AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Abnormal Psychology

© 2020 American Psychological Association
ISSN: 0021-843X

http://dx.doi.org/10.1037/abn0000640

Depression in Mothers and the Externalizing and Internalizing Behavior of Children: An Attempt to Go Beyond Association

William E. Pelham III, Stephen G. West, and Kathryn Lemery-Chalfant
Arizona State University

Sherryl H. Goodman
Emory University

Melvin N. Wilson
University of Virginia

Thomas J. Dishion
Arizona State University

Daniel S. Shaw
University of Pittsburgh

Hundreds of studies have documented an association between depression in mothers and behavior problems in children. Theory and empirical findings suggest this association may be confounded by other factors, but little attention has been paid to this issue. We used propensity score methods in a sample of 731 low-income families assessed repeatedly from child age 2 through 14 years to produce a weighted sample of families that were similar at child age 3 years except for mothers' depression. Depressive symptomatology was measured via self-report rating scale. Mothers were categorized as having clinically-elevated versus non-clinically-elevated scores based on an established threshold. Mothers with elevated versus nonelevated scores were equated on 89 other relevant characteristics (e.g., SES, child behavior, marital conflict). We then compared the equated groups on mother, secondary caregiver, and teacher ratings of child externalizing and internalizing behavior from child ages 4 to 14 years. Prior to equating, the mean *prima facie* effect of exposure to clinically-elevated mothers' depression scores at child age 3 years was $d = 0.45$ per mothers, $d = 0.26$ per secondary caregivers, and $d = 0.13$ per teachers. After equating, the mean adjusted effect was $d = 0.07$ per mothers, $d = 0.01$ per secondary caregivers, and $d = 0.03$ per teachers. Findings suggest that a substantial portion of the *prima facie* association between mothers' depression and later child behavior problems is accounted for by confounding variables rather than a causal effect of depressive symptoms per se. To fully understand why children of depressed mothers exhibit more behavior problems, a multicausal theory is needed that jointly considers the cluster of co-occurring clinical features that often accompany maternal depression.

General Scientific Summary
Children whose mothers reported clinically-elevated symptoms of depression at child age 3 years exhibited more externalizing and internalizing behavior between ages 4 and 14 years than children whose mothers did not report clinically significant symptoms of depression. Yet, when families with depressed versus nondepressed mothers were equated on many background variables, the size of the difference shrank substantially. Studies exploring developmental processes that link mothers' depression to children's behavior problems may produce misleading results when they do not account for the many ways in which families with depressed and nondepressed mothers differ.

Keywords: maternal depression, child psychopathology, internalizing, externalizing, propensity score

Supplemental materials: http://dx.doi.org/10.1037/abn0000640.supp

William E. Pelham III, Stephen G. West, and Kathryn Lemery-Chalfant, Department of Psychology, Arizona State University; Sherryl H. Goodman, Department of Psychology, Emory University; Melvin N. Wilson, Department of Psychology, University of Virginia; Thomas J. Dishion, Department of Psychology, Arizona State University; Daniel S. Shaw, Department of Psychology, University of Pittsburgh.

Data were collected with via grants from the National Institute on Drug Abuse (NIDA; DA036832, DA022773). Additional support was provided by a grant from the National Institute on Alcohol Abuse and Alcoholism (AA026768). This study was approved by Institutional Review Board at the University of Pittsburgh, IRB: REN18030126 / PRO13060500, title: Early Family Prevention of Adolescent Alcohol, Drug Use and Psychopathology.

Correspondence concerning this article should be addressed to William E. Pelham III, Department of Psychology, Arizona State University, 950 South McAllister Avenue, Tempe, AZ 85281. E-mail: wpelham@asu.edu

Exposure to a mother who is depressed has negative effects on children's subsequent mental health (Beardslee, Bemporad, Keller, & Klerman, 1983; Connell & Goodman, 2002; Downey & Coyne, 1990). This association is one of the best-documented findings in the field of developmental psychopathology, with the most recent meta-analysis including data from 193 studies comprising more than 80,000 families (Goodman et al., 2011). Results of this meta-analysis indicated that children of depressed mothers display more externalizing behavior ($d = 0.43$), internalizing behavior ($d = 0.47$), and general psychopathology ($d = 0.49$) in both clinical and community samples. The field has replicated the effect in many contexts, delineated the conditions under which the effect is magnified or attenuated, and explored the mechanisms through which the effect is statistically mediated (Goodman, 2020).

