Determination of the Di-(2-Ethylhexyl) Phthalate NOAEL for Reproductive Development in the Rat: Importance of the Retention of Extra Animals to Adulthood

Chad R. Blystone,* Grace E. Kissling,† Jack B. Bishop,* Robert E. Chapin,*¹ Gary W. Wolfe,‡² and Paul M. D. Foster*³

*National Toxicology Program and †Biostatistics Branch, National Institute of Environmental Health, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, North Carolina 27709 and ‡TherImmune Research Corporation, Gaithersburg, Maryland 20878

¹Present address: Pfizer Global Research, Groton, CT 06340.

²Present address: Aclairo Pharmaceutical Development Group Inc., Vienna, VA 22181.

³To whom correspondence should be addressed at National Toxicology Program, National Institute of Environmental Health, National Institutes of Health, Department of Health and Human Service, PO Box 12233 (MD K2-12), Research Triangle Park, NC 27709. Fax: (919)-541-4255. E-mail: Foster2@niehs.nih.gov.

Received April 19, 2010; accepted May 11, 2010

Deriving No Observed Adverse Effect Level (NOAEL) or benchmark dose is important for risk assessment and can be influenced by study design considerations. In order to define the di-(2-ethylhexyl) phthalate (DEHP) dose-response curve for reproductive malformations, we retained more offspring to adulthood to improve detection of these malformations in the reproductive assessment by continuous breeding study design. Sprague-Dawley rats were given a dietary administration of 1.5 (control), 10, 30, 100, 300, 1000, 7500, and 10,000 ppm DEHP. Male pups were evaluated for gross reproductive tract malformations (RTMs) associated with the “phthalate syndrome.” DEHP treatment had minimal effects on P0 males. There was a statistically significant increase in F1 and F2 total RTMs (testis, epididymides, seminal vesicle, and prostate) in the 7500-ppm dose group and F1 10,000-ppm dose group. The 10,000-ppm exposed F1 males did not produce an F2 generation. The NOAEL for F1 and F2 RTM combined data, because *in utero* exposures were similar, were 100 ppm (4.8 mg/kg/day), which was close to the 5% response benchmark dose lower confidence limit of 142 ppm. The utility of evaluating more pups per litter was examined by generating power curves from a Monte Carlo simulation. These curves indicate a substantial increase in detection rate when three males are evaluated per litter rather than one. A 10% effect across male pups would be detected 5% of the time if one pup per litter was evaluated, but these effects would be detected 66% of the time if three pups per litter were evaluated. Taken together, this study provides a well-defined dose response of DEHP-induced RTMs and demonstrates that retention of more adult F1 and F2 males per litter, animals that were already produced, increases the ability to detect RTMs and presumably other low-incidence phenomena.

Key Words: phthalate; endocrine disruption; RACB; dose response; reproductive tract; reproductive and development.

induced by the test article. Due to concerns about cost and animal use, studies are often limited in numbers of animals used such that calculating the power of detection based upon litter size and variation of the end point in reproduction studies is warranted. In studies that assess reproductive toxicity, which incorporate the production of rodent litters, the evaluation of postweaning end points in the offspring involves the selection of only one male and one female from each litter according to Organization for Economic Co-Operation and Development (OECD; OECD 416) and Environmental Protection Agency (EPA) guidelines (OPPTS 870.3800) for mating. The rest of the litter is removed during standardization (usually at post-natal day [PND] 4) or weaning. Histopathology is performed on animals selected for mating (OECD) or 10 randomly chosen males and females selected for mating (EPA). Because the litter is the statistical unit of analysis for these studies, the number of animals selected for evaluation in adulthood will influence the detection of adverse effects of low incidence (Hotchkiss *et al.*, 2008). In the present study, a reproductive assessment of di-(2-ethylhexyl) phthalate (DEHP) was conducted with seven dose levels and retained extra animals per litter, which would otherwise have been culled, to adulthood in order to better define the dose-response curve.

