

# Statistical Approaches for Estimating Sex-Specific Effects in Endocrine Disruptors Research

Jessie P. Buckley,<sup>1,2,3</sup> Brett T. Doherty,<sup>3</sup> Alexander P. Keil,<sup>3</sup> and Stephanie M. Engel<sup>3</sup>

<sup>1</sup>Department of Environmental Health and Engineering, Johns Hopkins University, Baltimore, Maryland, USA

<sup>2</sup>Department of Epidemiology, Johns Hopkins University, Baltimore, Maryland, USA

<sup>3</sup>Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

**BACKGROUND:** When a biologic mechanism of interest is anticipated to operate differentially according to sex, as is often the case in endocrine disruptors research, investigators routinely estimate sex-specific associations. Less attention has been given to potential sexual heterogeneity of confounder associations with outcomes. When relationships of covariates with outcomes differ according to sex, commonly applied statistical approaches for estimating sex-specific endocrine disruptor effects may produce divergent estimates.

**OBJECTIVES:** We discuss underlying assumptions and evaluate the performance of two traditional approaches for estimating sex-specific effects, stratification and product terms, and introduce a simple modeling alternative: an augmented product term approach.

**METHODS:** We describe the impact of assumptions regarding sexual heterogeneity of confounder relationships on estimates of sex-specific effects of the exposure of interest for three approaches: stratification, traditional product terms, and augmented product terms. Using simulated and applied examples, we demonstrate properties of each approach under a range of scenarios.

**RESULTS:** In simulations, sex-specific exposure effects estimated using the traditional product term approach were biased when confounders had sex-dependent associations with the outcome. Sex-specific estimates from stratification and the augmented product term approach were unbiased but less precise. In the applied example, the three approaches yielded similar estimates, but resulted in some meaningful differences in conclusions based on statistical significance.

**CONCLUSIONS:** Investigators should consider sexual heterogeneity of confounder associations when choosing an analytic approach to estimate sex-specific effects of endocrine disruptors on health. In the presence of sex-dependent confounding, our augmented product term approach may be advantageous over stratification when there is prior knowledge available to fit reduced models or when investigators seek an automated test for effect measure modification. <https://doi.org/10.1289/EHP334>

## Introduction

Research on health effects of endocrine-disrupting chemicals (EDs) has increased rapidly in the last decade. This work is predicated on the principle that chemicals that interact with the endocrine system may affect health outcomes that are regulated by that system (Colborn et al. 1993). When an ED is thought to interfere with the synthesis, activity, or elimination of sex hormones or other hormones involved in sexually dimorphic biologic processes (e.g., thyroid), researchers routinely employ statistical methods to assess whether the relationship between the ED and health outcome differs by the sex of the individual. This practice is logical for endpoints that are influenced by hormones and is particularly relevant for studies of developmental outcomes with well-described sexually dimorphic trajectories (Ingleby et al. 2014; Lenroot et al. 2007; Ronen and Benvenisty 2014). For example, sexual dimorphism in human brain anatomy (Lenroot et al. 2007) and epigenetic processes (Nugent and McCarthy 2011) have been proposed as potential explanations for sex-related differences in behavior and risk of developmental disabilities.

Although sexual heterogeneity of associations between EDs and outcomes is often of major interest, sexual heterogeneity of

covariate–outcome relationships is less often scrutinized. And yet, many commonly considered confounders in ED studies may themselves have sex-specific effects. In studies of EDs and child neurodevelopment, for example, instruments that evaluate parenting and the home environment are often included as confounders. Although parenting styles have been reported to have sex-specific associations with behavior, conduct problems, and language development (Braza et al. 2015; Tung et al. 2012; Valloton et al. 2012), the potential for sex differences in confounding by these instruments is typically overlooked. We suggest that common modeling approaches for estimating sex-specific effects of EDs may not yield equivalent estimates due to the presence of such sex-dependent confounding.

We review strategies for estimating sex-specific effects (i.e., stratification and product term models) and propose a simple modeling alternative: an augmented product term approach. Using causal diagrams and a simulation study, we elucidate underlying assumptions about covariate–outcome relationships implied by each approach and discuss performance of these methods in terms of bias, precision, mean squared error (MSE), confidence interval coverage, power, and heterogeneity testing. Finally, we describe a motivating example from our own previous research in which estimates generated using these approaches yielded some meaningful differences in conclusions.

## Description of the Problem

Two approaches are commonly applied to estimate sex-specific effects of exposures on outcomes: stratification and product term models. We define stratification as estimation of exposure–outcome associations separately in datasets comprised only of either males or females. We define product term models as estimation of exposure–outcome associations in a single dataset using an exposure by sex product term. To illustrate assumptions underlying these approaches, we consider a scenario in which we are interested in assessing relationships between a continuously measured exposure ( $X$ ), a binary confounder ( $Z$ ), a binary indicator of sex

---

Address correspondence to J.P. Buckley, Departments of Environmental Health and Engineering and Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Room W7513A, Baltimore, MD 21205 USA. Telephone: 410-502-6150. Email: [jbuck119@jhu.edu](mailto:jbuck119@jhu.edu)

Supplemental Material is available online (<https://doi.org/10.1289/EHP334>).

The authors declare they have no actual or potential competing financial interests.

Received 8 April 2016; Revised 29 November 2016; Accepted 12 December 2016; Published 23 June 2017.

**Note to readers with disabilities:** *EHP* strives to ensure that all journal content is accessible to all readers. However, some figures and Supplemental Material published in *EHP* articles may not conform to 508 standards due to the complexity of the information being presented. If you need assistance accessing journal content, please contact [ehponline@niehs.nih.gov](mailto:ehponline@niehs.nih.gov). Our staff will work with you to assess and meet your accessibility needs within 3 working days.

