Taking Selection Seriously in Correlational Studies of Child Development:

A Call for Sensitivity Analyses

Eric Dearing

Boston College

University of Oslo

Henrik Daae Zachrisson

University of Oslo

Published in Child Development Perspectives, online first, on Oct 5, 2019

Work on this article was supported by funding from the European Research Council Consolidator grant ERC-CoG-2018 EQOP (grant number 818425) and from the Research Council of Norway (grant number 276672).
Correspondence concerning this article should be addressed to Eric Dearing, Lynch School of Education and Human Development, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA 02476; e-mail: dearinge@bc.edu.

Key words: *correlational studies, omitted variable bias, sensitivity analysis*

Abstract

Correlational studies have played a major role in building our cumulative knowledge on child development. Yet as a result, we often have difficulty making causal inferences. The concern is selection effects: When children have not been randomly assigned to conditions, pre-existing biological, psychological, or social factors may bias correlations. In this article, we draw attention to sensitivity analyses, statistical techniques for estimating the robustness or fragility of results in light of potential selection effects. We highlight the coefficient of proportionality method recently developed by Oster (2019), which does not require assumptions about the number of omitted selection variables. The coefficient of proportionality provides an indication of how large the impact of unobserved selection factors would need to be—relative to observed covariates—to nullify a result. We offer two empirical examples to demonstrate the value of this method compared with other approaches used by child development researchers.

Much of what we know about individual differences in children's development is based on correlational studies. While randomized controlled experiments are ideal for establishing causal links in development, they are the exception rather than the rule in many sub-areas of study. This is because assigning children randomly to rearing conditions can raise practical or ethical dilemmas. Among the cases in point are the relatively frequent contributions of correlational studies (and relatively rare contributions of experimental studies) to what we know about the development of children raised in various parenting and family contexts. For the field, one consequence of this reliance on correlational rather than experimental studies is that we often have difficulty making causal inferences.

This difficulty is in large part due to concern for selection effects when children have not been randomly assigned to conditions. The concern is that pre-existing biological, psychological, or social factors may affect children's chances of selecting into certain contexts, as well as their likelihood of displaying developmental outcomes that are correlated with, but not really caused by, these contexts. More generally, this concern is called *omitted variable bias*. Causal interpretations of correlations rest on the assumption that pertinent selection factors have been controlled, and that uncontrolled selection factors are not powerful enough to alter results. In some cases, this assumption may be reasonable, but in many cases, it is not. Researchers in our field, however, rarely tackle the plausibility of this assumption directly. In this article, we argue that *sensitivity analysis* could help developmental researchers address a critical question: How serious would selection bias need to be to nullify or seriously alter a correlational finding?

Handling Selection Effects in Correlational Designs

The most popular approaches for handling selection effects in correlational studies rely on the measurement and statistical control of variables that theory and empirical work indicate may cause selection problems. For example, developmental researchers often use potential selection factors as covariates in multivariate statistical models or to create propensity scores for weighting estimates or matching participants. In this way, researchers attempt to isolate causal effects from selection effects via variance partitioning [AU: Please add a footnote and explain "variance partitioning."] or, in the case of propensity scores, balancing participants according to measured controls. However, these approaches fall short when they assume that all selection variables have been included in the statistical model. If pertinent selection factors are omitted (e.g., if they were not measured or were measured inadequately), associations estimated using covariate adjustment or propensity scores will be biased, either overestimating or underestimating the causal relation (e.g., Duncan, Magnuson, & Ludwig, 2004; Miller, Henry, & Votruba-Drzal, 2016). But the ideal design for avoiding selection problems—a randomized experiment—is not always feasible, for practical or ethical reasons.