DEHP is a member of a group of phthalate esters that induce male reproductive tract malformations (RTMs) via a disturbance in androgen signaling at critical periods in *in utero* sexual differentiation (Parks *et al.*, 2000). These male RTMs include malformations of the testis, epididymis, prostate, seminal vesicles, and external genitalia with common RTMs being epididymal agenesis, undescended testes, small fluid-filled testis, hypoplastic accessory sex organs, and hypospadias. This pattern of malformations in organs that require androgen for their normal development has been described as the phthalate syndrome (Gray and Foster, 2003). These malformations were not observed

In guideline toxicology studies, the design of the study must incorporate adequate power to detect potential adverse effects

with DEHP or other phthalates in standard rat developmental toxicity (teratology) study designs (exposure from gestational day [GD] 6 to 15 in the rat) due to lack of exposure throughout the period of sexual differentiation, and offspring evaluations conducted prior to expected delivery of the offspring before the reproductive organs have fully differentiated or developed.

A reproductive assessment by continuous breeding (RACB), a multigenerational study design, was conducted to evaluate the DEHP dose response of male RTMs. The Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC) and EPA recognized the deficiency in standard teratology study design and proposed the multigeneration reproduction study as the “definitive” Tier II test to evaluate endocrine active chemicals (EACs) (EDSTAC, 1998; EPA, 1998). This Tier II test would confirm or refute data obtained in the Tier I screens and provide essential dose-response information for risk assessment. However, it has been argued that the single male and female F1 animals evaluated per litter at adulthood, when RTMs are the most apparent, in the conventional, EPA regulatory multigeneration study design unnecessarily restricts the power to detect and characterize effects produced by EACs (Foster, 2002; Foster and McIntyre, 2002; McIntyre *et al.*, 2000). Low-powered studies hinder determination of an accurate no observed adverse effect level (NOAEL) and dose-response curve for chemicals (Hotchkiss *et al.*, 2008), and poor public health decisions can result.

The present study was designed to evaluate the postnatal effects of DEHP on reproductive development and define the dose-response relationship for the induction of RTMs. This study used more than the traditional three dose groups plus control and retained a greater number of F1 and F2 male rats until adulthood to better assess the shape of the dose-response curve and define the DEHP NOAEL for male RTMs. In conducting this study under Good Laboratory Practices (GLP), similar to guideline studies, we hope to demonstrate that it is feasible to retain extra animals per litter in a complex multigenerational design. This approach did not constitute the generation of any extra animals because the animals normally produced were simply not culled at PND 4 or weaning but were maintained alongside the animals selected for breeding. Furthermore, we generated power curves to quantify detection rates based upon the effect size within the litter and the number of pups examined (one up to six per sex per litter), which further illustrates the importance of retaining extra animals to detect effects of low incidence within a litter.

MATERIALS AND METHODS

The Continuous Breeding Protocol (Chapin and Sloane, 1997), a modified multigeneration reproduction study, was used to evaluate the reproductive toxicity of DEHP. Sprague-Dawley rats (Charles River Laboratories, Portage, MI) were delivered to TherImmune Research Corporation (Gaithersburg, MD) and 7 days after receipt were weighed and randomly assigned to treatment groups. Treatment groups were fed a diet with DEHP (99.8% pure by gas chromatography; Aldrich Chemical Company, Milwaukee, WI) concentrations

of 1.5 (control), 10, 30, 100, 300, 1000, and 7500 ppm (anticipated to be equivalent to ~0.1, 0.5, 1.5, 5, 15, 50, and 400 mg DEHP/kg/day). DEHP is a ubiquitous environmental contaminant and was present at trace levels even in our control NTP-07 diet. Additional groups, 1.5 ppm (control) and 10,000 ppm (anticipated to be > 500 mg DEHP/kg/day), were added after the start of the study, and these groups followed the same design.

Exposure started with the P0 generation and was continuous through the F1 and F2 generations. P0 males and females, ~5 weeks of age, were given the DEHP diet for 6 weeks prior to mating (premating exposure), then they were cohabitated (17 pairs per group) for 9 weeks with the same diet in order to produce three litters (F1a, F1b, and F1c). The first two litters produced (F1a and F1b) during the cohabitation period were counted and weighed at PND 1. These litters were euthanized by sodium pentobarbital overdose and discarded without necropsy on PND 1. The third litter born (F1c) was reared (without culling) by the dam until weaning on PND 21. On PND 16, up to six males and two females were randomly selected from each litter to be maintained to adulthood for histological evaluations. One to two males in the litter were used to breed for the production of the F2 generation while avoiding sibling matings. The additional nonbreeding males were maintained until necropsy as sexually mature adults. The non-mated males were necropsied ~2 weeks prior to the necropsy of their sibling mated male. The same methods for producing and evaluating the F1 generation were used for producing and evaluating the F2 generation such that three litters were produced in each generation and the third litter was evaluated postweaning. The F2c litters were selected for breeding and necropsy (same as the F1c), whereas one to two males of the F3c litter were terminated at PND 63/64. The F3c males were not evaluated for RTMs.