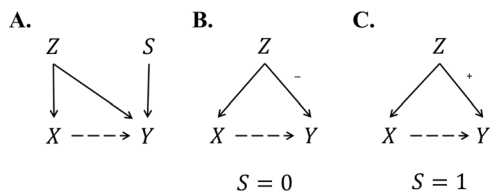
(*S*), and a continuous outcome of interest (*Y*). Figure 1A is a causal diagram representing one potential set of causal associations between these variables.

Under the causal diagram in Figure 1A, adjustment for *Z* is sufficient to estimate an unconfounded association between *X* and *Y*. However, a well-known limitation of causal diagrams is that effect measure modification (EMM) on the model scale is not easily represented (Weinberg 2007). The inability to depict heterogeneity on a single causal diagram may, in part, contribute to under-recognition of the problem of interest in this commentary. One way to express information on sex differences is to draw separate causal diagrams for males and females. Further, the use of signed causal diagrams allows us to represent differential effects by sex (VanderWeele and Robins 2010).

In Figures 1B and 1C, we use signed causal diagrams to demonstrate a reason for our central concern: the direction or magnitude of confounding may depend on *S*. Both causal diagrams assume no heterogeneity by *S* of the association between *Z* and *X*. However, we encode prior knowledge that the association between *Z* and *Y* differs among strata of *S*. For illustrative purposes, we specify a situation with opposing directions of association, although we could have specified a difference in magnitudes of a same-direction effect or no effect in one stratum of *S*. As shown in the diagrams of Figure 1, associations between *Z* and *Y* differ depending on the value of *S*. Such heterogeneity can potentially guide model fitting when evaluating associations between exposure and outcome, and may also affect how we should interpret estimates of heterogeneity within such models.

We describe three potential regression models for estimating the sex-specific effects of *X* with *Y*, controlling for *Z* (Table 1). As described above, the two most commonly applied methods are stratification and the use of an exposure by sex product term (traditional product term approach). The stratification approach is to fit separate models for the effect of *X* on *Y* in datasets stratified by levels of *S*. The traditional product term approach uses a product term between *X* and *S* to allow the effect of *X* on *Y* to differ by *S*.

In stratification, beta coefficients for both *X* and *Z* are estimated separately in datasets subset by strata of *S*. This allows not only the *X* → *Y* association to differ by sex, but also the *Z* → *Y* association to differ by sex. In contrast, the traditional product term approach estimates the beta coefficient for *Z* without respect to *S*, representing a weighted average of the sex-specific effects of *Z* on *Y*. If the *Z* → *Y* association differs by sex, as in our Figures 1B and 1C, the traditional product term approach does not fully control for confounding by *Z*. However, this approach is often preferred over stratification because one may easily conduct a test of heterogeneity for the *X* → *Y* association by *S* (e.g., a Wald test or likelihood ratio test comparing models with and without the *X* by *S* product term). Heterogeneity testing using the stratification approach is less common, but can be achieved with a two-sample z-test using model-estimated beta coefficients and variances (Paternoster et al. 1998).



**Figure 1.** Causal diagrams for relationships between exposure (*X*), outcome (*Y*), a measured covariate (*Z*), and sex (*S*) in the overall population (A), for *S*=0 (B), and for *S*=1 (C).

**Table 1.** Approaches for estimating sex-specific associations.

Approach	Linear regression equation
1. Stratification: Stratify by sex	$Y = \alpha + \beta_1 X + \beta_2 Z$ ; where $S = 0$ $Y = \alpha + \beta_1 X + \beta_2 Z$ ; where $S = 1$
2. Traditional product term: Exposure by sex product term	$Y = \alpha + \beta_1 X + \beta_2 S + \beta_3 XS + \beta_4 Z$
3. Augmented product term: Exposure by sex product term and confounder by sex product term	$Y = \alpha + \beta_1 X + \beta_2 S + \beta_3 XS + \beta_4 Z + \beta_5 ZS$

We propose an augmented product term approach as an alternative to these methods. Here, we include a *Z* by *S* product term in addition to the *X* by *S* product term. The sex-specific exposure effects (the beta coefficients for *X* and the *X* by *S* product term) are now adjusted for sex-specific effects of *Z*. In the simple scenario represented by Figures 1B and 1C, stratification and the augmented product term approach yield identical parameter estimates, but the latter approach allows the investigator to formally examine heterogeneity of the *X* → *Y* association by *S* (e.g., using a Wald test or likelihood ratio test for the *X* by *S* product term).

## Simulation Example

### Methods

We sought to illustrate differences in approaches using a simple but realistic simulation scenario. For this example, we simulated data from a hypothetical study assessing sex-specific associations of a log-normally distributed continuous exposure with a normally distributed continuous outcome in a sample size of 250, a common scenario and representative sample size for studies of ED research among children. We modeled exposure based on the distribution of third trimester maternal urinary mono-*n*-butyl phthalate (MnBP) concentrations observed in the Mount Sinai Children’s Environmental Health Center Study. We modeled the outcome after the Mental Development Index (MDI), an age- and sex-standardized neurodevelopmental scale. We included an unspecified binary covariate, *Z*<sub>1</sub>, to represent a confounder with sexually heterogeneous effects on MDI (such as parenting style), and included an additional binary covariate, *Z*<sub>2</sub>, in some scenarios.

For the primary simulated data scenario, we specified random variables as described in Table 2. Boys are denoted by *S*=0 and girls by *S*=1. The outcome was, on average, two units higher among girls ( $\beta_S = 2$ ). We specified modification of the association between *X* and *Y* on the additive scale, such that exposure had no effect in boys [ $\beta_{X(S=0)} = 0$ ], and each one-unit increase in natural log MnBP exposure reduced mean MDI scores by two points in girls [ $\beta_{X(S=1)} = -2$ ]. We also specified strong additive modification of the association between *Z*<sub>1</sub> and *Y* such that *Z*<sub>1</sub> was unassociated with MDI scores in boys [ $\beta_{Z_1(S=0)} = 0$ ], but was related to higher scores in girls [ $\beta_{Z_1(S=1)} = 5$ ].