With this dilemma in mind, developmental scientists have been encouraged to use statistical methods and research designs that improve the ability to probe causal hypotheses with nonexperimental data. These efforts have focused on methods that do not rely entirely on measured variables to control for selection, such as instrumental variables, fixed-effects, difference-in-difference, and regression discontinuity analysis (e.g., see Duncan et al., 2004; McCartney, Bub, & Burchinal, 2006; Miller et al., 2016, for further explanation of these issues). We agree with these ideas and have used these designs in our programs of research (e.g., Dearing, Zachrisson, & Nærde, 2015; Dearing, Zachrisson, Mykletun, & Toppelberg, 2018; Zachrisson & Dearing, 2015). Yet many instances remain in which developmental researchers cannot use these methods, despite their strengths. In terms of instrumental variable analysis, valid instruments are notoriously difficult to identify in the field of child development. Moreover, researchers often must rely on serendipity to provide natural experiments that provide an opportunity to use difference-in-difference or regression discontinuity methods. Thus, our field is likely to continue to address potential selection bias using covariates and propensity scores in some studies.

Choosing between covariate adjustment and propensity score methods may be less important than the more critical issue of identifying the correct variables to include in models. Researchers empirically tested the relative importance of mode of analysis (i.e., covariate adjustment versus propensity scores), covariate selection, and covariate reliability to try to recover the effects of a randomized trial from a nonrandomized study (Cook, Steiner, & Pohl, 2009; Steiner, Cook, Shadish, & Clark, 2010). They concluded unequivocally that the most important issue was choice of covariates. Next important was the reliability of the covariates. And the least important, with negligible consequences, was whether data were analyzed with covariate adjustment or propensity scores. The authors concluded that “the art of observational study design and practice is to discover the best covariates and to measure them really well” (Cook et al., 2009, p. 845).

Even setting aside measurement challenges that child development (and other social science) researchers routinely face, discovering the best covariates is rarely a straightforward matter. Theory, empirical work, and pilot studies aimed at discovering selection processes are recommended starting points (e.g., Steiner et al., 2010). Nonetheless, given the state of our cumulative knowledge about human behavior, pertinent selection effects may go undiscovered (or poorly measured), even when researchers invest in the discovery process. In the end, researchers analyzing correlational data will often be faced with a judgment call: What is the plausibility of their assumptions that sources of selection bias have been controlled? To address this question, we recommend that researchers conduct a *sensitivity analysis* (e.g., Frank, 2000; Imbens, 2003; Oster, 2019; VanderWeele, 2010).

Although much research has addressed the value of *sensitivity analysis* for correlational studies, its use in child development research is rare. In this article, we highlight an easy-to-implement and informative approach that estimates the sensitivity of results to unobserved selection bias using the impact of observed covariates on the coefficients and overall R-square in regression models (Oster, 2019). This method provides a *coefficient of proportionality* indicating how large the impact of unobserved selection would need to be—relative to the impact of observed covariates—to nullify a result.

Sensitivity Analysis

The term *sensitivity analysis* refers to methods used to estimate the consequences of violating the assumption that all relevant variables have been included in correlational statistical models (e.g., Frank, 2000; Imbens, 2003; Oster, 2019; VanderWeele, 2010). In other words, the methods address the vulnerability of an association, after controlling for observed confounders, to bias caused by omitted variables. Sensitivity analysis has long been discussed in statistics (Rosenbaum & Rubin, 1983) and advocated widely in some social and medical sciences (e.g., Altonji, Elder, & Taber, 2005a, 2005b; Oster, 2019; VanderWeele, 2010). It has also been recommended for psychological scientists (e.g., Imai, Keele, & Tingley, 2010). Although its use in child development research has been rare, notable exceptions (e.g., Ansari & Winsler, 2016; Crosnoe, 2009; McKellar, Marchand, Diemer, Malanchuk, & Eccles, 2019) have most often taken advantage of the impact threshold for a confounding variable (ITCV) method (Frank, 2000).

The ITCV method has been used to indicate how large (or small) the impact of a single unobserved confounder would need to be to nullify an association. The method allows researchers to determine how strong associations would need to be – between the confounding variable that has been omitted and the predictor, and between the omitted confounding variable and the outcome variable – to nullify a correlation between the predictor and outcome. In turn, the size of correlations for observed confounders that were included as covariates in regression models are used as benchmarks. If very large correlations with the confounder—relative to the size of correlations with observed covariates—would be required to invalidate a result, researchers have argued that their results are robust to unobserved selection bias (e.g., Ansari & Winsler, 2016; Crosnoe, 2009; McKellar et al., 2018).