Despite much progress, the literature relating mothers' depression to child behavior problems has retained a key weakness: The evidence remains correlational rather than causal. Families with depressed versus nondepressed mothers typically differ on many factors besides depression, including socioeconomic status, living environment, marital conflict, parenting practices, and their children's preexisting level of behavior problems. Each of the factors is a potential *confounding variable*: a variable that (a) is conceptually distinguishable from mothers' depression, (b) is correlated with mothers' depression, and (c) is itself an independent cause of later child behavior problems.[1] These confounding variables may explain part or all of the observed difference in later child behavior, undermining our understanding of the developmental processes truly at work (Gotlib, Goodman, & Humphreys, 2020; Rutter, 2007). For example, greater exposure to the marital conflict that is often concomitant with depression, rather than depression per se, may explain why the children of depressed mothers exhibit more aggression. Such a finding would suggest different priorities for future research and clinical intervention than if mothers' depression itself were causing child aggression.

Researchers have tried several strategies to address the problem of confounding variables. Some studies have controlled for a few readily available covariates such as child sex and age (e.g., Bagner, Pettit, Lewinsohn, & Seeley, 2010). Others have measured mothers' depression and children's behavior repeatedly and modeled their mutual interplay over time (e.g., Choe, Shaw, Brennan, Dishion, & Wilson, 2014). But no study has attempted to control for the broad range of child, mother, and family factors that might confound the relation of mothers' depression to child behavior. The failure to account for confounding variables may be due to the difficulty of including a large number of covariates in the traditional statistical approaches that have been used. For example, although investigators might be able to identify 50 + variables on which families with depressed and nondepressed mothers differ, they could not include 50 + variables in a regression analysis or analysis of covariance without a very large sample size.

Multicausal Theory Versus Monocausal Method

The methodological challenge is connected to a deeper conceptual question: Why do the children of depressed mothers exhibit more behavior problems? Kendler (2019) distinguishes monocausal versus multicausal explanations for psychiatric illness and traces the history of both approaches in medicine and psychiatry. In developmental psychopathology, theorists have long recognized that the etiology of most psychopathology is multicausal (Cicchetti, 1984, 1993; Sroufe & Rutter, 1984): Dysfunction arises from a network of additive and interactive influences of many different factors, at multiple levels of analysis. Yet, a multicausal reality poses difficulty for current methodological approaches, leading to a gap between theory and its empirical tests. If many different intercorrelated causes contribute to a developmental outcome, then isolating and understanding the effect of any specific cause would require accounting for the many other causes that may be simultaneously operating. Because the field's traditional statistical approaches cannot easily account for large number of independent variables, developmental psychopathologists are often forced to operationalize their elaborate, multicausal theory as a simplistic model with just a few causal determinants (e.g., the factor of interest and a few covariates). In some cases, models approach monocausality, in which the observed difference in outcome are attributed to a single preceding factor (e.g., mother's depression). Thus, limitations of traditional methodological approaches have led to a mismatch between theory and practice, precluding a clear understanding of the causal role of any specific risk factor within the larger network of many causes.

Framework for Causal Inference From Nonrandomized Data

The goal of this study was to estimate the causal effect of mothers' depressive symptoms on later child behavior problems by equating depressed and nondepressed mothers on a large set of other factors that might otherwise confound the observed association. We now explain our use of the word *causal* to avoid confusion and misinterpretation. The data we use do not come from a randomized experiment, but from a nonrandomized study following mothers and their children over a 12-year period. In the past, some scholars argued for a proscription on the use of the word *causality* when dealing with nonrandomized studies (e.g., Freedman, 1987), but this view has been superseded in the field of causal inference (e.g., Hernán & Robins, 2020; Imbens & Rubin, 2015; Morgan & Winship, 2014; Pearl & Mackenzie, 2018; Reichardt, 2019). A modern approach to causal inference in nonrandomized (i.e. correlational) studies emphasizes the use of causal language, identification of an explicit causal effect to be estimated, transparent discussion of the assumptions needed to justify a causal interpretation, extensive diagnostic checks of the assumptions, and appropriate caution in interpreting results (Ahern, 2018; Hernán, 2018) — we will pursue each of these elements in this article. Although this approach to causal inference arose outside psychology (in the fields of statistics, epidemiology, economics, and sociology), it has been recognized as a valuable framework for the study of developmental psychopathology (Foster, 2010; Ohlsson & Kendler, 2020; Rutter, 2007).