We note that in the scenario described by this causal diagram, sex does not affect the exposure or confounder (i.e., no arrow between *S* and *X* or between *S* and *Z*<sub>1</sub>). This is often a reasonable assumption, particularly in studies estimating effects of prenatal exposures, but the foundations of our argument apply to scenarios when *S* may also be a confounder. We ran additional simulations under the following alternative scenarios: Scenario 2: negative (rather than positive) *Z*<sub>1</sub> → *Y* association in girls [ $\beta_{Z_1(S=1)} = -5$ ]; Scenario 3: no modification by sex of *Z*<sub>1</sub> → *Y* association [ $\beta_{Z_1(S=0)} = \beta_{Z_1(S=1)} = 5$ ]; Scenario 4: sex not associated with *Y* ( $\beta_S = 0$ ); Scenario 5: weaker modification by sex of *Z*<sub>1</sub> → *Y* association [ $\beta_{Z_1(S=0)} = 0$ ;  $\beta_{Z_1(S=1)} = 2$ ]; Scenario 6:

**Table 2.** Simulation study parameters.

Parameter specification	Variable description
$S \sim \text{Bernoulli}(0.5)$	Child's sex (50% girls)
$Z \sim \text{Bernoulli}(0.5)$	Binary covariate (50% prevalence)
$X \sim \text{Normal}(3 + Z, 1.2)$	Continuous normally distributed log-transformed exposure dependent on Z, modeled after mono- <i>n</i> -butyl phthalate (mean = 3.5, SD = 1.2)
$Y \sim \text{Normal} [(101.75 + (\beta_{X(S=0)} * X) + (\beta_{X(S=1)} * X * S) + (2 * S) + (5 * Z * S)], 15]$	Continuous normally distributed outcome, modeled after the Mental Development Index (mean = 100, SD = 15)

Note: SD, standard deviation.

additional confounder with sex-specific associations with  $Y$  [ $\beta_{Z2(S=0)} = -5$ ;  $\beta_{Z2(S=1)} = 0$ ]; Scenario 7: additional confounder with sex-specific associations with  $Y$  [ $\beta_{Z2(S=0)} = 0$ ;  $\beta_{Z2(S=1)} = -5$ ]; and Scenario 8: additional confounder without sex-specific associations with  $Y$  [ $\beta_{Z2(S=0)} = \beta_{Z2(S=1)} = -5$ ].

We estimated associations between exposure and outcome using linear regression models fit with each of the three analysis methods detailed in Table 1. We also tested a reduced version of the augmented product term approach for Scenario 8, where the model included a sex product term for Z1 but not Z2. We report average beta coefficients, standard errors, and MSE (squared bias plus variance) across 10,000 simulated datasets. We calculated confidence interval coverage as the percent of simulations in which the 95% confidence interval included the true coefficient value. We calculated power for effect estimates as the percent of simulations in which the 95% confidence intervals excluded the null hypothesis (for the null effect in boys, this is equivalent to the type I error rate). We also report the power to assess heterogeneity as the percent of simulations that identified statistically significant EMM (two-sided alpha of 0.1). For stratified models, we tested EMM using a two-sample z-test to compare the coefficients for boys and girls, with the test statistic given as  $Z = [\beta_{X(S=0)} - \beta_{X(S=1)}] / \sqrt{[Var(\beta_{X(S=0)}) + Var(\beta_{X(S=1)})]^{-one-half}}$  (Paternoster et al. 1998), where model estimates of the coefficients and variances were used to calculate the statistics. For the traditional and augmented product term approaches, we determined power using model-based Wald  $p$ -values, which are output by default in SAS. As a sensitivity analysis to examine performance of the approaches in a larger sample size, all analyses were repeated using  $N = 500$ . We conducted analyses using SAS version 9.4 and fit linear regression models using the GENMOD procedure (SAS Institute Inc.). Simulation code is available in Supplemental Material, Simulation Code.

## Results

Simulation results are presented in Table 3. All results for stratification and the augmented product term approach were identical. In all scenarios, stratification and the augmented product term approach yielded unbiased exposure–outcome effect estimates for both boys and girls. The traditional product term approach yielded biased results except when there was no sex-dependent confounding (Scenario 3) or when two sex-dependent confounders had equal and opposite associations with the outcome, and bias was effectively canceled out (Scenario 6). In the primary scenario, the exposure coefficient estimated with the traditional product term approach was biased down and away from the null for boys (bias =  $-0.4$ ) and up and toward the null for girls (bias =  $0.4$ ).

Standard errors were consistently smaller for the traditional product term approach than they were for stratification or the augmented product term approach. MSEs were also slightly smaller for the traditional product term approach except in

Scenario 6, where coefficients estimated with this approach had large bias. When the sample size was increased to 500, which resulted in smaller standard errors for all approaches, stratification and the augmented product term approach had smaller MSE than the traditional approach in Scenarios 1, 2, 4, and 6 (see Table S1).

For main effects, 95% confidence intervals estimated using stratification and augmented product terms contained the true value in 94% to 95% of simulations for all approaches. The 95% confidence interval coverage using the traditional product term approach ranged between 90% and 95% in the main analyses (Table 3), and between 86% and 95% when the sample size was increased to 500 (see Table S1). The type I error rate for the null effect in boys (expressed as power) followed a similar pattern to the coverage statistic and was 5% to 6% for all scenarios at both sample sizes when using stratification or augmented product terms. The traditional product term approach had type I error rates  $>5\%$  in all scenarios where this approach yielded biased estimates; the type I error rate was highest when estimates were both biased and precise, ranging up to 14% in Scenario 7 with  $N = 500$ .