One strength of the ITCV method is its straightforward approach. Yet in correlational studies of child development, assuming a single source of selection bias may be unreasonable; confounding due to different factors may also occur. Moreover, benchmarking using measured covariates can be problematic because unobserved confounders might also bias these estimates (Cinelli & Hazlett, 2018). For this reason, we highlight a method for which the researcher may remain agnostic as to whether selection is due to one or more factors—the coefficient of proportionality approach (Oster, 2019).

Oster (2019) recently detailed a sensitivity analysis method that allows researchers to quantify (and benchmark) how much selection bias due to all possible unobserved selection factors would be necessary to invalidate a result. The method is based on changes in the coefficient for the predictor of interest *and* changes in the overall R-square in regression models, before and after adjusting for observed covariates. The amount of selection bias necessary to invalidate a result is then benchmarked against the multivariate contributions of all observed confounders. This ratio is the coefficient of proportionality. For example, if this ratio is 2-to-1, then unobserved selection factors would need to have twice the impact of observed confounders to reduce an estimate to zero. Oster suggests a coefficient of proportionality greater than 1 as a threshold for robustness.

Our confidence in robustness should be much greater when a coefficient remains stable before and after including a covariate set that has a sizable effect on R-square – total variance explained --[AU: Please define "R-square."] compared with scenarios in which covariate sets produce little to no movement in R-square. Put simply, there is less room for omitted sources of bias when R-square is larger. We and other child development researchers have used changes in predictor coefficients caused by covariates as an indicator of sensitivity without attention to R-square (e.g., Dearing, Kreider, Simpkins, & Weiss, 2006; [AU: Please add]; Lugo-Gil & Tamis-LeMonda, 2008; Magnuson, 2007; NICHD Early Child Care Research Network & Duncan, 2003). This approach has also been discussed in classic regression analysis textbooks (e.g., Cohen, Cohen, West, & Aiken, 2015), but is now discouraged (Oster, 2019). If changes in coefficients are not scaled by changes in the R-square, inferences concerning sensitivity are problematic. Predictor coefficients that change little before and after including covariates should be considered robust to selection bias only to the extent that these covariates explain variance in the outcome.

To illustrate this point, consider a hypothetical scenario in which a researcher wishes to estimate the effects of child care quality on achievement. Assume a correlation between child care quality and achievement, but also assume that two selection factors have been omitted: One of these factors explains 30% of the variance in achievement, while the other explains 10%. If one of these two selection factors is controlled in a regression model, the stability (or lack) of the child care coefficient will depend on which one. The child care coefficient will be more stable when controlling for the selection factor that explains less variance in achievement, but not for reasons related to bias. (For a similar example using wage returns to education, see Oster, 2019.)

Given the importance of R-square changes, a critical assumption underlying sensitivity analysis is the total amount of variance in the outcome that one can reasonably expect to explain. That is, how large would the model R-square be if all relevant background variables were included as predictors? In developmental science, the (un)reliability of our measures and the ubiquity of random error make assumptions of R-square equal to 1.0 unrealistic. Drawing on classic test theory (Lord & Novick, 1968), measure reliability (e.g., Cronbach's alpha) places an upper bound on R-square, assuming a known unidimensional mean or sum score; for example, the highest possible correlation for a measure having a reliability of .80 is .80, yielding a maximum R-square of .64.

To identify a lower bound for maximum R-square, empirical work can be useful. Studies that, by design, capture large portions of the systematic variation in an outcome are valuable for this purpose. For example, consider 1) within-child correlations from short-interval, longitudinal studies, 2) within-child correlations between different measures of the same construct, and 3) correlations for monozygotic twins reared together. Once squared, these correlations help identify the explanatory power of many relevant background factors without directly measuring them, because they capture all relevant genetic and environmental factors that are stable within the child or shared by twins. However, they underestimate the maximum R-square to the extent that unmeasured time-varying factors weaken within-child correlations and nonshared environmental factors weaken twin correlations. In the empirical example that follows, we illustrate the use of these upper and lower bounding approaches for maximum R-square when using the coefficient of proportionality method.

