classified as positive in an assay (ICCVAM 1997). Therefore, the term responsiveness is used herein. The striking observation was that these differences in responsiveness were test substance specific. For BPA, sc injection achieved statistical significance at lower doses than oral gavage, and the maximum induction was consistently higher by sc injection. Likewise for GN, most sc studies achieved statistical significance at a somewhat lower dose, 15 mg GN/kg/day, and with greater consistency than oral gavage. For NP, the majority of oral gavage and sc studies achieved statistical significance at similar doses, 75 mg NP/kg/day and 80 mg/kg/day, respectively. For MX, the majority of oral gavage studies achieved statistical significance at the lowest doses tested and were near their maximum induction at 20 mg MX/kg/day, whereas sc injection doses were higher and the maximum uterine weight increase was lower. For *o,p'*-DDT, oral gavage produced statistical significance at lower doses and higher maximum responses. In the sc administration studies, an overall difference was not discernable between the intact, immature version (protocol B) or the adult OVX version (protocol C). Additionally, the satellite oral gavage studies using OVX animals produced results similar to the intact, immature animals in protocol A in both the maximum fold induction of the uteri and the first dose reaching statistical difference. Collectively, these results suggest that no route of administration with various agonists will be consistently the most sensitive. The substantial equivalence of the results indicates that the choice of route of administration will then depend on the purpose for which the assay is used, such as detecting the activity of a substance at the lowest minimal effective dose or providing a route of administration relevant for human and wildlife exposure.

Five dose–response studies of 84 (not including the two studies in the laboratory that did not record terminal body weights) did not observe statistically significant increases with three substances: NP (three studies), BPA (one study), and *o,p'*-DDT (one study). These three substances are the lowest estrogen receptor–binding affinities, once MX metabolism in the liver to dihydroxymethoxychlor (HPTE) is considered (see Table 1 for binding affinities of each substance, including HPTE). A closer examination of the circumstances and data has been made to see if these cases were approaching statistical significance or what other circumstances may have intervened to prevent detection of statistical significance (Table 28). In these studies, statistical significance is achieved when the lower 95% confidence interval for the mean of the test substance is > 1.0-fold induction of the uterine weight.

Given that literature data and expert judgment were used to select the doses; that no range-finding studies were performed; that the range of the doses used was sometimes only a little more than an order of magnitude; and that these laboratories did not include the highest doses in their studies, these few studies lacking statistical significance may have been anticipated because of program design and not the performance shortcomings of the bioassay. In four of five cases, the studies did not test the highest dose of the five prescribed doses, reducing the opportunity for detecting a statistically significant response. These four cases are examined in detail.

In the first case involving NP in protocol B, the mean control blotted uterine weight in laboratory 6 was 58.0 mg, where the vehicle control means in most other immature control groups were < 40 mg. This would, in theory, be expected to diminish the study's responsiveness. Despite this possible impediment, the lower 95% confidence interval for the relative ratio for uterine weight increase was 0.91, indicating that the study was approaching statistical significance. In comparison with other protocol B NP studies, six of nine studies achieved statistical significance with uterine weight increases at 35 or 80 mg NP/kg/day, and a seventh achieved statistical significance at the highest dose of 100 mg NP/kg/day.

In the second case involving NP in protocol C, laboratory 6 was again approaching

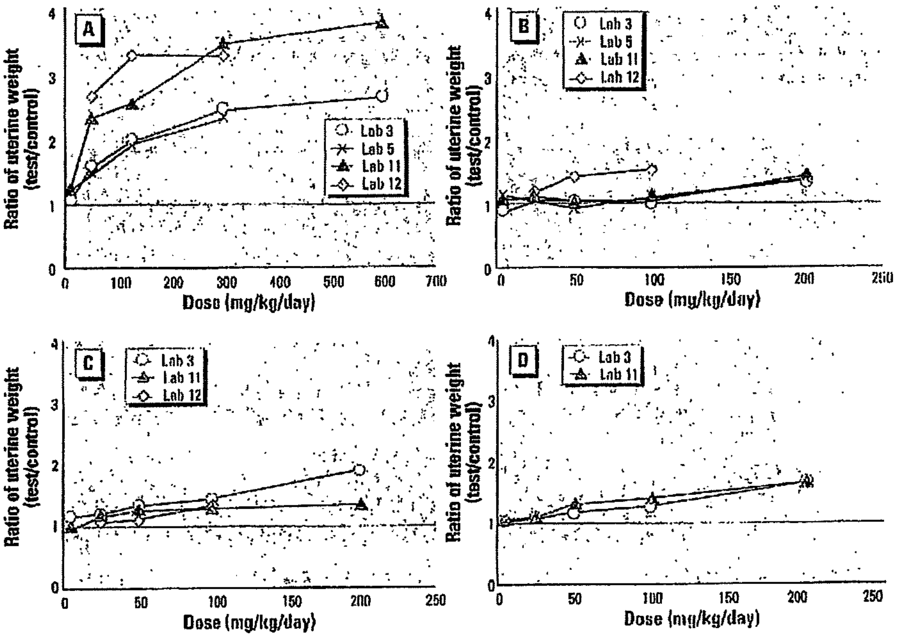

Figure 5. Ratio of the mean absolute blotted uterine weight in response to doses of *o,p'*-DDT (DDT) relative to the vehicle control group. (*A*) Participating laboratory results for protocol A using immature female rats, dosing by oral gavage for 3 consecutive days. (*B*) Participating laboratory results for protocol B using immature female rats, dosing by sc injection for 3 consecutive days. (*C*) Participating laboratory results for protocol C using adult OVX rats, dosing by sc injection for 3 consecutive days. (*D*) Participating laboratory results for protocol C using adult OVX rats and extending sc injection dosing to 7 days. In all cases, animals were humanely sacrificed 24 hr after the last dose administration, the uteri were removed and trimmed, and wet and blotted weights were recorded.

Table 26. Uterine weights, body weights, and ratio of the relative increase in uterine weights for *o,p'*-DDT in satellite OVX protocol by oral gavage.