Animals from all control litters ($n = 14$ litters for F1 and $n = 10$ for F2) and 8–17 litters per dose level were necropsied. The F1c non-mated males were necropsied on PND 194 ± 12 or PND 249 (the added 1.5-ppm and 10,000-ppm groups) and breeding males and females on PND 215 ± 12 or PND 263 (the added 1.5-ppm and 10,000-ppm groups). The F2c non-mated males were necropsied on PND 241 ± 10 and the mated males and females were necropsied on PND 248 ± 10 . During necropsy, RTMs observed by gross evaluation were recorded for testes, epididymides, prostate, and seminal vesicles.

Pregnancy and litter end points (e.g., litter size, pup weights) were compared between the 1.5-ppm (control) and dose groups. Quantitative data were analyzed using the nonparametric multiple comparisons procedure of Dunn (1964) or Shirley (1977), as modified by Williams (1986). Shirley's test is designed to detect treatment-related differences when the response to treatment consistently increased (or decreased) with increasing dose, which was verified by Jonckheere's test (Jonckheere, 1954). If the p value from Jonckheere's trend test was < 0.01, Shirley's test was used; otherwise, Dunn's test was used.

Male RTMs were recorded as present or absent for each F1c and F2c litter according to whether any of the malformations were observed in the litter. RTM incidences, with the litter as the unit of analysis, were analyzed using Fisher's exact test to determine statistically significant differences between control and each treated group. Evaluations were conducted separately on the F1c and F2c litters, as well as on the pooled litters from the F1c and F2c groups (since their *in utero* exposure period was identical). One-tailed p values < 0.05 were considered significant. Benchmark dose (BMD) and benchmark dose lower confidence limit (BMDL) were calculated using EPA's BMDs 2.1.1 software (Build 11-6-09), after evaluating the fit of the various models to the observed data. BMDs were calculated for a 5% response using a dichotomous Weibull model with extra risk of RTM incidence in the F1c, F2c, and combined F1c + F2c litters.

Power curves at $\alpha = 0.05$ were generated for Fisher's exact test via Monte Carlo simulations. Computer-generated samples of 1, 2, 3, 4, 5, or 6 pups per litter for 20 “treated” litters were compared with a “control” group having no abnormalities in 20 litters, using a one-sided Fisher's exact test. Each treated pup had the same probability of having an abnormality, with probabilities ranging between 0.01 and 0.90. A litter was positive if at least one pup had the abnormality and negative if none of the pups had the abnormality. A total of 5000 samples of 20 treated and 20 control litters were generated for each

combination of number of pups per litter and probability of an abnormality. Power was estimated as the proportion of samples for which the one-sided Fisher's exact test p value was ≤ 0.05 . Each power curve shows the probability of getting a significant result at the 0.05 level by retaining and examining the specified number of pups per litter, assuming that the chemical produces in each pup an abnormality with the probability plotted on the x-axis.

RESULTS

Based upon feed consumption of the 1.5, 10, 30, 100, 300, 1000, 7500, and 10,000 ppm DEHP diet, average consumed doses were 0.12, 0.78, 2.4, 7.9, 23, 77, 592, and 775 mg/kg/day for the P0 animals; 0.09, 0.48, 1.4, 4.9, 14, 48, 391; and 543 mg/kg/day for the F1 animals, and 0.1, 0.47, 1.4, 4.8, 14, 46, and 359 mg/kg/day for the F2 animals (there were no F2 animals at 10,000 ppm). P0 food consumption was increased and decreased at certain time points during exposure, but there was no dose- or time-related pattern (i.e., no consistent effect across time and/or dose level; data not shown). In the P0 generation, male body weights were significantly decreased by 5–6% at 10,000 ppm during weeks 21 and 23 and dam body weights at 10,000 ppm were significantly reduced at delivery of each of the three litters. In the F1 generation, body weights of males and females were significantly reduced throughout exposure in the 10,000-ppm dose group. Body weights were also reduced in 7500-ppm adult F1 males and F2 males and females. Food consumption (grams per animal in kilogram) was generally increased throughout exposure in the F1 7500 and 10,000 ppm males and F2 7500 ppm males and females.