In this article, we will use the phrase "the causal effect of mothers' depression" to indicate a hypothetical quantity of scientific and practical interest—the expected effect of changing a mother from nondepressed to depressed (or vice versa) at a specific point in time (Rubin, 2005). We cannot directly estimate this hypothetical quantity, but we can estimate adjusted effects that use

[1] See Greenland and Robins (1986) for a formal definition of a confounding variable.

statistical methods to equate depressed and nondepressed mothers on the many other factors that would ideally be held fixed for the comparison. Thus, throughout the article, we will maintain a distinction between *adjusted* effects (what we can estimate from these data) and *causal* effects (the hypothetical quantity of scientific interest). In the Introduction and Methods sections, we will invoke causality during theoretical discussions of what we are trying to estimate and how best to estimate it. In the Results section, we will refer only to the adjusted effect sizes that can be estimated from these data. In the Discussion section, we will discuss the adjusted effects and consider to what extent we have satisfied assumptions that would justify giving these adjusted effects a causal interpretation.

Isolating the Causal Effect of Depressive Symptoms

From the standpoint of causal inference, the ideal way to estimate the causal effect of mothers' depressive symptoms on child behavior problems would be to conduct a randomized trial. At baseline, mothers would be randomized to either be depressed or nondepressed, creating two groups that were initially similar except for the presence of mothers' depression. Their children's mental health outcomes would be measured repeatedly over time. Because the families in the groups with depressed versus nondepressed mothers would be similar (in expectation, identical) at baseline on all measured and unmeasured covariates, any between-groups differences in later child behavior could be causally attributed to exposure to mothers' depression.

Although this hypothetical study would be ideal for causal inference, it is unrealizable in practice—ethical and practical concerns preclude randomization of mothers to levels of depression. However, we can pursue this ideal using methods based on propensity scores, an approach first developed by Rosenbaum and Rubin (1983). Propensity score methods provide machinery for equating exposed and unexposed groups on measured characteristics in nonrandomized studies (also called "correlational" studies). Exposed (e.g., to mothers' depression) and unexposed cases receive weights that are a function of the propensity score, the estimated probability of a case being exposed given its values on a set of baseline covariates. This strategy is called inverse probability of treatment (IPT) weighting (Rosenbaum, 1987). Statistical theory indicates that after IPT weighting, the exposed and unexposed cases will be (in expectation) balanced on all measured baseline characteristics that were included in the estimation of the weights. Using IPT weighting, we can reduce the problem of equating the exposed and unexposed groups on a very large number of potential confounding variables to that of weighting them by a function of a single summary variable—the propensity score.

The propensity score approach consists of two steps (Rubin, 2007, 2008). In Step 1, we estimate IPT weights such that after weighting, the exposed and unexposed groups in a nonrandomized study are balanced on all measured baseline covariates potentially related to the outcome. In Step 2, weighted mean outcomes in the exposed versus unexposed groups are compared over time to estimate an adjusted effect of the exposure that can be interpreted as causal if the appropriate assumptions are satisfied. Thus, the propensity score approach permits us to mimic a hypothetical randomized experiment in which mothers are initially randomized to levels of depression and children's mental health outcomes are assessed repeatedly over time.