Compared with stratification and the augmented product term approach, the traditional product term approach had lower power to detect the main effect in girls and effect measure modification in all five scenarios where sex-dependent confounding caused coefficients estimated using this approach to be biased toward the null (Scenarios 1, 4, 5, 7, and 8). In Scenario 3 (no sex-dependent confounding) and Scenario 6 (two sex-dependent confounders with equal and opposite associations with the outcome), the traditional product term approach had slightly better power because estimates were unbiased and standard errors were smaller than the other two approaches. In Scenario 2, the traditional product term model exhibited greater power than the other approaches, due in part to smaller standard errors, but also to substantial bias away from the null, which increased the number of simulations detecting a statistically significant association.

In Scenario 8, where Z1 had sex-dependent effects but Z2 did not, a reduced version of the augmented product term approach performed well. Estimates were unbiased, had 94% to 95% confidence interval coverage, and the average standard errors for boys and girls were both 1.08. The average MSE was 2.37 for boys and 2.39 for girls, which was higher than the traditional, but lower than the full augmented product term approach. The reduced model had the expected 5% type I error rate for the null effect in boys and greater power than the other approaches to detect the main effect in girls (46%) and EMM (39%).

## Applied Example

The present study was motivated by our previous experiences estimating sex-specific effects of endocrine-disrupting compounds in the Mount Sinai Children's Environmental Health Center study. To illustrate the impact of different approaches in real data, we reanalyzed results of a previous study of prenatal



**Table 3.** Simulation results implementing three approaches to estimate sex-specific effects of  $X$  on  $Y$  ( $N = 250$ ).

Scenario/parameter	Stratification <sup>a</sup>					Traditional product term <sup>b</sup>					Augmented product term <sup>c</sup>				
	$\beta$	SE	MSE	Coverage	Power	$\beta$	SE	MSE	Coverage	Power	$\beta$	SE	MSE	Coverage	Power
<b>Scenario 1</b>															
$\beta$ (boys)	0.0	1.12	2.56	95%	5%	-0.4	1.08	2.51	93%	7%	0.0	1.12	2.56	95%	5%
$\beta$ (girls)	-2.0	1.12	2.57	94%	44%	-1.6	1.08	2.53	93%	33%	-2.0	1.12	2.57	94%	43%
EMM					36%					23%					36%
<b>Scenario 2</b>															
$\beta$ (boys)	0.0	1.12	2.56	95%	5%	0.4	1.08	2.52	94%	6%	0.0	1.12	2.56	95%	5%
$\beta$ (girls)	-2.0	1.12	2.57	94%	44%	-2.4	1.08	2.53	93%	59%	-2.0	1.12	2.57	94%	43%
EMM					36%					59%					36%
<b>Scenario 3</b>															
$\beta$ (boys)	0.0	1.12	2.56	95%	5%	0.0	1.08	2.36	95%	5%	0.0	1.12	2.56	95%	5%
$\beta$ (girls)	-2.0	1.12	2.57	94%	44%	-2.0	1.08	2.37	95%	46%	-2.0	1.12	2.57	94%	43%
EMM					36%					40%					36%
<b>Scenario 4</b>															
$\beta$ (boys)	0.0	1.12	2.56	95%	5%	-0.4	1.08	2.51	93%	7%	0.0	1.12	2.56	95%	5%
$\beta$ (girls)	-2.0	1.12	2.57	94%	44%	-1.6	1.08	2.53	93%	33%	-2.0	1.12	2.57	94%	43%
EMM					36%					23%					36%
<b>Scenario 5</b>															
$\beta$ (boys)	0.0	1.12	2.56	95%	5%	-0.1	1.08	2.39	95%	5%	0.0	1.12	2.56	95%	5%
$\beta$ (girls)	-2.0	1.12	2.57	94%	44%	-1.9	1.08	2.40	94%	41%	-2.0	1.12	2.57	94%	43%
EMM					36%					32%					36%
<b>Scenario 6</b>															
$\beta$ (boys)	0.0	1.12	2.58	94%	6%	0.0	1.05	2.25	95%	5%	0.0	1.12	2.58	95%	6%
$\beta$ (girls)	-2.0	1.12	2.58	95%	43%	-2.0	1.05	2.26	95%	47%	-2.0	1.12	2.58	95%	43%
EMM					35%					42%					35%
<b>Scenario 7</b>															
$\beta$ (boys)	0.0	1.12	2.58	94%	6%	-0.7	1.05	2.69	90%	10%	0.0	1.12	2.58	95%	6%
$\beta$ (girls)	-2.0	1.12	2.58	95%	43%	-1.3	1.05	2.69	90%	25%	-2.0	1.12	2.58	95%	43%
EMM					35%					14%					35%
<b>Scenario 8</b>															
$\beta$ (boys)	0.0	1.12	2.58	94%	6%	-0.3	1.05	2.35	93%	7%	0.0	1.12	2.58	95%	6%
$\beta$ (girls)	-2.0	1.12	2.58	95%	43%	-1.7	1.05	2.35	93%	36%	-2.0	1.12	2.58	95%	43%
EMM					35%					26%					35%

Notes: EMM, effect measure modification; MSE, mean squared error; SE, standard error;  $\beta$ (boys)=0 and  $\beta$ (girls) = -2. Scenario 1: Primary results. Scenario 2: Negative (rather than positive)  $Z1 \rightarrow Y$  association in girls ( $\beta_{Z1(S=1)} = -5$ ) Scenario 3: No modification by sex of  $Z1 \rightarrow Y$  association ( $\beta_{Z1(S=0)} = \beta_{Z1(S=1)} = 5$ ) Scenario 4: Sex not associated with  $Y$  ( $\beta_S = 0$ ). Scenario 5: Weaker modification by sex of  $Z1 \rightarrow Y$  association ( $\beta_{Z1(S=0)} = 0$ ;  $\beta_{Z1(S=1)} = 2$ ) Scenario 6: Additional confounder with sex-specific associations with  $Y$  ( $\beta_{Z2(S=0)} = -5$ ;  $\beta_{Z2(S=1)} = 0$ ). Scenario 7: Additional confounder with sex-specific associations with  $Y$  ( $\beta_{Z2(S=0)} = 0$ ;  $\beta_{Z2(S=1)} = -5$ ). Scenario 8: Additional confounder without sex-specific associations with  $Y$  ( $\beta_{Z2(S=0)} = \beta_{Z2(S=1)} = -5$ ).