Laboratory	Measure	Vehicle	Dose 1 (10 mg/kg/day)	Dose 2 (50 mg/kg/day)	Dose 3 (125 mg/kg/day)	Dose 4 (300 mg/kg/day)	Dose 5 (600 mg/kg/day)
12	Wet weight (mg, mean ± SD)	101.1 ± 16.93	Not done	228.1 ± 64.85	472.4 ± 242.66	683.7 ± 143.62	Not done
	Blotted weight (mg, mean ± SD)	95.0 ± 16.43		191.2 ± 23.03	253.6 ± 70.91	275.6 ± 31.42	
	bw (g, mean ± SD)	295.5 ± 11.09		286.1 ± 16.83	278.1 ± 10.52	269.4 ± 3.92	
	Absolute ratio			2.01	2.67	2.90	
	bw adjusted ratio			2.03*	2.64*	2.94*	
	(Lower CL, upper CL)[a]			(1.53, 2.71)	(1.92, 3.61)	(2.05, 4.23)	

[a]Lower and upper 95% confidence limits for ratio of blotted uterine weights based on body weights as a covariable. *Level of significance, $p < 0.05$.

Table 27. Laboratory dotails for animals, diet, vehicles, and bedding.

	Strain of rat[a]		Animal diet[a]		For oral gavage		For sc injection		
Laboratory	Immature rats	OVX rats	Immature	OVX	Vohicle 1	Vehicle 2	Vehicle 1	Vehicle 2	Bedding[a]
1	Crj:CD(SD)IGS SPF/VAF	Crj:CD(SD)IGS SPF/VAF	CRF-1, Oriental Yeast	CRF-1, Oriental Yeast	Corn oil	NA	Corn oil	NA	ALPHA-dri (Pulp) for immature None for OVX
2	Crj:CD(SD)IGS	Crj:CD(SD)IGS	CRF-1, Oriental Yeast	CRF-1, Oriental Ycast	Corn oil	Ethanol	Corn oil	Ethanol	Arufa-dry for immature None for OVX
3	Crj·CD(SD)IGS	Crj:CD(SD)IGS	MF pelleted diet, Oriental Yeast	MF pelleted diet, Oriental Yeast	Sosame oil	Ethanol (99.5%)	Sesame oil	Ethanol (99.5%)	Autoclaved hardwood chips, Beta Chips, for immature and OVX
4	CRL: WI (GLX/BRL/ HAN) IGS BR	NA	Kliba rat/mouse/ hamster, Provimi	NA	Olive oil EP/DAB	Olive oil EP/DAB	NA	NA	SSNIFF (type 3/4)
5	CRL: CD (SD) IGSBR	NA	PMI certified rodent diet 5002	NA	Ethanol (95%)	Corn oil	NA	NA	ALPHA DRI for immature DACB PAPER for OVX
6	Crl CD® (SD) IGS BR	Crl CD® (SD) IGS BR	A04 C pellet maintenance diet, UAR	A04 C pellet maintenance diet, UAR	NA	NA	Corn oil	NA	Autoclaved sawdust for immature and OVX
7	Crj:CD(SD)IGS	Crj:CD(SD)IGS	CE-2, CLEA	CE-2, CLEA	Corn oil	NA	Corn oil	NA	No bedding used for immature or OVX
8	Alpk:APfSD[b]	Alpk:APfSD	R&M3 to weaning R&M1 postweaning, Special Diet Services	R&M1, Special Diet Services	Peanut oil (arachis oil)	NA	Peanut oil (arachis oil)	NA	Shredded paper for immature and OVX
9	Crj: CD (SD) IGS	Crj: CD (SD) IGS	MF pelleted diet, Oriental yeast	MF pelleted diet, Oriental yeast	Olive oil	Ethanol (99.5%)	Olive oil	Ethanol (99.5%)	Sunflake for immature and OVX
11	Wistar (BrlHan: WIST@Jcl)	Wistar (BrlHan: WIST@Jcl)	CE-2, CLEA	CE-2, CLEA	Corn oil	NA	Corn oil	NA	Sunflake for immature None for OVX
12	Crl:CD®(SD) IGS BR	Crl:CD®(SD) IGS BR	PMI certified rodent diet 5002	PMI certified rodent diet 5002	Corn oil	NA	Corn oil	NA	Ground corncobs "Bed-O'Cobs" for immature and OVX
13	SPF-bred Wistar, HSD/Cph: WU	NA	Altromin 1324, Altromin	NA	Corn oil plus min. ethanol	NA	Corn oil plus min. ethanol	NA	Low-dust wood granules Type BK 8/15
14	SD ICO:OFA SD (IOPS Caw)	NA	Pellet AO4C 10, UAR	NA	Corn oil	NA	Corn oil	NA	UAR
15	Wistar Hsd Cpb:WU	NA	R&M3, Special Diet Services	NA	NA	NA	Corn oil plus min. ethanol	NA	No bedding used
18	Sprague-Dawley	NA	PMI certified rodent diet 5014	NA	NA	NA	Corn oil plus min. ethanol	NA	Elm tree (autoclaved)
20	Hsd: Sprague Dawley	NA	Altromin MT, Altromin	NA	NA	NA	Corn oil	10% ethanol	Nesting material
21	Crl:CD(SD)BR	NA	GLP4RF25 top certificate, Mucedola	NA	NA	NA	Corn oil	NA	Dust-free poplar/fir wood chips, heat processed

Abbreviations: min., minimal; NA, not applicable.
[a]Detailed information is available from the corresponding author of this article. [b]Wistar-derived strain that is Sprague-Dawley (SD) fostered.

Table 28. Data for laboratories that did not observe a statistically significant increase in utorine weights with treatment.