In the mated cohorts, the pregnancy index (number of females delivering/number of cohabiting pairs) was unaffected by treatment in the P0 generation, but a clear effect was evident at the 10,000 ppm dose level in the F1 generation (Table 1). No litters were produced by the F1 10,000-ppm dose group so exposure continued only for the 1.5-, 10-, 30-, 100-, 300-,

1000-, and 7500-ppm dose groups. In the production of the F2 generation, the pregnancy index of F1 dams was significantly decreased at the 7500 ppm dose. The litter size (number of live pups) produced by the P0 generation was significantly reduced in the 7500-ppm dose group (Table 2), but this effect was not consistent across generations or dose dependent. Of the three litters produced by each generation, average pup weights on PND1 (combined male and female) were generally unaffected by treatment below 7500 ppm. Body weights at PND 1 were reduced in the F1 pups at 10,000 ppm and in the F2 pups at 7500 ppm (Table 2).

Two P0 10,000 ppm males had a mild testicular lesion with no consequent effects on fertility in this group and one 7500 ppm animal had a prostate lesion. In stark contrast, all the F1 10,000 ppm males had an RTM (Table 3), and no F2 litters were produced (Table 2). At 7500 ppm, there were significant increases compared with controls in the incidence of RTMs in the F1 and F2 males. RTMs were also observed in the 1000- and 300-ppm groups of the F1 and F2 generations. A single incidence of a seminal vesicle and prostate malformation was observed in the F1 10-ppm group and 30-ppm group groups. A single testis malformation (tunica albuginea aplasia) was noted in the control (1.5 ppm) group in breeding male of the F2 generation but was not considered for the analysis of malformations because it was not typical of the phthalate syndrome. There was a significant increase in total RTMs, consistent with the phthalate syndrome, at and above 300 ppm when malformations over the F1 and F2 generations were combined (Table 3). The dose-response curves were similar for the F1 and F2 generations and when the RTM of the two generations were combined (Fig. 1). A Hill model (Graphpad Prism 5.01, San Diego, CA) of the 1.5–10,000 ppm RTM data, which was constrained between 0 and 100% incidence, calculated a half maximal concentration (EC_{50}) of 2771 ppm ($R^2 = 0.9139$) for the F1 generation, an EC_{50} of 1480 ppm ($R^2 = 0.9894$) for the F2 generation, and the EC_{50} for

TABLE 1
Summary of Pregnancy Index^a from Three Matings of Cohabited Breeding Pairs^b from Each Generation Exposed to DEHP

Litter	Dose (ppm)							
	1.5	10	30	100	300	1000	7500	10,000
F1a	17/17 (100)	17/17 (100)	17/17 (100)	17/17 (100)	17/17 (100)	16/17 (94)	17/17 (100)	16/17 (94)
F1b	17/17 (100)	16/17 (94)	16/17 (94)	17/17 (100)	17/17 (100)	16/17 (94)	16/17 (94)	15/17 (88)
F1c	14/17 (82)	12/17 (71)	15/17 (88)	14/17 (82)	17/17 (100)	14/16 (88)	13/17 (76)	10/17 (59)
F2a	15/17 (88)	17/17 (100)	16/17 (94)	17/17 (100)	16/17 (94)	16/17 (94)	12/17 (71)	17/17 (100)
F2b	13/17 (76)	13/17 (76)	15/17 (88)	15/17 (88)	13/16 (81)	14/17 (82)	10/17 (59)	15/15 (100)
F2c	10/17 (59)	11/17 (65)	10/17 (59)	10/17 (59)	9/16 (56)	11/17 (65)	10/17 (59)	12/15 (80)
F3a	17/17 (100)	17/17 (100)	17/17 (100)	17/17 (100)	17/17 (100)	17/17 (100)	9/17 (53)*	—
F3b	17/17 (100)	17/17 (100)	17/17 (100)	17/17 (100)	17/17 (100)	16/17 (94)	8/17 (47)*	—
F3c	11/17 (65)	16/17 (94)*	13/17 (76)*	12/17 (71)	8/17 (47)	12/17 (71)	6/17 (35)	—

^aPregnancy index = number of females delivering/number of cohabiting pairs (percent pregnant). Each dose group is compared with the control group with Fisher's exact test (*one-sided $p < 0.05$).