<sup>a</sup>Stratify by sex.

<sup>b</sup>Exposure by sex product term.

<sup>c</sup>Exposure by sex product term and covariate by sex product term.

phthalate metabolite concentrations and child neurodevelopment (Doherty et al. 2017).

## Methods

Details of the Mount Sinai Children's Environmental Health Center study have been described previously (Engel et al. 2007). Briefly, primiparous women with singleton pregnancies were enrolled at the Mount Sinai Diagnostic and Treatment Center and two private practices in New York City, and delivered at Mount Sinai Medical Center between 1998 and 2001. In the third trimester, women completed an interviewer-administered questionnaire and provided a spot urine sample. Birth data were obtained from a computerized perinatal database. The pregnancy cohort includes 404 mother-infant pairs. Using procedures described previously (Kato et al. 2005), prenatal urine samples were analyzed at the Centers for Disease Control and Prevention (CDC) for the following phthalate metabolites: monoethyl phthalate (MEP), mono-*n*-butyl phthalate (MnBP), mono-isobutyl phthalate (MiBP), mono(3-carboxypropyl) phthalate (MCPP), monobenzyl phthalate (MBzP), mono(2-ethylhexyl) phthalate (MEHP), mono(2-ethyl-5-hydroxyhexyl) phthalate (MEHHP), mono(2-ethyl-5-oxohexyl) phthalate (MEOHP), and mono(2-ethyl-5-carboxypentyl) phthalate (MECPP). The four metabolites of di(2-ethylhexyl) phthalate (MEHP, MEHHP, MEOHP, MECPP) are expressed as a molar sum ( $\sum$  DEHP).

As described by Doherty et al. (2017), 276 children who attended a follow-up visit at 24 months of age were administered the Bayley Scales of Infant Development II, which produces two indices of cognitive and psychomotor development: the MDI and the Psychomotor Development Index (PDI). Age- and sex-standardized scores for the MDI and PDI have a mean of 100 (range: 50 to 150) and a standard deviation of 15 (Bayley 1993).

Women provided informed consent prior to participation, and the study received approval from the Mount Sinai School of Medicine Institutional Review Board. The involvement of the CDC laboratory was determined not to constitute engagement in research of human subjects. The current analysis was approved by University of North Carolina at Chapel Hill Office of Human Research Ethics.

This analysis builds on the work of Doherty et al. (2017) to illustrate the use of stratification, traditional product term, and augmented product term approaches to estimate sex-specific effects. The analysis includes 258 children with measured prenatal phthalate metabolite concentrations and valid MDI/PDI scores. We followed the analysis strategy of the previous study to estimate sex-specific associations of each phthalate metabolite or DEHP sum with child MDI and PDI (Doherty et al. 2017). Briefly, we creatinine-standardized natural log phthalate biomarker concentrations as recommended by O'Brien et al. (O'Brien et al. 2016), and estimated associations in linear regression models adjusted for natural log-transformed creatinine (mg/dL),

prepregnancy body mass index (kg/m<sup>2</sup>), maternal race/ethnicity (Hispanic/White/Black/other), maternal education (less than high school/high school graduate/some college/college graduate), Home Observation for Measurement of the Environment (HOME) score (Caldwell and Bradley 1979), duration of breastfeeding in months, maternal age in years, child age at testing in months, and maternal marital status (single/married/living with father). Child's sex was evaluated in models as an effect measure modifier.

For the stratification approach, we estimated exposure associations in separate datasets for boys and girls. For the traditional product term approach, we included sex as a covariate and estimated exposure associations by including an exposure by sex product term. For the augmented product term approach, we included sex as a covariate, estimated exposure associations by including an exposure by sex product term, and also included product terms between sex and all adjustment variables. We also fit models using a reduced version of the augmented product term approach. For these models, we did not include product terms between sex and three variables: natural log creatinine, maternal age, and maternal body mass index. We selected product terms to remove for this post hoc analysis based on our prior expectation that these variables would not have sex-dependent effects on neurodevelopment, which we confirmed by examining associations in our data.

For the stratification approach, we calculated the *p*-value for EMM by child's sex using model-based beta coefficients and standard errors, as above. For the product term approaches, we report the model-based Wald *p*-value. We considered modification to be statistically significant at an alpha level of 0.1. We fit linear regression models using the GENMOD procedure in SAS version 4.0 (SAS Institute Inc.).

## Results

Estimated sex-specific associations of phthalate biomarkers with MDI scores are reported in Table 4. Beta coefficients from the stratified approach are identical to those from the augmented product term approach. In both sexes, precision differed between these two approaches because the stratified approach allows the variance of the error term to differ by sex, whereas the augmented product term approach does not. We note that standard errors from the two approaches will not differ in expectation when the residual variance in the outcome is similar in both groups (as demonstrated in the simulation).

Estimates from the traditional product term approach had the smallest standard errors, and beta coefficients were generally farther from the null for boys and closer to the null for girls than estimates from other approaches. Using the stratified approach, three of the six phthalate biomarkers were significantly associated with lower MDI scores among girls, whereas there were no statistically significant associations among girls using the traditional product term approach (Table 4). The augmented and reduced augmented product term approaches yielded statistically significant associations for two and one phthalate biomarkers, respectively, but beta coefficients were more comparable to the stratified model than were the beta coefficients from the traditional product term model.

Conclusions regarding EMM did not differ by approach at our *a priori* alpha level of 0.1. Small differences between the EMM *p*-values reported for stratification and the augmented product term approach were due to rounding error in the hand calculation for stratification. We observed similar patterns for the PDI, though differences in beta coefficients were smaller, and all approaches led to the same conclusions based on statistical tests (see Table S2).