			Doses					
Laboratory	Substance	Protocol	Dose 1	Dose 2	Dose 3	Dose 4	Dose 5	Comments
6	NP	B	5 mg/kg/day[a] Not done	15 mg/kg/day 0.84 (0.62, 1.13)	35 mg/kg/day 1.03 (0.76, 1.40)	80 mg/kg/day 1.24 (0.91, 1.69)	100 mg/kg/day Not done	Five labs performing tho NP dose response reached statistical significance only at doso 4 and a sixth only at dose 5 (100 mg/kg/day).
6	NP	C	5 mg/kg/day Not done	15 mg/kg/day 1.02 (0.79, 1.30)	35 mg/kg/day 1.14 (0.89, 1.46)	80 mg/kg/day 1.16 (0.90, 1.48)	100 mg/kg/day Not done	One of five labs performing the NP dose response reached statistical significance only at dose 4 and two others only at dose 5 (100 mg/kg/day).
12	DDT	C	5 mg/kg/day Not done	25 mg/kg/day 1.07 (0.79, 1.45)	50 mg/kg/day 1.10 (0.81, 1.49)	100 mg/kg/day 1.31 (0.96, 1.78)	200 mg/kg/day Not done	One of the other two labs performing the DDT dose response did not reach statistical significance until dose 5 (200 mg/kg/day).
12	BPA	Satellite C by po	60 mg/kg/day Not done	200 mg/kg/day 1.16 (0.86, 1.56)	375 mg/kg/day 1.27 (0.97, 1.68)	600 mg/kg/day 1.29 (0.94, 1.75)	1000 mg/kg/day Not done	With immature animals, one of four labs performing the BPA dose response reached statistical significance only at dose 4 and a second only at doso 5 (1,000 mg/kg/day).
20	NP	B	5 mg/kg/day 0.68 (0.46, 1.00)	15 mg/kg/day 0.62 (0.42, 0.91)	35 mg/kg/day 0.68 (0 46, 1.01)	80 mg/kg/day 0.75 (0.51, 1.11)	100 mg/kg/day 0.71 (0.48, 1.05)	Five laboratories performing NP dose response reached statistical significance only at dose 4 (80 mg/kg/day) and a sixth only at dose 5 (100 mg/kg/day). In this case, the mean vehicle blotted uterine weights were greater than the means of all treated groups, as can be seen in the columns.

[a]First row: treatment dose of the given chemical. Second row: mean relative increase in blotted uterine weight of treatment versus vehicle controls. Third row: (lower, upper 95% confidence intervals).

statistical significance, with a lower 95% confidence interval of 0.90. Two other protocol C NP studies out of five had not achieved statistical significance at 80 mg NP/kg/day, but both of these laboratories then did so at the highest dose in the series of 100 NP mg/kg/day. The maximum inductions at the 80-mg NP/kg/day dose in other protocol C NP studies had relatively low ratio values of 1.2 to 1.65. Again, the control uterine weights in laboratory 6 were higher than average, potentially reducing responsiveness.

In the third case involving DDT in protocol C, laboratory 12 was close to achieving statistical significance at 100 mg DDT/kg/day, with a lower 95% confidence interval of 0.96 (Table 24). The maximum induction in uterine size at this dose in other DDT studies using sc administration was low: 1.3–1.5 (Tables 23, 24, and 25). One other study using OVX animals had not achieved statistical significance at this dose, but did so at the highest dose in the series of 200 mg DDT/kg/day.

In the fourth case involving BPA, the satellite oral gavage study with OVX animals approached statistical significance at the dose of 600 mg BPA/kg/day, with the lower 95% confidence interval value of 0.94 (Table 6). For comparison, one of four protocol A BPA laboratories required the highest dose of 1,000 mg BPA/kg/day to achieve statistical significance (Table 2).

In retrospect, the several cases lacking or having borderline significance appear to be one of the doses selected and not one of the performance capabilities of the uterotrophic bioassay. For example, the doses selected apparently were too low and were too narrowly spaced for both the immature and the OVX versions in the case of the sc doses of NP and DDT (see Figures 4 and 5, respectively). Important conclusions from these observations are that range-finding studies should be considered when working with unknown test substances, and that a more widely interspersed set of doses could be used, for example, spacing at 0.5-log intervals, as was done with EE in phase 1, so that five doses would cover two orders of magnitude. In addition, the range-finding studies may be useful in avoiding conditions that exceed the maximum tolerated dose.

One final case deserves examination. There was a second instance in protocol B where NP failed to achieve statistical significance. First, in this case, the control blotted uterine weight mean in laboratory 20 was 54.3 mg, which again would be expected to diminish responsiveness. Analysis of this laboratory's diet showed that the phytoestrogen content measured by combined GN, daidzein, and coumestrol was greater than 500 μg/g diet (Owens et al. 2003). Second, the blotted uterine weight means of all NP test-substance dose groups were less than the controls, ranging from 37.1 mg to 41.3 mg, even up to the highest NP doses. Third, this was the laboratory with the mirror opposite BPA dose response showing statistical significance at the two lowest BPA doses and even statistically significant decreases at the two highest doses, i.e., the upper 95% confidence levels were < 1. Fourth, this laboratory had not participated in phase 1 or previously demonstrated its proficiency to conduct the protocol with a set of reference EE doses. It must, therefore, be concluded that the results in this laboratory for BPA and NP are an anomaly and cannot be attributed to the inherent performance capabilities of the specific protocol or the uterotrophic bioassay in general.

In these studies, the uterotrophic results from protocol A using oral gavage appear to relevant and conservatively predictive when compared with available chronic data from dietary studies. NP was negative at 15 mg/kg/day in all laboratories and positive at 75 mg/kg/day (Table 17). This compares favorably with LOEL observations of about 35 mg/kg/day in two multigeneration studies (Chapin et al. 1999; Nagao et al. 2001). BPA was negative at 200 mg/kg/day in all laboratories and was positive over a range of 375–1000 mg/kg/day. This compares favorably with the absence of estrogen-mediated effects at doses up to 500 mg/kg/day, as well in the controversial low–dose range in two multigeneration studies (Ema et al. 2001; Tyl et al. 2002). The GN and MX uterotrophic results in protocol A are also consistent when compared with available developmental and other studies (Casanova et al. 1999; Chapin et al. 1997; Newbold et al. 2001).

In conclusion, both the intact immature and the OVX uterotrophic versions of the uterotrophic bioassay and all protocols appear robust, reproducible, and transferable across laboratories. They are able to detect weak estrogen agonists where sufficient doses are administered and control uterine weights are sufficiently low to provide responsiveness. These results will be submitted along with other data for independent peer review to provide support for the validation of the uterotrophic bioassay. These results will also be used to develop a draft OECD test guideline for the uterotrophic bioassay. Subsequently, the guideline will be available for acceptance and implementation by regulatory authorities.