^bOnly one male and one female pup from each F1c and F2c litter was selected to breed to produce the next generation.

TABLE 2

Summary of Average Litter Size, Proportion of Pups Born Alive, and Pup Weights (combined male and female) from the Three Litters Produced by the P0, F1, and F2 Generations

End point	Dose (ppm)								
	1.5	10	30	100	300	1000	7500	1.5	10,000
P0 litter size ^{a,b}	15.2 ± 0.6	15.5 ± 0.5	13.6 ± 0.8	14.4 ± 0.6	14.3 ± 0.6	14.0 ± 0.6	12.5 ± 0.7*	14.4 ± 0.7	13.1 ± 0.9
F1 litter size	13.3 ± 0.6	11.3 ± 0.9	14.1 ± 0.5	14.0 ± 0.6	12.5 ± 0.8	12.6 ± 0.5	12.3 ± 1.2	—	—
F2 litter size	12.1 ± 0.9	11.9 ± 0.9	13.3 ± 0.6	14.1 ± 0.5	13.6 ± 0.8	12.2 ± 0.7	11.5 ± 1.3	—	—
P0 live offspring ^{b,c}	99 ± 1%	98 ± 1%	94 ± 4%	97 ± 1%	98 ± 1%	95 ± 2%	96 ± 2%	100 ± 0%	96 ± 2%*
F1 live offspring	99 ± 1%	97 ± 1%	98 ± 1%	99 ± 1%	98 ± 1%	97 ± 1%	95 ± 2%	—	—
F2 live offspring	97 ± 1%	95 ± 2%	98 ± 1%	97 ± 1%	96 ± 3%	95 ± 2%	97 ± 1%	—	—
F1 pup body weight (g) ^{b,d}	6.52 ± 0.10	6.47 ± 0.09	6.65 ± 0.08	6.76 ± 0.09	6.65 ± 0.08	6.52 ± 0.09	6.30 ± 0.11	6.62 ± 0.08	6.00 ± 0.14*
F2 pup body weight (g)	6.66 ± 0.13	6.64 ± 0.12	6.64 ± 0.08	6.63 ± 0.08	6.61 ± 0.15	6.64 ± 0.12	5.97 ± 0.14*	—	—
F3 pup body weight (g)	6.34 ± 0.11	6.53 ± 0.10	6.43 ± 0.07	6.46 ± 0.12	6.23 ± 0.12	6.51 ± 0.10	6.09 ± 0.21	—	—

^aMean ± SE of live pups (number of pregnant pairs: P0 *n* = 12–17, F1c *n* = 12–17, F2c *n* = 9–17).

^bEach dose group is compared with the control group with Shirley's test when a trend is present (*p* < 0.01 from Jonckheere's trend test), otherwise Dunn's test used.

^cMean of the average proportion of pups born alive to each fertile pair ± SE (number of pregnant pairs: P0 *n* = 16–17, F1c *n* = 12–17, F2c *n* = 9–17).

^dMean ± SE of pup weight at PND 1 (number of pregnant pairs: P0 *n* = 12–17, F1c *n* = 12–17, F2c *n* = 9–17).

*One-sided *p* < 0.05.

the combined F1 + F2 generation was 2094 ppm ($R^2 = 0.9564$). The roughly twofold difference in EC₅₀ between the F1 and F2 generations is due to the low-incidence observed in the F1 1000-ppm group, and when removed, the F1 EC₅₀ is 1406 ppm ($R^2 = 0.9564$) similar to the F2 generation EC₅₀. After evaluating

various types of BMD models, the Weibull model provided the best fit across the F1, F2, and F1 + F2 data sets (see Supplementary data). The BMD and BMDL values for the F1, F2, and combined F1 + F2 data were 257 and 169 ppm, 233 and 77 ppm, and 198 and 142 ppm, respectively.