**Table 4.** Change in Mental Development Index score per unit change in ln-transformed standardized phthalate metabolite concentration in boys and girls.

Approach/metabolite	Boys ( <i>n</i> = 131)		Girls ( <i>n</i> = 116)		EMM <i>p</i> -value
	β (SE)	95% CI	β (SE)	95% CI	
<b>Stratification<sup>a</sup></b>					
MEP	1.0 (0.7)	-0.5, 2.4	-0.1 (0.8)	-1.7, 1.5	0.3
MnBP	1.7 (0.8)	0.1, 3.3	-2.8 (1.1)	-5.0, -0.5	0.001
MiBP	1.5 (1.0)	-0.4, 3.5	-2.3 (1.0)	-4.3, -0.2	0.008
MCPP	2.0 (1.0)	0.0, 4.1	-2.4 (1.2)	-4.7, 0.0	0.005
MBzP	1.8 (0.9)	0.1, 3.6	-0.6 (0.9)	-2.3, 1.1	0.05
∑ DEHP	0.1 (0.8)	-1.5, 1.7	1.8 (1.0)	-0.1, 3.7	0.2
<b>Traditional<sup>b</sup></b>					
MEP	1.1 (0.7)	-0.2, 2.4	-0.4 (0.8)	-2.1, 1.2	0.1
MnBP	1.9 (0.8)	0.3, 3.4	-2.2 (1.2)	-4.6, 0.1	0.004
MiBP	1.6 (0.9)	-0.2, 3.4	-2.0 (1.1)	-4.1, 0.2	0.009
MCPP	1.8 (1.0)	-0.2, 3.7	-2.0 (1.3)	-4.5, 0.5	0.02
MBzP	1.8 (0.8)	0.2, 3.4	-0.6 (0.9)	-2.3, 1.1	0.03
∑ DEHP	0.3 (0.8)	-1.2, 1.8	1.2 (1.0)	-0.7, 3.2	0.5
<b>Augmented<sup>c</sup></b>					
MEP	1.0 (0.7)	-0.4, 2.3	-0.1 (0.9)	-1.8, 1.6	0.3
MnBP	1.7 (0.8)	0.2, 3.2	-2.8 (1.2)	-5.2, -0.3	0.002
MiBP	1.5 (0.9)	-0.3, 3.4	-2.3 (1.1)	-4.5, 0.0	0.009
MCPP	2.0 (1.0)	0.1, 4.0	-2.4 (1.3)	-4.9, 0.2	0.007
MBzP	1.8 (0.8)	0.2, 3.5	-0.6 (0.9)	-2.4, 1.2	0.05
∑ DEHP	0.1 (0.8)	-1.4, 1.6	1.8 (1.1)	-0.3, 3.9	0.2
<b>Reduced augmented<sup>d</sup></b>					
MEP	1.0 (0.7)	-0.4, 2.3	-0.1 (0.8)	-1.8, 1.6	0.3
MnBP	1.7 (0.8)	0.2, 3.2	-2.7 (1.2)	-5.1, -0.3	0.002
MiBP	1.5 (0.9)	-0.3, 3.3	-2.1 (1.1)	-4.3, 0.1	0.01
MCPP	2.0 (1.0)	0.1, 3.9	-2.2 (1.2)	-4.6, 0.2	0.007
MBzP	1.8 (0.8)	0.2, 3.4	-0.5 (0.9)	-2.3, 1.3	0.06
∑ DEHP	0.2 (0.7)	-1.3, 1.6	1.4 (1.0)	-0.5, 3.3	0.3

Notes: CI, confidence interval; EMM, effect measure modification; SE, standard error; Beta coefficient (SE) and 95% CI per natural log increase in creatinine-standardized phthalate biomarker concentrations estimated in linear regression models adjusted for ln-transformed urinary creatinine concentration, prepregnancy body mass index, maternal race/ethnicity, maternal education, Home Observation for Measurement of the Environment (HOME) score, duration of breastfeeding, maternal age, child age at testing, and maternal marital status.

<sup>a</sup>Stratify by sex.

<sup>b</sup>Exposure by sex product term.

<sup>c</sup>Exposure by sex product term and covariate by sex product terms for all covariates (results published by Doherty et al. 2017).

<sup>d</sup>Exposure by sex product term and covariate by sex product terms for the following covariates: maternal race/ethnicity, maternal education, HOME score, duration of breastfeeding, child age at testing, and maternal marital status.

In our data, we believe the differences in parameter estimates are driven by residual confounding by sex-dependent associations of the covariates with child neurodevelopment. We observed heterogeneity of associations between maternal race/ethnicity, maternal education, and HOME score with MDI score by child's sex. For example, higher levels of maternal education were strongly associated with higher MDI scores in girls but not boys (EMM *p*-value < 0.05 for girls in each of the two highest education levels compared to boys). Sex differences in associations of these variables with MDI may be related to parenting styles, which have been associated with sex differences in child neurodevelopment (Braza et al. 2015; Tung et al. 2012; Vallotton et al. 2012) or other factors that are currently understudied. This difference in confounding by sex is unspecified in the traditional product term approach, which likely leads to bias of stratum-specific estimates.

## Discussion

Our findings provide guidance to investigators who observe different results when fitting stratified and traditional product term models. In our experience, investigators intuitively expect the two approaches to yield similar findings and may not fully appreciate the cause of any differences in results. Our analyses suggest

that stratification and the augmented product term approach yield identical estimates that are less biased than the traditional product term approach when there are sex-dependent confounders. This finding has implications for studies in which sex-dependent effects of EDs are of interest, and it specifically highlights that the exposure by sex product term  $p$ -value from a traditional product term model does not quantify the degree of EMM in estimates produced using stratification. These are different models with different underlying assumptions.