We acknowledge the dedicated efforts and work of the participating laboratories in generating these data, namely, the Chemical Evaluation and Research Institute, Japan; the Food Drug Safety Centre, Japan; the Institute of Environmental Toxicology, Japan; Mitsubishi Chemical Safety Institute, Japan; Japan Bioassay Research Centre, Japan; Sumitomo Chemical Company, Japan; Syngenta Toxicology Laboratory, United Kingdom; TNO, the Netherlands; WIL Research Laboratories, United States; BASF, Germany; Bayer AG, Germany; CIT, France; Aventis Crop Science, France; Exxon Biomedical Sciences Inc., United States; National Institute of Toxicological Research, Korea; Research Toxicology Center, Italy; and Institute of Biomedical Research, Italy. The acquisition of the chemical substances and the chemical repository for this assay validation were 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 J. Haseman for guidance and assistance with the design and conduct of statistical analysis, and both 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.

Branham WS, Dial SL, Moland CL, Hass BS, Blair RM, Fang H, et al. 2002. Phytoestrogen and mycoestrogen binding to rat uterine estrogen receptor. J Nutr 132:658–664.

Bulger WH, Muccitelli RM, Kupfer D. 1978. Studies on the *in vivo* and *in vitro* estrogenic activities of methoxychlor and its metabolites. Role of hepatic mono-oxygenase in methoxychlor activation. Biochem Pharmacol 27:2417–2423.

Casanova M, You L, Gaido KW, Archibeque-Engle S, Janszen DB, d'A Heck H. 1999. Developmental effects of dietary phytoestrogens in Sprague-Dawley rats and interactions of genistein and daidzein with rat estrogen receptors α and β *in vitro*. Toxicol Sci 51:236–244.

Chang HC, Churchwell MI, Delclos KB, Newbold RR, Doerge DR. 2000. Mass spectrometric determination of genistein tissue distribution in diet-exposed Sprague-Dawley rats. J Nutr 130:1963–1970.

Chapin RE, Delaney J, Wang Y, Lanning L, Davis B, Collins B, et al. 1999. The effects of 4-nonylphenol in rats: a multigeneration reproduction study. Toxicol Sci 52:80–91.

Chapin RE, Harris MW, Davis BJ, Ward SM, Wilson RE, Mauney MA, et al. 1997. The effects of perinatal/juvenile methoxychlor exposure on adult rat nervous, immune, and reproductive system function. Fundam Appl Toxicol 40:138–157.

Coldham NG, Sauer MJ. 2000. Pharmacokinetics of [^{14}C]-genistein in the rat: gender-related differences, potential mechanisms of biological action, and implications for human health. Toxicol Appl Pharmacol 164:206–215.

Ema M, Fuji S, Furukawa M, Kiguchi M, Ikka T, Harazona A. 2001. Rat two-generation reproduction study of bisphenol A. Reprod Toxicol 15:505–523.

Fennell TR, MacNeela JP, Manaugh CA. 1998. Pharmacokinetics of *p*-nonylphenol in male and female rats [Abstract]. Toxicologist 42:213.

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 1. 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: coded single-dose studies. Environ Health Perspect 111:1550–1558.

Miyakoda H, Tabata M, Onodera S, Takeda K. 2000. Comparison of conjugative activity, conversion of bisphenol A to bisphenol glucuronide, in fetal and mature male rat. J Health Sci 46.269–274.

Müller S, Schmid P. Schlatter C. 1998. Pharmacokinetic behavior of 4-nonylphenol in humans. Environ Toxicol Pharmacol 5:257–265.

Newbold RR, Banks PD, Bullock B, Jefferson WN. 2001. Uterine adenocarcinoma in mice treated neonatally with genistein. Cancer Res 61:4325–4328.

Nagao T, Wada K, Marumo H, Ysohimura S, Ono H. 2001. Reproductive effects of nonylphenol in rats after gavage administration: a two-generation study. Reprod Toxicol 15:293–315.

OECD (Organisation for Economic Co-operation and Development). 1998a. Report of the First Meeting of the OECD Endocrine Disrupter Testing and Assessment (EDTA) Working Group, 10 11 March 1998. ENV/MC/CHEM/RA(98)5. Paris:OECD.

——. 1998b. 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/MONO (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.' Environ Health Perspect 111:1559–1567.

Pottenger LH, Domoradzki JY, Markham DA, Hansen SC, Cagen SZ, Waechter JM Jr. 2000. The relative bioavailability and metabolism of bisphenol A in rats is dependent upon the route of administration. Toxicol Sci 54:3–18.

Tyl RW, Myers CB, Marr MC, Thomas BF, Keimowitz AR, Brine DR, et al. 2002. Three-generation reproductive toxicity evaluation of bisphenol A (BPA) in CD (Sprague-Dawley) rats. Toxicol Sci 68:121–146.

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 Owens[6]

[1]National Institute of Health Sciences, Tokyo, Japan; [2]Environmental Health and Safety Division, Organisation for Economic Co-operation and Development, Paris, France; [3]National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA; [4]Syngenta Central Toxicology Laboratory, Macclesfield, Cheshire, United Kingdom; [5]BASF Aktiengesellschaft, Ludwigshafen, Germany; [6]The 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□-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□-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 Koëter, 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 decribed. 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 [^{3}H]17[]-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 µg/kg bw/day for oral gavage and 0.3 and 1 µg/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.

	Protocol							
	A		B		C		D	
	1 µg/kg/day		0.3 µg/kg/day		0.3 µg/kg/day		0.3 µg/kg/day	
Lab	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)[c,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. [a]Results from studies with coded or blinded doses for each substance. [b]Results from the dose–response studies reported in the accompanying paper (Kanno et al. 2003). [c]Ratio 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). [d]This study did not achieve statistical significance. [e]This 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. [f]This 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.

	Protocol							
	A		B		C		D	
	3 µg/kg/day		1 µg/kg/day		1 µg/kg/day		1 µg/kg/day	
Lab	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. [a]Ratio 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). [b]This 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. [c]This 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. [d]This laboratory incorrectly diluted the high EE dose in all studies. [e]This 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.