TABLE 3

The Incidence of RTMs Observed Grossly Across the F1 and F2 Generations

End point	Dose (ppm)							
	1.5	10	30	100	300	1000	7500	10,000
F1 generation								
Number of litters (number of male pups)	14 (56)	13 (46)	16 (49)	15 (51)	17 (55)	15 (52)	13 (40)	10 (44)
Number of male pups per litter	2–5	1–5	1–5	2–5	1–4	1–5	1–5	3–6
Testis ^a	0 (0)	0 (0)	0 (0)	0 (0)	3 (3)	0 (0)	9*** (17)	0 (0)
Epididymis	0 (0)	0 (0)	0 (0)	0 (0)	2 (2)	0 (0)	1 (1)	0 (0)
Seminal vesicle	0 (0)	1 (1)	0 (0)	0 (0)	2 (2)	0 (0)	1 (1)	0 (0)
Prostate	0 (0)	0 (0)	1 (1)	0 (0)	0 (0)	2 (4)	2 (2)	0 (0)
Total RTM	0 (0)	1 (1)	1 (1)	0 (0)	4 (4)	2 (4)	9*** (18)	0 (0)
F2 generation								
Number of litters (number of male pups)	10 (37)	10 (35)	10 (35)	8 (31)	8 (31)	10 (35)	9 (30)	
Number of male pups per litter	2–5	1–5	1–5	3–5	1–6	1–5	1–5	
Testis	0 (0)	0 (0)	0 (0)	0 (0)	1 (1)	3 (3)	9*** (20)	
Epididymis	0 (0)	0 (0)	0 (0)	0 (0)	1 (1)	3 (3)	9*** (16)	
Seminal vesicle	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	
Prostate	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	
Total RTM	0 (0)	0 (0)	0 (0)	0 (0)	1 (1)	3 (3)	9*** (20)	
Total RTM, F1 + F2	0/24 (0/93)	1/23 (1/81)	1/26 (1/84)	0/23 (0/82)	5/25* (5/86)	5/25* (7/87)	18/22*** (38/70)	

^aNumber of litters affected (total number of pups affected).

*Differs from the proportion of affected litters in the 1.5-ppm group by one-sided Fisher's exact test at *p* < 0.05; ****p* < 0.001.

FIG. 1. Dose response of the RTMs in the F1 (○), F2 (□), and the F1 + F2 combined (Δ) data with a fitted Hill model. The EC_{50} is 2771 ppm ($R^2 = 0.9139$) for the F1 generation, 1480 ppm ($R^2 = 0.9894$) for the F2 generation, and 2094 ppm ($R^2 = 0.9564$) for the combined F1 + F2 generation.

The Monte Carlo simulations demonstrated that when more males per litter were examined using the non-mated cohort, the ability to detect an effect (i.e., statistical power) increased (Fig. 3). Detection was defined as getting a p value of ≤ 0.05 using Fisher's exact test with the litter as the unit of analysis. If the probability of an abnormality is 20% in each treated pup and 0% in each control pup, this effect can be detected 99.5% of the time when 3 male pups are examined from each of 20 litters per group versus only 37.9% of the time when 1 pup per litter was examined. A lower probability of an abnormality of 10% would be detected only 4.7% of the time when one male pup per litter is examined but detected 66.4 and 86.5% of the time when three or four pups are examined, respectively.

DISCUSSION

DEHP produced a dose-dependent increase in the incidence of male RTMs at 300 ppm (~15 mg/kg/day) and higher when F1 and F2 offspring were pooled together. Although not statistically significant, the presence of multiple malformations in multiple litters within the 300-ppm dose group of F1 generation was considered to be an indication of an adverse effect. Based upon the combined F1 and F2 number of litters with malformations, the NOAEL was considered to be 100 ppm or 4.8 mg/kg/day and the lowest observed adverse effect level (LOAEL) was 300 ppm or 14 mg/kg/day. The single RTM present in the 10- and 30-ppm group of the F1 generation were considered to be equivocal responses because it was unclear if these single RTMs present only in the F1 generation were related to treatment. However, these specific malformations observed were part of the phthalate syndrome and therefore their relationship to treatment with DEHP cannot be completely dismissed.