Empirical Example: The Sensitivity of Child Care Quality and Quantity Effects

We illustrate the use of the coefficient of proportionality method for examining sensitivity to omitted variable bias with two empirical examples that use secondary analyses of the NICHD Study of Early Child Care and Youth Development (SECCYD). Our choice of analyses was informed by two published studies using the SECCYD that demonstrated robust associations between 1) quality of child care between 6 and 24 months and children's cognitive achievement at 54 months and 2) quantity of early child care between 6 and 24 months and levels of externalizing behavior problems at 24 months (McCartney et al., 2010; NICHD Early Child Care Research Network & Duncan, 2003). Details on the SECCYD sample, research design, and measures can be found in these published studies, as well as through the Inter-university Consortium for Political and Social Research (2018). The child care variables we created are identical to the original studies and we used the same outcome measure for externalizing behavior problems (i.e., externalizing subscale scores from the Child Behavior Checklist). For simplicity, we used only one measure of cognitive achievement (i.e., Woodcock-Johnson Picture Vocabulary) rather than an average of several measures, as was done in the original study. For our analyses, we chose a covariate set similar to the most heavily controlled models in the original studies (see Table A1).

We regressed externalizing problem scores at 24 months on hours in child care between 6 and 24 months. In turn, we regressed vocabulary scores at 54 months on quality of child care observed from 6 to 24 months. In Table 1, we provide the child care regression coefficients and model R-squares, with and without covariates. Two findings are key; the first is related to the coefficient changes and the second is related to the R-square changes. Adding the covariates reduced the size of the child care quality -> achievement coefficient by more than 50%, but resulted in a relatively small reduction in the size of the child care quantity -> externalizing coefficient. Thus, at first glance, the coefficient for child care quality appeared less stable—in absolute terms—than the coefficient for child care quantity. However, inferences about sensitivity require attention to the changes in R-square, too. Including the covariates increased the R-square nearly six and half times in the achievement model, compared with an increase of a little more than two and half times in the externalizing problem model.

Table 1

Model summaries for covariate adjusted regression estimates of the association between (a) child care quality and cognitive scores at 54 months and (b) child care quantity and externalizing behavior problems at 24 months.

		Achievement Scores at 54 months (N = 680)	
		Unconditional Model	Conditional Model
R-Square		.05	.32
Quality of Care	*b*	1.19***	.58**
	(*SE*)	(.20)	(.19)

		Externalizing Problems at 24 months (N = 531)	
		Unconditional Model	Conditional Model
R-Square		.05	.13
Quantity of Care	*b*	.19***	.17***
	(*SE*)	(.04)	(.04)

Note. Results for model covariates are provided in Table A1.

From these models, we calculated coefficients of proportionality. To do so, we identified a range of maximum R-square values (.60, .70, and .80) using studies reporting squared correlations between multiple measures of the outcome constructs, test-retest reliability, or correlations for monozygotic twin studies (e.g., Achenbach, 1992; Bartels et al., 2003; Goodman & Scott, 1999; McArdle & Woodcock, 1997). Based on these maximum R-square values and the changes in coefficients and R-square from Table 1, we calculated coefficients of proportionality using the Stata program *psacalc* (Oster, 2019; see https://www.brown.edu/research/projects/oster/). This program can be used in *R* through the package RStata (http://github.com/lbraglia/RStata), which allows R users to execute Stata commands.

The results are in Figure 1. While a simple comparison of changes in the child care coefficients might lead to the conclusion that findings related to quantity of care were more robust, the coefficients of proportionality indicate the opposite: Although none of the models reached Oster's recommended level of 1.00, findings related to quality of care appeared more robust to omitted selection factors than findings related to quantity of care across the range of maximum R-square values. The smallest coefficient of proportionality for child care quality -> achievement (i.e., .42) was still larger than any of these coefficients for child care quantity -> externalizing problems, the largest of which was .39. In other words, there was consistent evidence that the child care quantity finding would be invalidated if omitted selection factors were 39% as powerful as those we included in the model; for child care quality, omitted selection factors would need to be somewhere between 42-71% as powerful as those we included.