	Protocol							
	A		B		C		D	
	600 mg/kg/day		300 mg/kg/day		300 mg/kg/day		300 mg/kg/day	
Lab	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. [a]Ratio 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). [b]This study did not achieve statistical significance. [c]In 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. [d]In 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.

	Protocol							
	A		B		C		D	
	300 mg/kg/day		35 mg/kg/day		35 mg/kg/day		35 mg/kg/day	
Lab	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)*[b]	—	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)*[b]	—	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. [a]Ratio 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). [b]In the coded single-dose studies at this dose, one animal died in laboratory 2 and one in laboratory 14. *Level of significance, $p < 0.05$.

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

	Protocol							
	A		B		C		D	
	300 mg/kg/day		500 mg/kg/day		500 mg/kg/day		500 mg/kg/day	
Lab	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)*[b]	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)*[c]	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. [a]Ratio 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). [b]In the dose–response studies at this dose, one animal died in laboratory 3. [c]In 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.

	Protocol							
	A		B		C		D	
	250 mg/kg/day		80 mg/kg/day		80 mg/kg/day		80 mg/kg/day	
Lab	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)*[b]	—	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)*[b]	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)*[b]	—	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)*[b]	—	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)*[b]	—	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,c]	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)*[b]	—	1.08 (0.79, 1.48)[c]	—	—	—	—	—
14	2.05 (1.17, 3.59)*[b]	—	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. [a]Ratio 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). [b]In 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. [c]This study did not achieve statistical significance. [d]In 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.

	Protocol							
	A		B		C		D	
	300 mg/kg/day		100 mg/kg/day		100 mg/kg/day		100 mg/kg/day	
Lab	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)*[c]	—	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)*[e]	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)*[c]	—	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. [a]Ratio 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). [b]This study did not achieve statistical significance. [c]In the coded single-dose studies, all six animals died in laboratory 12, and two animals died in laboratory 14. [d]With the lower confidence level number > 1.00, the result is statistically significant. [e]In 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.

	Protocol			
	A	B	C	D
Lab	1000 mg/kg/day	500 mg/kg/day	500 mg/kg/day	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. [a]Ratio 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). [b]This 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]

	Strain of rat[b]		Animal diet[b]		For oral gavage[b]		For sc injection[b]		
Lab	Immature rats	OVX rats	Immature	OVX	Vehicle 1	Vehicle 2	Vehicle 1	Vehicle 2	Bedding
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

[a]The 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). [b]Detailed 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).

Table 10. Global analysis of results.

	Protocol			
Substance/dose	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.

[a]Ratio 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).

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 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 reslts 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 1. 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.

ENVIRONMETRICS
Environmetrics 2003; **14**: 213–222 (DOI: 10.1002/env.578)

A statistical method for judging synergism: application to an endocrine disruptor animal experiment

Nobuhito Matsunaga[1], Jun Kanno[2] and Isao Yoshimura[3,*,†]

[1] *Kyowa Hakko Kogyo Co., Ltd 1-6-1, Ohtemachi, Chiyoda-ku, Tokyo 1008185, Japan*
[2] *National Institute of Health Sciences 1-18-1, Kamiyoga, Setagaya-ku, Tokyo 1588501, Japan*
[3] *Tokyo University of Science, 1-3 Kagurazaka, Shinjuku-ku, Tokyo 1628601, Japan*

SUMMARY

This article proposes a statistical method for judging whether or not the combined action of chemicals is synergistic, being focused on the case in which two or more endocrine disruptors are made to act simultaneously. After defining synergism, the synergistic relation of two chemicals is formulated for a higher response than that expected under an exchangeable relation between them. Using this formulation as a basis, we then rationalize the triangular design for an animal experiment in which all dose settings are controlled within a triangle domain that prescribes the sum of doses of simultaneously applied chemicals less than a certain level. In addition, a statistical test is proposed for judging the synergism among chemicals used in animal experiments, i.e. the test evaluates the discrepancy between the observed mean response from simultaneous administration groups of chemicals and an estimated response under the null hypothesis of zero interaction based on data from single administration groups. Finally, test performance is examined using a simulation study and a case study—the rodent uterotrophic assay. The simulation study revealed that the test is not superior in power to the standard analysis of variance test based on a linear model with interaction term, yet robust in the sense that type I errors under variance heterogeneity were better controlled using Welch correction than the analysis of variance test. The application of the proposed statistical test to an animal experiment is considered acceptable based on results. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: animal experiment; endocrine disruptor; design of experiment; statistical test; synergism

1. INTRODUCTION

To protect people from the harmful effects of chemicals, society has begun regulating environmental pollutants and toxicants at levels having negligible impact. In the past these regulation levels were determined based on the knowledge or toxicity data of a single administration of an individual chemical. Recently, however, synergic effects due to combining chemicals have become apparent and regulations are now considered to be based on the knowledge or data on their combined action. Accordingly, many studies have been carried out to clarify synergism of harmful effects by

*Correspondence to: Isao Yoshimura, Faculty of Engineering, Tokyo University of Science, 1-3 Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan.
†E-mail: isao@ms.kagu.tus.ac.jp

Contract/grant sponsor: Japan Society for the Promotion of Science; Contract/grant number: (c)11680328.

Received 23 December 2001
Accepted 22 March 2002

simultaneous administration of chemicals (see, for example, Reif, 1984; Hasegawa *et al.*, 1996). One particular application is that for the synergism of endocrine disruptors.

As one of the authors has been engaged in endocrine disruptor studies underway in Japan (Kanno *et al.*, 2001), it was necessary to determine how to obtain and analyze data from animal experiments concerned with synergism. Under this requirement, here we investigate data collection/analysis allowing evaluation of synergism, applying the devised method to an animal experimental study conducted in Japan.

Section 2 explains the issues elicited in the above-mentioned study, while Section 3 discusses the concept of synergism adopted in the analysis. Sections 4 and 5 subsequently describe the experiment design and statistical test used for analysis of the endocrine disruptor study, after which Sections 6 and 7 respectively present the results of the simulation study, which examines the performance of the proposed test, and a case study. Section 8 provides a conclusion and discussions.

