA, GN, MX, NP, and DDT), and the negative substance (DBP). The results of the dose–response and coded single-dose studies are in agreement. No substantive performance differences were found between the different versions or their protocols that would support one version being consistently superior to another. Therefore, both the intact immature and OVX versions of the uterotrophic bioassay and the protocols herein are judged to be qualitatively equivalent to one another. Low rates of false negatives and false positives were observed. The false negatives occurred with very weak agonists (BPA,

DDT, and NP) in the lowest portions of the their dose–response curves. The false-positive rate with DBP was just over 8%, with mean relative weight increases of 30–40%, suggesting the importance of controlling group-to-group variations in the baseline and using a weight-of-the-evidence approach in interpreting very modest responses. These and other results from the dose–response studies and the dietary analyses will be used to develop the draft OECD test guideline for the uterotrophic bioassay. These results will be submitted along with other data for independent peer review to provide support for the validation of the uterotrophic bioassay.

We acknowledge the dedicated efforts and work of the participating labs in generating these data, namely, the Institute of Environmental Toxicology, Japan; Mitsubishi Chemical Safety Institute, Japan; Japan Bioassay Research Centre, Japan; Sumitomo Chemical Company, Japan; Syngenta Central Toxicology Laboratory, United Kingdom; WIL Research Laboratories, United States; BASF, Germany; Bayer AG, Germany; Aventis Crop Sciences, France; Exxon Biomedical Sciences Inc., United States; Free University of Berlin, Germany; National Institute of Toxicological Research, Korea; Huntingdon Life Sciences, United Kingdom; Institute of Food Safety and Toxicology, Denmark; Research Toxicology Center, Italy; and the Institute of Biomedical Research, Italy. The acquisition of the chemical substances and the chemical repository for this assay validation was sponsored by the European Chemical Industry Council and the Japanese Chemical Industry Association. Nonylphenol was kindly donated by Schenectady International Inc., (Schenectady, NY, USA) and bisphenol A was kindly donated by Bayer AG. We thank E. Zeiger and H. Koëter for reviewing the manuscript.

REFERENCES

- Blair R, Fang H, Branham WS, Hass B, Dial SL, Moland CL, et al. 2000. Estrogen receptor relative binding affinities of 188 natural and xenochemicals: structural diversity of ligands. *Toxicol Sci* 54:138–153.
- Christian MS, Hoberman AM, Bachmann S, Hellwig J. 1998. Variability in the uterotrophic response assay (an *in vivo* estrogenic response assay) in untreated control and positive control (DES-DP, 2.5 µg/kg, BID) Wistar and Sprague-Dawley rats. *Drug Chem Toxicol* 21(suppl 1): 51–100.
- Ema M, Miyawaki E. 2001. Adverse effects on development of the reproductive system in male offspring of rats given monobutyl phthalate, a metabolite of dibutyl phthalate, during late pregnancy. *Reprod Toxicol* 15:189–194.
- Ema M, Miyawaki E, Kawashima K. 2000. Critical period for adverse effects on development of reproductive system in male offspring of rats given di-*n*-butyl phthalate during late pregnancy. *Toxicol Lett* 111:271–278.
- ICCVAM (Intraagency Coordinating Committee on the Validation of Alternative Methods). 1997. Validation and regulatory acceptance of toxicological test methods. A report of the Ad Hoc Interagency Coordinating Committee on the Validation of Alternative Methods, NIH Report No. 97-3981. Research Triangle Park, NC:National Institute of Environmental Health Sciences.
- 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 I. *Environ Health Perspect* 109:785–794.
- Kanno J, Onyon L, Peddada S, Ashby J, Owens W. 2003. The OECD program to validate the rat uterotrophic bioassay. Phase 2: dose–response studies. *Environ Health Perspect* 111:1530–1549.
- Mylchreest E, Cattley RC, Foster PMD. 1998. Male reproductive tract malformations in rats following gestational and lactational exposure to di(*n*-butyl) phthalate: an antiandrogenic mechanism? *Toxicol Sci* 43:47–60.
- Mylchreest E, Sar M, Cattley RC, Foster PMD. 1999. Disruption of androgen-regulated male reproductive development by di(*n*-butyl) phthalate during late gestation in rats is different from flutamide. *Toxicol Appl Pharmacol* 156:81–95.
- OECD (Organisation for Economic Co-operation and Development). 1998. The Validation of Test Methods Considered for Adoption as OECD Test Guidelines. ENV/MC/CHEM(98)6, Paris:OECD.
- . 2000. Guidance Document on the Recognition, Assessment, and Use of Clinical Signs as Humane Endpoints for Experimental Animals Used in Safety Evaluation. OECD Environmental Health and Safety Publications. Series on Testing and Assessment, No. 19. ENV/JM/MOMO(2000)7. Paris:OECD.
- . 2002. OECD Series on Principles of Good Laboratory Practice and Compliance Monitoring. Twelve documents are available for downloading, including “OECD principles of Good Laboratory Practice.” Available: <http://webnet1.oecd.org/EN/document/0,,EN-document-526-14-no-24-6553-0,00.html> [accessed 1 August 2002].
- Owens W, Ashby J, Odum J, Onyon L. 2003. The OECD program to validate the rat uterotrophic bioassay. Phase 2: dietary phytoestrogen analyses. Schulz VD, Phillips S, Sar M, Foster PMD, Gaido KW. 2001. Altered gene profiles in fetal rat testes after *in utero* exposure to di(*n*-butyl)phthalate. *Toxicol Sci* 64:233–242.
- Zacharewski TR, Meek MD, Clemons JH, Wu ZF, Fielden MR, Matthews JB. 1998. Examination of the *in vitro* and *in vivo* estrogenic activities of eight commercial phthalate esters. *Toxicol Sci* 46:282–293.