A more accurate dose response (Fig. 1) of DEHP-induced RTMs was generated by examining more pups per litter. The

LOAEL from this study is consistent with a recently published LOAEL of 11 mg/kg/day after a GD 8 to PND 17 exposure to Sprague-Dawley rat dams in which all the male pups within a litter were evaluated in adulthood (Gray *et al.*, 2009). Although the length of exposure differed between the current study and Gray *et al.*, 2009, both studies exposed animals to DEHP during the important in utero window of male rat sexual differentiation. In the current study, the lack of fertility in the F1 10,000-ppm DEHP group is likely due to the widespread incidence of RTMs within males. The BMD values of a 5% response of male RTM incidence by litter did not differ greatly (198–257 ppm) when using data from the individual generations or when the data were combined and were lower than the identified LOAEL. There was more variation in the BMDL values (77–169 ppm) across the F1, F2, and F1 + F2 data sets due to larger variation in the F2 data set. However, these BMDL values were close to the NOAEL of 100 ppm, which may be due to the use of seven dose groups and spacing of doses within this design. Some limitations of the NOAEL approach are its dependency on the number of dose groups, spacing of doses, level of detection, and limitation to one of the experimental doses (Barnes *et al.*, 1995; Sand *et al.*, 2002) such that a design with a low number of dose groups spaced greatly apart can lead to an artificially low NOAEL or insufficiently powered study could lead to an inappropriately high NOAEL. The current study reduced some of these limitations by making use of more pups per litter (i.e., increased detection) and by including additional dose groups.

The power calculations demonstrate that the sampling of additional males per litter can add substantial value to detecting low-incidence effects. When only one male offspring is examined, the dose-response curve of a test article could artificially be steep which will affect the determination of NOAEL and BMD, whereas sampling more males per litter may smooth out the dose-response curve. As mentioned previously, examination of these extra animals makes use of pups already produced that would normally be removed during culling. The utility of adding extra dose groups in order to define the dose response could be hampered by low power when only one male per litter is examined and lead to wastage of animals in the extra dose groups. In the current study, the extra dose groups in combination with evaluating more pups per litter contributed to a well-defined dose response for DEHP. However, if only one pup was evaluated per litter, the DEHP-induced effects in the lower dose litters would likely have been missed so the utility of those extra dose groups would be little to none.

Power calculations demonstrate that increasing the number of animals sampled per litter increases the probability of detecting low-incidence malformations and suggests that if more animals are sampled per litter, fewer litters would be needed thus reducing the overall animal use and production (Gray *et al.*, 2009; Hotchkiss *et al.*, 2008). Although this study focused on power to detect noncontinuous outcomes

FIG. 2. BMD curves for the F1 (A), F2 (B), and F1 + F2 (C) incidence of litters with a reproductive lesion. A BMD (solid line) and BMDL (dashed line) were calculated using a Weibull model for a 5% response. F1 BMD and BMDL: 257 and 169 ppm; F2 BMD and BMDL: 233 and 77 ppm; combined F1 + F2 BMD and BMDL: 198 and 142 ppm.

(i.e., specific androgen-dependent malformations), the ability to detect effects in continuous outcomes (e.g., organ weights) is also influenced by the number of pups sampled per litter (Gray *et al.*, 2009; Hotchkiss *et al.*, 2008).

These data illustrate the increased ability to detect RTMs at lower dose levels by the retention of more adult F1 and F2 males per litter following *in utero* exposure to the endocrine disruptor, DEHP. This study was a feed exposure multigenerational reproductive toxicity assessment conducted under GLP conditions, which incorporates the technical and logistical necessities that are similar to guideline studies. We demonstrate here that it is possible to examine more than one pup per sex per litter in adulthood under these conditions. Retaining these extra littermates makes better use of the animals produced and strengthens the power of the study. While additional pups are examined grossly at weaning under the current EPA and OECD guidelines, malformations in the reproductive tract may not be apparent at this young age. The current sampling of one adult male per litter under the current OECD guidelines would have a 5% power of detecting an effect in males at low litter effect sizes of 10%.

In order to improve the dose-response information from RACB studies, the National Toxicology Program will retain more adult offspring in all its future studies. Regulatory entities should consider incorporating in their respective guidelines the

FIG. 3. Power curves for detecting an effect (one-sided Fisher's exact test $p \leq 0.05$) given the effect size (probability of an abnormality in treated pups) and the number of pups examined per litter, using 20 treated litters and 20 control litters. None of the control litters were affected.

evaluation of an increased number of male and female offspring (e.g., $n = 3$ per sex per litter) in adulthood. This change in study design will improve the dose-response information on postnatal development and provide more informative reproductive toxicity data for use in human health risk assessment.