2. ENDOCRINE DISRUPTOR ISSUE

Chemicals that induce a hormonal effect are referred to as hormonally active agents (HAAs)—see, for example, Committee report (1999), EDSTAC (2001) and Solicitation (2001). Endocrine disrupting chemicals (EDCs) are defined as HAAs that induce adverse effects. As most hormonal effects are well known to be mediated by hormone receptors, endocrine disruption can therefore be defined as a 'receptor-mediated adverse effect or toxicity'.

A question arises concerning what are the major differences between traditional toxicity and receptor-mediated toxicity, especially that occurring through nuclear receptors such as estrogen and androgen receptors, or through ligand-inducible transcription factors such as dioxin receptors. It must be realized that the effects are mediated by the 'signal', and that the 'toxicants' do not need to be at the site of the adverse effect. In addition, with regard to the 'redundancy' of the receptor system, such receptors bind a variety of chemicals having various structures. Naturally, then, the affinity is different among chemicals and usually much lower than that of intrinsic natural hormones such as estradiol (see, for example, Yamasaki *et al.*, 2002). However, binding does occur, and if the concentration of the ligand goes above a certain level, then it usually has the capability to transduce the signal just as natural hormones do.

Since the signal transduction system basically amplifies the signal, it is believed that this occurs at a lower dose range than that exhibited in traditional toxicity studies. Expansion of this aspect may indicate that a system exists in which there is no threshold in response. Another aspect of redundancy is that each particular chemical can change the conformation of the ligand-bound receptor molecule according to the shape of each ligand molecule. If true, this may lead to different signaling properties especially when considering interactions with DNA and/or co-factor molecules.

Ligand-bound receptor molecules need to bind to a specific DNA sequence and recruit co-factors and other transcription machinery molecules in order to induce actual biological effects. In this context, the combined effect of multiple chemicals can be slightly different from what we expect from the monitored effect due to a single chemical.

Moreover, because more than one signaling system is present in humans, and because many other nuclear receptors/transcription factors are redundant in such ways, there may be an interaction between different signal pathways which leads to possible synergism for certain biological endpoints. Therefore, the definition currently needed for the expected combined effect is that if two treatments produce the same endpoint, they can be exchanged by any ratio to produce the same magnitude of the effect. The definition of an unexpected combined effect is that the effect due to such a combination is much larger than the particular effect induced by each treatment alone.

3. DEFINITION OF SYNERGISM

There are numerous discussions on the definition of additivity, synergism, and antagonism (see, for example, Rothman, 1980; Saracct, 1980; Reif, 1984; Berenbaum, 1989; Kodell and Pounds, 1990; Machado and Robinson, 1994; Laska *et al.*, 1997; Gennings *et al.*, 1997; Roy and Estieve, 1998).

From the 1920s to the 1960s, pharmacologists attempted to classify mechanisms representing the mode of combined action of two chemicals, which is the case considered here. Such trials subsequently generated numerous technical terms such as 'independent joint action,' 'similar joint action,' 'synergistic action,' 'dissimilar joint action,' 'potentiation,' 'depotentiation', and 'augmenter.' Due to the complexity of the concepts and difficulties in actual verification, such mechanistic analyses had virtually ended until a simple definition was introduced (Sakuma, 1996).

It is illustrated in a pharmacology textbook (see, for example, Laurence and Bennett, 1980) as a chart representing a 'Mountain of Happiness', which is an isobolic expression of happiness given after drinking a certain amount of wine followed by coffee. On this chart, a combination of a certain amount of wine and coffee realizes the apex of the response, which cannot be expected by the single administration of wine or coffee, while an excessive administration ends to dullness or sleep. It implies that the pharmacologically useful endpoint is to determine the best combination of two treatments (wine and coffee) regardless of mechanistic considerations. Synergism can be used to express such a peak in an isobologram, which also indicates that too much wine and/or coffee reduces happiness.

In general, toxicologic events are also complex, multi-step phenomena that are not fully understood; hence, it is reasonable to surmise that mechanistic considerations are not established for predicting the combined adverse effect of two chemicals. The definition of synergism regarding hazard identification must therefore be based on a non-mechanistic approach analogous to the Mountain of Happiness, although we are obviously not interested in the best combination of two chemicals that produce the strongest adverse effect. Our interest under the above-mentioned situation concerns the low dosage range in which two chemicals show combined adverse effects at a higher magnitude than that expected when two chemicals are equal in a particular response, i.e. they are exchangeable by any ratio. This viewpoint leads the following formulation adopted here.

Let $\mathrm{f}(d_\mathrm{A}, d_\mathrm{B})$ be the response at the combined dosage $(d_\mathrm{A}, d_\mathrm{B})$ of two chemicals A and B, and D_A and D_B be such that $\mathrm{f}(D_\mathrm{A}, 0) = \mathrm{f}(0, D_\mathrm{B})$ under the assumption that f is a continuously monotone increasing function of either coordinate. If chemicals A and B are exchangeable, then $\mathrm{f}(d_\mathrm{A}, d_\mathrm{B}) = \mathrm{f}(D_\mathrm{A}, 0) = \mathrm{f}(0, D_\mathrm{B})$ is expected for $(d_\mathrm{A}, d_\mathrm{B})$ on the line connecting $(D_\mathrm{A}, 0)$ and $(0, D_\mathrm{B})$. Accordingly, we define the response of the two chemicals to be 'synergistic' if $\mathrm{f}(d_\mathrm{A}, d_\mathrm{B}) > \mathrm{f}(D_\mathrm{A}, 0) = \mathrm{f}(0, D_\mathrm{B})$ for $(d_\mathrm{A}, d_\mathrm{B}), (D_\mathrm{A}, D_\mathrm{B})$ such that

$$\frac{d_\mathrm{A}}{D_\mathrm{A}} + \frac{d_\mathrm{B}}{D_\mathrm{B}} = 1 \qquad (1)$$

The case where the equality $\mathrm{f}(d_\mathrm{A}, d_\mathrm{B}) = \mathrm{f}(D_\mathrm{A}, 0) = \mathrm{f}(0, D_\mathrm{B})$ holds, implies 'zero interaction'.