The frequency of meaningful differences in interpretation among these approaches is unknown. Investigators rarely report estimates using more than one approach, and sex differences in associations for many outcomes are not well characterized in the literature. In our area of prenatal ED exposures and child development, variables commonly considered to be confounders, such as gestational stress and tobacco smoke exposure, have also been reported to have sex-dependent associations with child health outcomes (Barrett and Swan 2015; Chang et al. 2012; Fergusson et al. 1998; Golding et al. 2014; Petkovsek et al. 2014; Riedel et al. 2014). While more work is needed to determine how often sex-dependent confounding occurs, we suspect that it may be common in the ED literature and necessitates appropriate analysis techniques.

Compared with either stratification or the augmented product term approach, our simulations showed that the traditional product term approach is susceptible to bias in the presence of confounders with sex-dependent effects on the outcome. However, this approach had consistently smaller standard errors and outperformed other methods in terms of MSE in many of the scenarios we explored. We recommend the use of stratification or augmented product terms when bias is of primary concern or when there is prior knowledge that confounders have sex-specific relationships with the outcome under study.

Beta coefficients estimated using the augmented product term approach are equivalent to stratification when the analyst includes product terms between the stratifying variable (i.e., sex) and all other variables and interaction terms of interest. Standard errors are also equivalent when the residual variation in the outcome is similar among categories of the stratifying variable. As we have discussed, a given analysis typically includes many confounders, of which only a subset has sex-dependent associations with the outcome. The reduced version of the augmented product term approach offers a payoff in terms of variance reduction compared to the stratified approach while maintaining improved validity over the traditional product term approach. Indeed, this approach performed well in our simulation when one confounder had sex-dependent effects but the other did not. We recommend reduced models be specified *a priori* by excluding product terms for variables with strong evidence indicating no sex differences based on relationships reported in the literature. Alternatively, a data-driven approach could be cautiously applied to avoid well-known limitations of automated model reduction procedures (Greenland 1989). We suggest prespecifying product terms of interest using prior knowledge and reporting details of any additional post hoc analyses for model reduction. In general, the augmented product term approach has advantages over stratification when there is *a priori* knowledge available to fit reduced models or when investigators seek an automated test for EMM of fully stratified estimates.

In our applied example, absolute differences in beta coefficients were small, but we observed some meaningful differences in conclusions between estimation approaches. In other studies, the variability in estimated parameters between approaches will depend on the degree of sexual heterogeneity of confounder associations. The common practice of hypothesis testing based on a

specified *a priori* heterogeneity criterion (e.g.,  $p$ -value  $< 0.1$ ) may cause subtle modeling differences to alter conclusions for estimates near the threshold, particularly in the setting of small sample sizes that characterize ED studies. In general, however, it is more informative to report estimates of heterogeneity (such as product term coefficients or alternative measures, such as the relative excess risk due to interaction) rather than results of hypothesis tests (Greenland et al. 2008). We emphasize this point because, as shown in our simulations, increased power to detect modification can result either from a more sensitive hypothesis test or from bias.

This work examined differences between analysis methods under a relatively narrow range of simulated scenarios. We selected parameters to illustrate our issue of concern using strong sex-dependent associations of covariates with the outcome, and we assessed the robustness of our findings to varying parameters within our simple example. Although many more complex situations can be conceived, the qualitative results likely hold over a range of parameter values. The relative utility of the various approaches will depend on the question at hand. For example, when two or more modifiers are of simultaneous interest, investigators may choose to stratify into groups cross-classified by the modifiers (e.g., White males, White females, Black males, Black females), as the augmented product term approach may become complex in this setting. Alternatively, an investigator could combine approaches, stratifying by one variable and using augmented product terms for the other. Rather than providing prescriptive guidance on which method to use, we aimed to describe the performance of available options in various situations so that investigators may choose an optimal approach for their analysis.

## Conclusions

Our findings suggest that variation in estimation approaches may contribute to between-study heterogeneity in findings and that establishing a consistent methodology is warranted to facilitate comparisons. More generally, we recommend against the traditional product term approach when covariates may have sexually heterogeneous associations with the outcome, as it can result in biased estimates. Investigators interested in estimating sex-specific effects of EDs should examine prior literature to understand potential confounding due to sex differences in covariate–outcome associations and apply appropriate methods to account for such heterogeneity, such as stratification or the augmented product term approach.

## Acknowledgments

We thank M. Wolff for her leadership of the Mount Sinai Children's Environmental Health and Disease Prevention Research Center. We thank Antonia Calafat, Manori Silva, Ella Samandar, and Jim Preau for the measurement of the phthalate metabolites. B.T.D. and A.P.K. were funded by a training grant from the National Institute of Environmental Health Sciences (NIEHS, T32 ES007018). J.P.B. and S.M.E. were supported by a grant from NIEHS (5R01ES021777). The Mount Sinai Children's Environmental Health Study was supported by grants from NIEHS (ES009584), U.S. EPA (R827039 and RD831711), ATSDR, and The New York Community Trust.

## References

- Barrett ES, Swan SH. 2015. Stress and androgen activity during fetal development. *Endocrinology* 156(10):3435–3441, PMID: 26241065, <https://doi.org/10.1210/en.2015-1335>.
- Bayley N. 1993. *Bayley Scales of Infant Development*, 2nd edition. San Antonio, TX: Psychological Corporation.