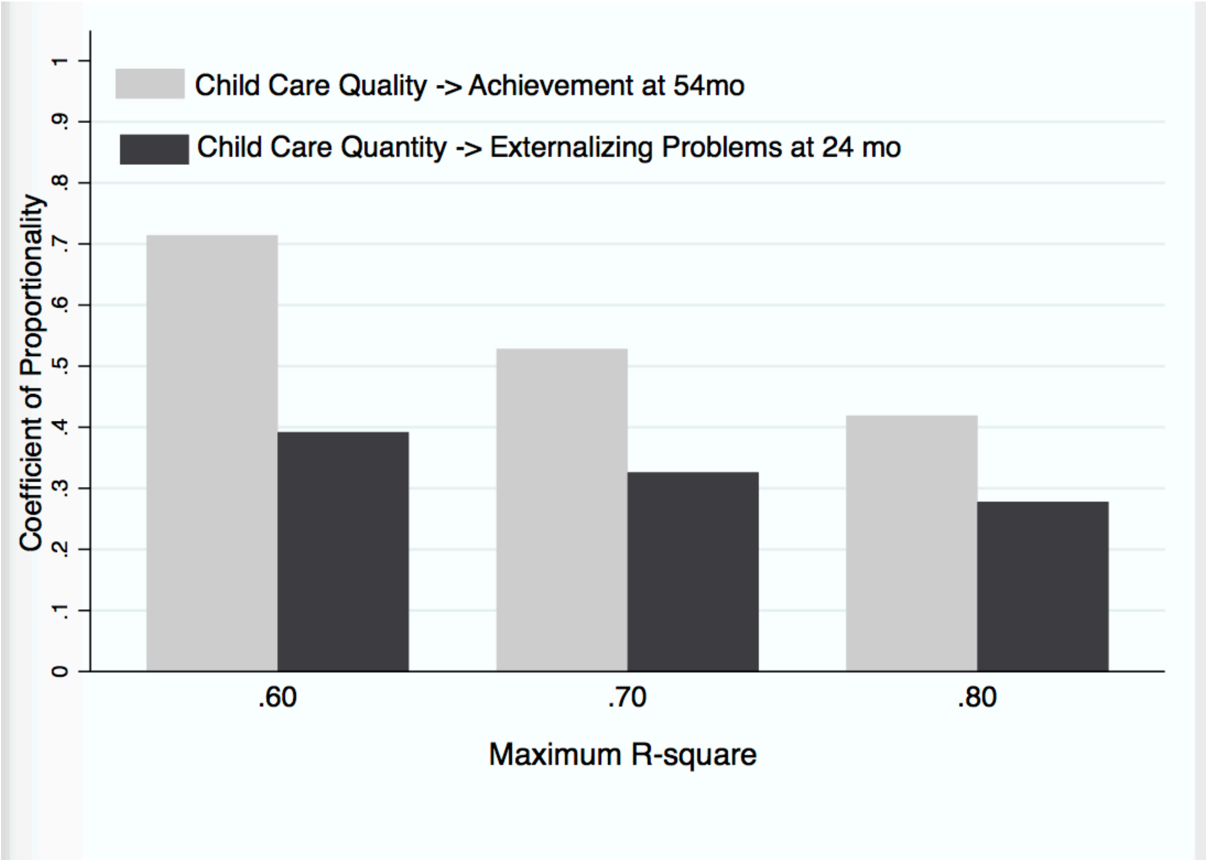

Figure 1

Coefficients of proportionality for associations between (1) quality of child care and cognitive scores at 54 months and (2) quantity of care and externalizing behavior problems at 24 months in the NICHD Study of Early Child Care and Youth Development. Coefficients provide a ratio comparing the strength of selection effects due to unobserved variables needed to invalidate results relative to the strength of selection effects evident in the observed covariates; larger ratios indicate more robust results. The horizontal axis provides three estimates of the maximum R-square (i.e., given measurement error, what percent of variance can be explained in the outcomes if all relevant variables were included?).