SUPPLEMENTARY DATA

Supplementary data are available online at <http://toxsci.oxfordjournals.org/>.

FUNDING

Intramural Research Program of the National Institutes of Health; National Institute of Environmental Health Sciences under Research Project Number 1 Z01 ESO45004-11 BB.

REFERENCES

- Barnes, D. G., Daston, G. P., Evans, J. S., Jarabek, A. M., Kavlock, R. J., Kimmel, C. A., Park, C., and Spitzer, H. L. (1995). Benchmark Dose Workshop: criteria for use of a benchmark dose to estimate a reference dose. *Regul. Toxicol. Pharmacol.* **21**, 296–306.
- Chapin, R. E., and Sloane, R. A. (1997). Reproductive assessment by continuous breeding: evolving study design and summaries of ninety studies. *Environ. Health Perspect.* **105**(Suppl. 1), 199–205.
- Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics* **6**, 241–252.
- Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC). (1998). *Endocrine Disruptor Screening and Testing Advisory Committee Final Report*. U.S. EPA, Washington, DC. Available at: <http://www.epa.gov/endo/pubs/edsreview/finalrpt.htm>. Accessed May 27, 2010.
- EPA. (1998). Endocrine Disruptor Screening Program; Proposed Statement of Policy. *Fed. Regist.* **63**, 71542–71568.
- Foster, P. M. D. (2002). Response to comments of Richard H. McKee. *Toxicol. Pathol.* **30**, 757.
- Foster, P. M., and McIntyre, B. S. (2002). Endocrine active agents: implications of adverse and non-adverse changes. *Toxicol. Pathol.* **30**, 59–65.
- Gray, L. E., Jr, Barlow, N. J., Howdeshell, K. L., Ostby, J. S., Furr, J. R., and Gray, C. L. (2009). Transgenerational effects of Di (2-ethylhexyl) phthalate in the male CRL: CD(SD) rat: added value of assessing multiple offspring per litter. *Toxicol. Sci.* **110**, 411–425.
- Gray, L. E., and Foster, P. M. D. (2003). Significance of experimental studies for assessing adverse effects of endocrine-disrupting chemicals. *Pure Appl. Chem.* **75**, 2125–2141.
- Hotchkiss, A. K., Rider, C. V., Blystone, C. R., Wilson, V. S., Hartig, P. C., Ankley, G. T., Foster, P. M., Gray, C. L., and Gray, L. E. (2008). Fifteen years after “Wingspread”—environmental endocrine disruptors and human and wildlife health: where we are today and where we need to go. *Toxicol. Sci.* **105**, 235–259.
- Jonckheere, A. R. (1954). A distribution-free k-sample test against ordered alternatives. *Biometrika* **41**, 133–145.
- McIntyre, B. S., Barlow, N. J., Wallace, D. G., Maness, S. C., Gaido, K. W., and Foster, P. M. (2000). Effects of in utero exposure to linuron on androgen-dependent reproductive development in the male Crl:CD(SD)BR rat. *Toxicol. Appl. Pharmacol.* **167**, 87–99.
- Parks, L. G., Ostby, J. S., Lambright, C. R., Abbott, B. D., Klinefelter, G. R., Barlow, N. J., and Gray, L. E., Jr. (2000). The plasticizer diethylhexyl phthalate induces malformations by decreasing fetal testosterone synthesis during sexual differentiation in the male rat. *Toxicol. Sci.* **58**, 339–349.
- Sand, S., Filipsson, A. F., and Victorin, K. (2002). Evaluation of the benchmark dose method for dichotomous data: model dependence and model selection. *Regul. Toxicol. Pharmacol.* **36**, 184–197.
- Shirley, E. (1977). A non-parametric equivalent Williams’ test for contrasting increasing dose levels of a treatment. *Biometrics* **33**, 386–389.
- Williams, D. A. (1986). A note on Shirley’s nonparametric test for comparing several dose levels with a zero-dose control. *Biometrics* **42**, 183–186.