Current Study

The current study evaluated the extent to which it is exposure to mothers' depressive symptoms per se that explains why the children of depressed mothers exhibit more externalizing and internalizing behavior. Using propensity score methods, we attempted to isolate the specific causal effect of mothers' depression at child age 3 years from those of the many co-occurring clinical features. Data were drawn from a prospective, longitudinal, multisite dataset in which 731 families were recruited from the Women, Infants, and Children (WIC) Nutritional Supplement program at child age 2 years and assessed repeatedly until age 14 years. Mothers' depression was measured via self-report symptom rating scale at child age 3 years, and mothers were categorized into groups with clinically-elevated versus non-clinically-elevated symptoms using an established threshold. In Step 1, we estimated IPT weights such that in the weighted sample, depressed and nondepressed mothers at child age 3 years were balanced on 89 concurrently measured covariates that might otherwise confound the relation of mothers' depression to child behavior problems. In Step 2, we compared depressed and nondepressed mothers in the weighted sample on child behavior problems over time in order to estimate the adjusted effect of exposure to mothers' depression during early childhood. Finally, we conducted six sensitivity analyses to probe whether results were robust to changes in analytical approach.

Method

Sample

The Early Steps trial has followed 731 at-risk families recruited from the WIC program beginning when children were 2 years old (Dishion et al., 2008; Shaw, Connell, Dishion, Wilson, & Gardner, 2009). Families were recruited at three sites—Eugene, OR; Charlottesville, VA; and Pittsburgh, PA. Families were approached at the WIC office and invited to complete a screening procedure if they had a child between the ages of 2 years 0 months and 2 years 11 months. Families were offered enrollment in the study if they possessed risk factors for the development of child conduct problems from at least two of the following categories: (a) child behavior problems, (b) family problems, and (c) sociodemographic risk. One qualifying family problem risk factor was mothers' depression, contributing to the high rates of mothers' depression in the present study.

At study entry, 37% of primary caregivers were married, and 60% reported having a live-in partner. Twenty-four percent of primary caregivers did not have a high school degree; only 3% had a degree from a 4-year college. Two-thirds of families reported income below $20,000 annually. 50% of children were male, 50% were European American, 28% were African American, and 13% were Hispanic American. These analyses used data from 629 of the 731 families. 74 families were excluded because they did not participate in the study wave at child age 3 years; 27 families were excluded because the primary caregiver was not the child's mother; 1 family was excluded because of a missing value for whether mother had a live-in partner at child age 3 years.

Families were assessed at child ages 2, 3, 4, 5, 7.5, 8.5, 9.5, 10.5, and 14 years, with retention exceeding 75% at all waves. As part of the trial, families were randomized at child age 2 years to be offered (or not offered) an annual family based intervention designed to prevent development of child conduct problems and subsequent early onset substance use (Family Check-Up; Dishion & Kavanagh, 2003). Randomization to intervention was found to reduce later maternal depression (Shaw et al., 2009) and child problem behavior (Dishion et al., 2014). In the current study, we included randomization to intervention among the covariates on which we equated the depressed versus nondepressed groups in order to account for its potential role as a confounding variable.

Measurement of Mothers' Depressive Symptoms at Child Age 3 Years

Mothers' depressive symptoms at child age 3 years were assessed via the Center for Epidemiological Studies on Depression Scale (CES-D; Radloff, 1977), a 20-item, self-report measure asking respondents to rate their depressive symptoms over the past week. The CES-D is a reliable and valid measure of depressive symptoms (Vilagut, Forero, Barbaglia, & Alonso, 2016; coefficient $\alpha = .91$ in the current sample). The total CES-D score is a sum potentially ranging from 0 to 60. Following published literature (Vilagut et al., 2016), total scores of 16 or greater were taken to indicate significant probability of major depressive disorder. This threshold was used to create two groups: mothers who exhibited clinically significant depressive symptoms at child age 3 years (42%) and mothers who did not (58%). The term "depression" will be used throughout this article to indicate a clinically significant level of depressive symptoms; "depression" should not be interpreted as meeting *DSM* 5 criteria for major depressive disorder.

Measurement of Covariates at Child Age 3 Years

Covariates for equating the depressed and nondepressed groups were drawn from the measures collected at the visit when the child was 3 years old. The mother completed 18 questionnaires, or self-report survey instruments. When available (63% of cases), the mother's live-in partner completed eight questionnaires. Study staff completed home environment inventories. Two additional covariates were collected after child age 3 years: polygenic scores indexing the child's genetic risk for (a) aggression and (b) internalizing problems. Both were based on the EAGLE Consortium's genome-wide association analyses (Pappa et al., 2016) and genotyping that occurred at child age 14 years.