It is important to note, however, we did not attempt to replicate the study's findings (McCartney et al., 2010; NICHD Early Child Care Research Network & Duncan, 2003). For simplicity and parsimony, our regression models differed in some ways from the models previously published. The original studies included all children in the SECCYD sample, some of whom did not attend child care, and then used estimation strategies to focus the analyses exclusively on those in child care. Our models included only children in child care. In addition, while the original studies included multiple child care predictors of interest in their regression models, we examined only child care quality and quantity from 6 to 24 months because these were associated robustly with children's outcomes in the original studies. Thus, while the pattern of results reported here is substantively very similar to the original findings, including statistical significance and effect sizes, our coefficients of proportionality are limited to the models estimated here.

Sensitivity Analysis for other Modeling Strategies, with Caveats

Our discussion is concerned with covariate-adjusted regression models with manifest variables, in large part because this is one of the most popular types of analysis in correlational child development research. In the last six issues of *Child Development* (Vol. 90, issues 1-2 and Vol. 89, issues 3-6), 46% of the 115 correlational studies used this method. Yet in many instances, this approach is not ideal (e.g., when examining latent constructs and growth). For these analyses, sensitivity analysis (and *R* code; e.g. Imai et al., 2010) can be conducted with structural equation models. Researchers who prefer propensity score matching or weighting can also find guidance on sensitivity analysis (e.g., VanderWeele & Arah, 2011). Both these approaches allow for benchmarking similar to the Oster method.

The term sensitivity analysis has, in some cases, been used to discuss within-study robustness checks. For example, researchers can look at recommendations for examining the robustness of findings in light of alternative estimation strategies, control variables, and methods for handling missing data (Duncan, Engel, Claessens, & Dowsett, 2014). These types of within-study robustness checks and the type of sensitivity analyses we have described are complementary. Researchers working with correlational data should do both.

However, any correlational study that relies on observed confounding variables to handle selection is in a precarious position with regard to causal inference. This is also true for regression, structural equation, and growth curve models. Sensitivity analysis is not a substitute for study designs that rule out unobserved sources of bias. When it comes to probing causal hypotheses in correlational designs that rely on measuring selection variables, the most critical decision is deciding what to measure (Cook et al., 2009; Steiner et al., 2010). Our statistical models cannot remedy unobserved selection bias in correlational designs that rely on observed covariates. The value of sensitivity analysis is in helping gauge how skeptical we should be that our correlational estimates are unbiased.

Conclusion

Correlational studies have played a major role in building our cumulative knowledge on child development and are likely to continue to do so for years. However, the value of correlational findings rests on the assumption (or hope) that uncontrolled omitted variables are not powerful enough to alter results seriously. Given the array of forces that may act on children's lives, this assumption is unreasonable in many cases. Sensitivity analyses provide a way to tackle the plausibility of this assumption in our correlational studies. More specifically, sensitivity analyses can provide an easy-to-estimate benchmark for how serious selection bias would need to be to nullify a correlational finding. In this article, we have focused on the coefficient of proportionality (Oster, 2019) because it is applicable to either single or multiple omitted selection variables. The value of this method compared with past approaches is evident in our empirical examples, in which stability of regression coefficients (without attention to R-square changes) would provide misleading results concerning robustness.

When it comes to internal validity, sensitivity analysis is not a substitute for experiments or even well-designed quasi-experiments. But it provides a powerful method for evaluating correlational results. We hope our call for using this approach encourages developmental scientists to include sensitivity analyses in their toolkit. Our field should expect explicit discussions of the plausibility of causal assumptions in our correlational designs.

References

Achenbach, T. M. (1992). *Manual for the child behavior checklist/2-3 and 1992 profile*. Burlington, VT: University of Vermont Department of Psychiatry.

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005a). An evaluation of instrumental variable strategies for estimating the effects of Catholic schooling. *Journal of Human Resources, XL*, 791-821. doi:10.3368/jhr.XL.4.791

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005b). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy, 113*, 151-184. doi:10.1086/426036

Ansari, A., & Winsler, A. (2016). Kindergarten readiness for low-income and ethnically diverse children attending publicly funded preschool programs in Miami. *Early Childhood Research Quarterly, 37*, 69-80. doi: 10.1016/j.ecresq.2016.06.002