The combined action of two chemicals considered here is, within a certain dose range, the same as the simple similar action for quantal response discussed by Hewlett and Plackett (1959) (see also Piegorsch and Bailer (1997) for summarized explanation), but is slightly different in the sense that it is formulated through an isobolic relation. This formulation is meaningful for proposing a triangular design, for we need not worry about the combined action of simultaneous administration of chemicals in the dose which is the maximum in the groups with individual chemical administration, while the formulation by Hewlett and Placket was too strict to apply to toxicity evaluation.

4. DESIGN OF EXPERIMENT

From a statistical viewpoint, synergism is examined experimentally using one-sided hypothesis testing for the null hypothesis of zero interaction. Note that a linear model can, without loss of generality, be assumed to express the dose–response relationship under the above-mentioned situation, i.e. in the exchangeable case.

This is true because the dose dependency of the response to both chemicals can be linearized by a suitable scale adjustment and a certain transformation of response, i.e. by the use of a function such as a link function in a generalized linear model that makes the dose–response relationship linear so that the relation $\mathrm{f}(d_{\mathrm{A}}, d_{\mathrm{B}}) = \beta_0 + \beta_{\mathrm{A}} d_{\mathrm{A}} + \beta_{\mathrm{B}} d_{\mathrm{B}}$ holds.

While the factorial design shown in Table 1 is most often used for statistically evaluating interaction, it is not appropriate in our case, for the linearization should be confined within a certain dose range. In a two dimensional (2D) plane having coordinates that respectively indicate the dose of each chemical, responses outside the line connecting the maximum dose of the two chemicals do not provide any information on the synergism, so that the dose settings outside this triangle domain are useless for evaluating synergism.

In fact, even when the response for Groups (10), (12), (13), (14), (15), and (16) in Table 1 is quite high, it cannot be used to evaluate synergism because corresponding zero interaction response to be compared with them cannot be estimated. Consequently, we propose to use the triangular design, which eliminates the above-mentioned groups as shown in Table 2, for an animal experiment under the condition that the number of doses given by the administration of individual chemical is the same between the two chemicals. The number of simultaneous administration groups, which is 3 in Table 2, may well be dependent on the purpose of the experiment, but this is not our principal concern here.

5. STATISTICAL METHOD

The one-sided statistical test for evaluating the discrepancy between the observed response and the response estimated under the null hypothesis of zero interaction is considered reasonable as the statistical method for data analysis.

Table 1. An example of factorial design with 4 dose levels of each chemical. Animals are randomly allocated to each of 16 groups. Groups (1) through (7) correspond to single administration groups, whereas Groups (8) through (16) represent simultaneous administration groups

		Dose of chemical A			
		d_{A1}	d_{A2}	d_{A3}	d_{A4}
Dose of chemical B	d_{B1}	(1)	(2)	(3)	(4)
	d_{B2}	(5)	(8)	(9)	(10)
	d_{B3}	(6)	(11)	(12)	(13)
	d_{B4}	(7)	(14)	(15)	(16)

Table 2. An example of triangular design with 4 dose levels of each chemical. Animals are randomly allocated to each of 10 groups. Groups (1) through (7) correspond to single administration groups, whereas Groups (8) through (10) represent simultaneous administration groups

		Dose of chemical A			
		d_{A1}	d_{A2}	d_{A3}	d_{A4}
Dose of chemical B	d_{B1}	(1)	(2)	(3)	(4)
	d_{B2}	(5)	(8)	(9)	
	d_{B3}	(6)	(10)		
	d_{B4}	(7)			

With the endocrine disruptor issue in mind, we assume that the observed variable y_{ij} of jth individual of ith group is distributed as normal with mean μ_i and variance σ_i^2 and that the ys are independent. Let $\bar{y}$ be the observed mean response for simultaneous administration groups (Groups (8), (9), and (10) in the case of Table 2) of two chemicals and $\hat{y}$ be the estimated response corresponding to $\bar{y}$ using data for groups (Groups (1) through (7) in the case of Table 2) with the administration of individual chemicals under the assumption of zero interaction. Naturally, $\bar{y}$ and $\hat{y}$ are statistically independent.

We propose using the following test statistic:

$$T = \frac{\bar{y} - \hat{y}}{\sqrt{\mathrm{Var}(\bar{y}) + \mathrm{Var}(\hat{y})}} \tag{2}$$

where $\mathrm{Var}(\bar{y})$ and $\mathrm{Var}(\hat{y})$ are the estimated variances of $\bar{y}$ and $\hat{y}$, respectively.

If we assume that all σs are equal, the denominator of the statistic T should be pooled within variance, with the degrees of freedom ν being equal to 'the total number of observations – the number of groups' and the critical value with significance level α is the upper 100α percentage point, $t(\nu, \alpha)$, of a t-distribution with degrees of freedom ν. Else if we assume that σs are homogeneous within simultaneous administration groups or groups with individual chemical administration, but heterogeneous between two classes, the two terms in the denominator of T should be separately estimated as the within-class sum of squares divided by 'the total number of observations of the class—the number of the groups in the class'. In the latter case, the critical value is set at $t(\nu, \alpha)$ with the degrees of freedom ν adjusted by Welch correction (see Welch, 1938, or Satterthwaite, 1946).

In the real situation of toxicity experiments, the variances are likely to be heterogeneous and even the latter assumption may be violated. However, since the heterogeneity of variances cannot be exactly estimated, we propose to use the latter test (Proposed-W) as the statistical method for judging synergism, or the former test (Proposed-T) when the homogeneity of variance is confirmed, the performance of these tests being compared with a regression test in the next section.

Thus, the flow of the proposed method is as follows:

Step 0. Check the linearity of the dose–response relationship for the groups with individual chemical administration. If a non-linear dose–response relationship is observed, transformations that linearize the relation are applied.

Step 1. Fit a linear response plane, i.e.

$$y = \beta_0 + \beta_A d_A + \beta_B d_B \tag{3}$$

to the groups with single chemical administration using a least squares method (assuming zero interaction).