- Braza P, Carreras R, Muñoz JM, Braza F, Azurmendi A, Pascual-Sagastizábal E, et al. 2015. Negative maternal and paternal parenting styles as predictors of children's behavioral problems: Moderating effects of the child's sex. *J Child Fam Stud* 24(4):847–856, <https://doi.org/10.1007/s10826-013-9893-0>.
- Caldwell BM, Bradley RH. 1979. *HOME Observation for Measurement of the Environment*. Little Rock, AK:University of Arkansas Press.
- Chang L, Cloak CC, Jiang CS, Hoo A, Hernandez AB, Ernst TM. 2012. Lower glial metabolite levels in brains of young children with prenatal nicotine exposure. *J Neuroimmune Pharmacol* 7(1):243–252, <https://doi.org/10.1007/s11481-011-9311-6>.
- Colborn T, vom Saal FS, Soto AM. 1993. Developmental effects of endocrine-disrupting chemicals in wildlife and humans. *Environ Health Perspect* 101(5):378–384, PMID: 8080506.
- Doherty BT, Engel SM, Buckley JP, Silva MJ, Calafat AM, Wolff MS. 2017. Prenatal phthalate biomarker concentrations and performance on the Bayley Scales of Infant Development-II in a population of young urban children. *Environ Res* 152:51–58, PMID: 27741448, <https://doi.org/10.1016/j.envres.2016.09.021>.
- Engel SM, Berkowitz GS, Barr DB, Teitelbaum SL, Siskind J, Meisel SJ, et al. 2007. Prenatal organophosphate metabolite and organochlorine levels and performance on the Brazelton neonatal behavioral assessment scale in a multiethnic pregnancy cohort. *Am J Epidemiol* 165(12):1397–1404, <https://doi.org/10.1093/aje/kwm029>.
- Fergusson DM, Woodward LJ, Horwood LJ. 1998. Maternal smoking during pregnancy and psychiatric adjustment in late adolescence. *Arch Gen Psychiatry* 55(8):721–727, PMID: 9707383.
- Golding J, Northstone K, Gregory S, Miller LL, Pembrey M. 2014. The anthropometry of children and adolescents may be influenced by the prenatal smoking habits of their grandmothers: A longitudinal cohort study. *Am J Hum Biol* 26(6):731–739, <https://doi.org/10.1002/ajhb.22594>.
- Greenland S. 1989. Modeling and variable selection in epidemiologic analysis. *Am J Public Health* 79(3):340–349, PMID: 2916724.
- Greenland S, Lash T, Rothman K. 2008. Concepts of interaction. In: *Modern Epidemiology*. 3rd Edition. PA:Lippincott Williams & Wilkins.
- Ingleby FC, Flis I, Morrow EH. 2014. Sex-biased gene expression and sexual conflict throughout development. *Cold Spring Harb Perspect Biol* 7(1):1–17, PMID: 25376837, <https://doi.org/10.1101/cshperspect.a017632>.
- Kato K, Silva MJ, Needham LL, Calafat AM. 2005. Determination of 16 phthalate metabolites in urine using automated sample preparation and on-line preconcentration/high-performance liquid chromatography/tandem mass spectrometry. *Anal Chem* 77(9):2985–2991, PMID: 15859620, <https://doi.org/10.1021/ac0481248>.
- Lenroot RK, Gogtay N, Greenstein DK, Wells EM, Wallace GL, Clasen LS, et al. 2007. Sexual dimorphism of brain developmental trajectories during childhood and adolescence. *Neuroimage* 36(4):1065–1073, PMID: 17513132, <https://doi.org/10.1016/j.neuroimage.2007.03.053>.
- Nugent BM, McCarthy MM. 2011. Epigenetic underpinnings of developmental sex differences in the brain. *Neuroendocrinology* 93(3):150–158, PMID: 2141982, <https://doi.org/10.1159/000325264>.
- O'Brien KM, Upson K, Cook NR, Weinberg CR. 2016. Environmental chemicals in urine and blood: Improving methods for creatinine and lipid adjustment. *Environ Health Perspect* 124(2):220–227, <https://doi.org/10.1289/ehp.1509693>.
- Paternoster R, Brame R, Mazerolle P, Piquero A. 1998. Using the correct statistical test for the equality of regression coefficients. *Criminology* 36(4):859–866, <https://doi.org/10.1111/j.1745-9125.1998.tb01268.x>.
- Petkovsek MA, Boutwell BB, Beaver KM, Barnes JC. 2014. Prenatal smoking and genetic risk: Examining the childhood origins of externalizing behavioral problems. *Soc Sci Med* 111:17–24, <https://doi.org/10.1016/j.socscimed.2014.03.028>.
- Riedel C, Fenske N, Müller MJ, Plachta-Danielzik S, Keil T, Grabenhenrich L, et al. 2014. Differences in BMI z-scores between offspring of smoking and nonsmoking mothers: a longitudinal study of German children from birth through 14 years of age. *Environ Health Perspect* 122(7):761–767, PMID: 24695368, <https://doi.org/10.1289/ehp.1307139>.
- Ronen D, Benvenisty N. 2014. Sex-dependent gene expression in human pluripotent stem cells. *Cell Rep* 8(4):923–932, PMID: 25127145, <https://doi.org/10.1016/j.celrep.2014.07.013>.
- Tung I, Li JJ, Lee SS. 2012. Child sex moderates the association between negative parenting and childhood conduct problems. *Aggress Behav* 38(3):239–251, PMID: 22531998, <https://doi.org/10.1002/ab.21423>.
- Vallotton CD, Harewood T, Ayoub CA, Pan B, Mastergeorge AM, Brophy-Herb H. 2012. Buffering boys and boosting girls: The protective and promotive effects of Early Head Start for children's expressive language in the context of parenting stress. *Early Child Res Q* 27(4):696–707, PMID: 23166405, <https://doi.org/10.1016/j.ecresq.2011.03.001>.
- VanderWeele TJ, Robins JM. 2010. Signed directed acyclic graphs for causal inference. *J R Stat Soc Series B Stat Methodol* 72(1):111–127, PMID: 25419168, <https://doi.org/10.1111/j.1467-9868.2009.00728.x>.
- Weinberg CR. 2007. Can DAGs clarify effect modification?. *Epidemiology* 18(5):569–572, PMID: 17700243, <https://doi.org/10.1097/EDE.0b013e318126c11d>.