Bartels, M., Hudziak, J. J., Boomsma, D. I., Rietveld, M. J. H., Van Beijsterveldt, T. C. E. M., & Van den Oord, E. J. C. G. (2003). A study of parent ratings of internalizing and externalizing problem behavior in 12-year-old twins. *Journal of the American Academy of Child & Adolescent Psychiatry, 42*, 1351-1359. doi: 10.1097/01.CHI.0000085755.71002.5d

Cinelli, C., & Hazlett, C. (2018). Making sense of sensitivity: Extending omitted variable bias. Working paper. Retrieved August 23, 2019, from https://www.researchgate.net/profile/Chad_Hazlett/publication/322509816_Making_Sense_of_Sensitivity_Extending_Omitted_Variable_Bias/links/5b483bb10f7e9b4637d1db59/Making-Sense-of-Sensitivity-Extending-Omitted-Variable-Bias.pdf

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2015). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). New York, NY: Routledge.

Cook, T. D., Steiner, P. M., & Pohl, S. (2009). How bias reduction is affected by covariate choice, unreliability, and mode of data analysis: Results from two types of within-study comparisons. *Multivariate Behavioral Research, 44*, 828-847. doi:10.1080/00273170903333673

Crosnoe, R. (2009). Family-school connections and the transitions of low-income youths and English language learners from middle school to high school. *Developmental Psychology, 45*, 1061-1076. doi:10.1037/a0016131

Dearing, E., Kreider, H., Simpkins, S., & Weiss, H. B. (2006). Family involvement in school and low-income children's literacy: Longitudinal associations between and within families. *Journal of Educational Psychology*, *98*, 653. doi: 10.1037/0022-0663.98.4.653

Dearing, E., Zachrisson, H. D., Mykletun, A., & Toppelberg, C. O. (2018). Estimating the consequences of Norway's national scale-up of early childhood education and care (beginning in infancy) for early language skills. *AERA Open, 4*, 1-16. doi:10.1177/2332858418756598

Dearing, E., Zachrisson, H. D., & Nærde, A. (2015). Age of Entry Into Early Childhood Education and Care as a Predictor of Aggression Faint and Fading Associations for Young Norwegian Children. *Psychological Science, 26*, 1595-1607. doi: 10.1177/0956797615595011

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology, 50*, 2417-2425. doi:10.1037/a0037996

Duncan, G. J., Magnuson, K. A., & Ludwig, J. (2004). The endogeneity problem in developmental studies. *Research in Human Development, 1*, 59-80. doi:10.1080/15427609.2004.9683330

Frank, K. A. (2000). Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research, 29*, 147-194. doi:10.1177/0049124100029002001

Goodman, R., & Scott, S. (1999). Comparing the strengths and difficulties questionnaire and the child behavior checklist: Is small beautiful? *Journal of Abnormal Child Psychology, 27*, 17-24. doi:10.1023/A:1022658222914

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*, 309-334. doi: 10.1037/a0020761

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review, 93*, 126-132. doi:10.1257/000282803321946921

Inter-university Consortium for Political and Social Research (2018). *NICHD Study of Early Child Care and Youth Development Series.* Retrieved August 23, 2019, from: https://www.icpsr.umich.edu/icpsrweb/ICPSR/series/00233

Lord, F., & Novick, M. R. (1968). *Statistical theories of mental tests*. New York, NY: Addison-Wesley.

Lugo-Gil, J., & Tamis-LeMonda, C. S. (2008). Family resources and parenting quality: Links to children's cognitive development across the first 3 years. *Child Development, 79*, 1065-1085. doi:doi:10.1111/j.1467-8624.2008.01176.x

Magnuson, K. (2007). Maternal education and children's academic achievement during middle childhood. *Developmental Psychology, 43*, 1497-1512. doi:10.1037/0012-1649.43.6.1497

McArdle, J. J., & Woodcock, R. W. (1997). Expanding test-retest designs to include developmental time-lag components. *Psychological Methods, 2*, 403-435. doi:10.1037/1082-989X.2.4.403

McCartney, K., Bub, K. L., & Burchinal, M. (2006). Selection, reflection, detection. *Monographs of the Society for Research in Child Development, 71*, 105-126. doi:10.1111/j.1540-5834.2006.00407.x