Step 2. Calculate the mean response $\bar{y}$ for simultaneous administration groups and the estimate $\hat{y}$ of expected response corresponding to $\bar{y}$ under zero interaction.

Step 3. Calculate the test statistic T and the critical value $t(\nu, \alpha)$ with the adjusted degrees of freedom ν using the Welch correction, where α is the significance level.

Step 4. If $T > t(\nu, \alpha)$, then the relationship is judged as synergistic with significance level α.

6. SIMULATION STUDY

A simulation study was performed to evaluate the performance of the proposed tests.

6.1. Common setup

Let y_{ij} be the response variable obtained from the jth animal of the ith group, where the number of groups with individual chemical administration is seven, being the same as the triangular design shown in Table 2, whereas that of the simultaneous administration groups is one, two, and three for Cases 1, 2, and 3, respectively. The total number k of groups is therefore eight, nine, or ten, depending on the case. The number of animals was fixed at six to coincide with the number used in the endocrine disruptor experiment for the case study in the next section.

It is assumed that y_{ij}, $i = 1, 2, \ldots, k, j = 1, 2, \ldots, 6$, were distributed independently as normal with mean μ_i and variance σ_i^2 and that the dose–response relationship was linear when each chemical was singly administered. As an alternative to the proposed test, we considered an analysis of variance test for interaction in a regression model with interaction, i.e. the null hypothesis was $H_0 : \beta_{AB} = 0$ for the following model:

$$E\{y_{ij}\} = \beta_0 + \beta_A d_A + \beta_B d_B + \beta_{AB} d_A d_B \tag{4}$$

where d_A, d_B are the doses of chemicals A and B, respectively, administered to the ith group. Robustness was examined by comparing the proposed t-tests with Welch correction (Proposed-W test) and without Welch correction (Proposed-T test) with the analysis of variance test (Regression test). Other common simulation conditions were as follows:

- repetition of simulation, 10 000 times
- dose setting for singly administered groups,

$$(d_A, d_B) = (0,0)\ (0,1)\ (0,2)\ (0,3)\ (1,0)\ (2,0) \text{ or } (3,0).$$

- parameter values: $\beta_0 = \beta_A = \beta_B = 1$
- nominal significance level: 5 per cent.

6.2. Alternative hypothesis

Three cases of simultaneous administration groups were considered, i.e.:

Case 1: One group with $(d_A, d_B) = (1.0,\ 1.0)$.
Case 2: Two groups with $(d_A, d_B) = (1.0,\ 1.0),\ (1.5,\ 1.5)$.
Case 3: Three groups with $(d_A, d_B) = (1.0,\ 1.0),\ (1.0,\ 2.0),\ (2.0,\ 1.0)$.

Table 3. Setup of the strength of synergy in alternative hypothesis. Δ_i represents the strength of synergy on the ith group with simultaneous administration of two chemicals. Model (1) corresponds to the null hypothesis, whereas models (2) through (9) represent alternative hypothesis

		Model								
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Case 1	Δ_1	0.0	0.3	0.5	1.0	1.5	2.0			
Case 2	Δ_1	0.0	0.3	0.5	1.0	0.3	0.5	1.0	0.3	0.5
	Δ_2	0.0	0.45	0.75	1.5	0.675	1.125	2.25	0.9	1.5
Case 3	Δ_1	0.0	0.3	0.5	1.0	0.3	0.5	1.0	0.3	0.5
	Δ_2	0.0	0.45	0.75	1.5	0.6	1.0	2.0	0.9	1.5
	Δ_3	0.0	0.45	0.75	1.5	0.6	1.0	2.0	0.9	1.5

The number of simultaneous administration groups is therefore different, depending on the case, and $\bar{y}$ is the mean of the observed responses of 1, 2, or 3 groups, depending on Cases 1, 2, or 3, respectively. The strength of the synergism is represented by the parameters Δ_1, Δ_2, and Δ_3, which are defined as the difference between the expected value of y_{ij} of the simultaneous administration groups and the one under the null hypothesis, i.e. Equation (3). If we adopt Equation (4) as an alternative model such as models (5), (6) and (7) in Table 3, then $\Delta_i = \beta_{\mathrm{AB}} d_{\mathrm{A}} d_{\mathrm{B}}$, where $d_{\mathrm{A}}, d_{\mathrm{B}}$ are the doses of A and B of the ith group, respectively.

Δs in Table 3 were selected as the simulation setting. As all Δs are obviously zero for model (1), this implies the null hypothesis. For models (2)–(4) the Δs are proportional to $d_{\mathrm{A}} + d_{\mathrm{B}}$, while for models (5)–(7) the Δs are proportional to $d_{\mathrm{A}} d_{\mathrm{B}}$, being advantageous for the analysis of variance test. Models (8) and (9) use steeper Δs.

6.3. *Power under variance homogeneity*

Table 4 summarizes the results of the simulation, where all the σ_i^2s are the same. It is theoretically natural that powers in Case 1 are the same between the two tests. In other settings, it is noted that the proposed tests are slightly inferior to the analysis of variance test in power under variance homogeneity.

Table 4. Probability (%) to realize significance. Type I error in model (1) and powers in other models. 'Proposed-W' and 'Proposed-T' are proposed tests with or without Welch correction, respectively

		Model								
Case	Test	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Case 1	Proposed-W	5.4	9.7	17.3	49.9	82.5	96.7			
	Proposed-T	4.9	10.6	20.2	61.0	91.7	99.5			
	Regression	4.9	10.6	20.2	61.0	91.7	99.5			
Case 2	Proposed-W	5.2	18.0	41.6	92.2	26.3	60.6	99.3	37.9	77.4
	Proposed-T	4.9	18.8	44.0	93.7	27.2	63.4	99.4	39.6	80.3
	Regression	5.0	17.9	42.5	93.4	30.1	68.8	99.8	46.9	88.2
Case 3	Proposed-W	4.9	24.1	54.1	98.5	33.5	73.0	100.0	57.3	94.9
	Proposed-T	5.1	24.2	54.4	98.6	34.4	73.8	100.0	58.1	95.2
	Regression	4.8	24.0	54.5	98.5	36.1	75.8	100.0	62.8	97.0