McCartney, K., Burchinal, M., Clarke-Stewart, A., Bub, K. L., Owen, M. T., Belsky, J., & NICHD Early Child Care Research Network (2010). Testing a series of causal propositions relating time in child care to children's externalizing behavior. *Developmental Psychology, 46*, 1-17. doi:10.1037/a0017886

McKellar, S. E., Marchand, A. D., Diemer, M. A., Malanchuk, O., & Eccles, J. S. (2019). Threats and supports to female students' math beliefs and achievement. *Journal of Research on Adolescence, 29,* 449-465. doi:10.1111/jora.12384

Miller, P., Henry, D., & Votruba-Drzal, E. (2016). Strengthening causal inference in developmental research. *Child Development Perspectives, 10*, 275-280. doi:10.1111/cdep.12202

National Institute of Child Health and Human Development Early Child Care Research Network & Duncan, G. J. (2003). Modeling the impacts of child care quality on children's preschool cognitive development. *Child Development, 74*, 1454-1475. doi:10.1111/1467-8624.00617

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, *37,* 187-204. doi:10.1080/07350015.2016.1227711

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55. doi: 10.1093/biomet/70.1.41

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15*, 250-267. doi:10.1037/a0018719

VanderWeele, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology, 4,* 540-551. doi: 10.1097/EDE.0b013e3181df191c

Zachrisson, H. D., & Dearing, E. (2015). Family income dynamics, early childhood education and care, and early child behavior problems in Norway. *Child Development, 86*(2), 425-440. doi:10.1111/cdev.12306

SUPPLEMENTARY MATERIAL
Taking selection seriously in correlational studies of child development:
A call for sensitivity analyses
Eric Dearing & Henrik Daae Zachrisson

Table A1
Results for covariates from regression models summarized in Table 1

Covariate	Picture Vocabulary Score 54 mo *b* (*SE*)	Externalizing Behavior Problems 24 mo *b* (*SE*)
Child characteristics		
Gender = boy[a]	-2.47 (.89)	-.88 (.84)
White[a]	-.65 (2.01)	-2.10 (2.04)
African American[a]	-6.47 (2.52)	-.99 (2.53)
Latin/x American[a]	-5.42 (2.01)	.01 (2.00)
Temperament[b]	.50 (1.26)	-.05 (1.17)
Maternal characteristics		
Partnered[b]	-1.82 (1.61)	.25 (1.53)
Depressive symptoms[b]	-.04 (.07)	.08 (.08)
Years of education[a]	.68 (.22)	-.59 (.24)
Verbal intelligence[c]	.19 (.04)	-.02 (.03)
Separation Anxiety[b]	-.00 (.04)	.01 (.04)
Employment Attitudes[b]	-.01 (.08)	-.03 (.07)
Parenting Beliefs[a]	-.06 (.04)	.01 (.04)
Sensitivity[b]	.46 (.28)	-.08 (.27)
Agreeableness[b]	.07 (.10)	.10 (.10)
Neuroticism[b]	.04 (.09)	.00 (.08)
Extraversion[b]	-.09 (.09)	-.17 (.08)
Family characteristics		
Family income-to-needs[b]	.45 (.17)	.26 (.16)
Study site 1	-.92 (2.05)	.12 (1.81)
Study site 2	1.47 (1.96)	3.13 (1.87)
Study site 3	3.09 (1.92)	.73 (1.75)
Study site 4	3.30 (2.04)	2.86 (1.91)
Study site 5	1.17 (1.96)	-.37 (1.88)
Study site 6	1.55 (2.03)	.19 (2.06)
Study site 7	.36 (2.02)	-.48 (2.04)
Study site 8	.98 (1.92)	-1.20 (1.83)
Study site 9	3.44 (1.96)	-2.64 (1.81)
Child care characteristics		
Hours (6-24mo)	-.01 (.04)	
Quality (6-24mo)		-.10 (.19)

Note. Descriptions of covariate measures are provided in McCartney et al. (2010) and NICHD ECCRN & Duncan (2003). [a]Covariates assessed when children were 1 month of age. [b]Covariates assessed when children were 6 months of age. [c]Covariate assessed when child was 36 months of age.