Table S1 lists all the collected measures and the key reference for each. We reviewed all variables measured at child age 3 years for inclusion as a potential confounding variable upon which to equate the depressed and nondepressed groups. We selected confounding variables based on four criteria:

A. *The variable was conceptually distinguishable from the construct of depression*. We equated families with depressed versus nondepressed mothers on all potential confounding variables. Thus, it was important that the pool of potential confounding variables not include any constructs that are a core feature of the construct of depression (Miller & Chapman, 2001; Shadish, Cook, & Campbell, 2002). Equating the depressed and nondepressed groups on such a feature would render the residual difference between the depressed and nondepressed groups difficult to interpret.[2]

B. *The variable correlated with mothers' depression*. The variable must be associated with mothers' depression. Variables that do not relate to mothers' depression cannot confound the relation of mothers' depression to child behavior problems because there is no difference between depressed and nondepressed mothers on these variables.

C. *The variable predicted later child behavior outcomes*. The variable must predict later child behavior outcomes in order to potentially confound the relation of mothers' depression to child behavior problems. Because child outcomes were rated at multiple timepoints by multiple informants, we screened variables for prediction of any of these multiple measurements of child behavior problems.

D. *The variable was not itself measured after the measurement of mothers' depression*. Variables measured at a later wave might themselves have been affected by mothers' depression at child age 3 years. Adjusting for these postexposure variables (i.e. potential mediators or intermediate outcomes) may bias estimates of the causal effect of mothers' depression (Rosenbaum, 1984; Rubin, 2004). There was one exception to this criterion: Polygenic scores were included in the balancing pool because they reflect inborn characteristics that could not be changed by exposure to mothers' depression at child age 3 years.

We reviewed every collected variable (see Table S1) with these four criteria in mind and selected a total of 89 covariates upon which to equate the depressed and nondepressed groups at child age 3 years. Table 1 lists the variables and reports descriptive statistics. When possible, we favored summary scores rather than individual items (e.g., total count of behavior problems rather than indicators for each individual behavior problem). The 89 selected covariates spanned the domains of demographics (e.g., sex, race, ethnicity, income, marital status), areas of family strength (e.g., support from extended family), negative impact factors (e.g., re-

[2] For example, we identified family income as a potential confounding variable. We viewed the question, "Does later behavior differ in children of depressed vs. nondepressed mothers when the mothers are otherwise similar on family income?" as a meaningful comparison. Family income is not a part of the construct of depression, so we can conceptualize a depressed and a nondepressed mother with the same family income and consider that a meaningful comparison. In contrast, we excluded mothers' life satisfaction from the pool of potential confounding variables. We viewed the question, "Does later behavior differ in children of depressed vs. nondepressed mothers when the mothers are equivalent on life satisfaction?" as not being a meaningful comparison. Lower satisfaction with life is a defining part of the construct of depression. Equating the two groups on life satisfaction would fundamentally change the nature of the "depression" construct that remained as a difference between them.

Table 1
Descriptive Statistics for Covariates at Child Age 3 Years Upon Which the Depressed and Nondepressed Groups Were Equated

Variable	*N*	*M*	*SD*	Min	Max	SA#5
Binary variables						
Family randomized to intervention condition in trial	629	50%	—	—	—	
Site is Charlottesville, VA	629	27%	—	—	—	
Site is Pittsburgh, PA	629	37%	—	—	—	
M: Child is male	629	50%	—	—	—	
M: Child is Hispanic	627	13%	—	—	—	
M: Child is black	628	28%	—	—	—	
M: Child is biracial	628	14%	—	—	—	
M: Mother was teen parent	629	23%	—	—	—	
M: Mother is Hispanic	627	11%	—	—	—	
M: Mother is black	628	28%	—	—	—	
M: Mother is married	629	40%	—	—	—	
M: Mother has a live-in partner	629	63%	—	—	—	
M: Mother has religious / spiritual beliefs	629	69%	—	—	—	
M: Family is below poverty line	625	67%	—	—	—	
M: Family receives food stamps	629	61%	—	—	—	
M: Family receives medical assistance	629	69%	—	—	—	
M: Family receives social security income	629	13%	—	—	—	
M: Family receives child support	629	21%	—	—	—	
M: Family owns their home	627	19%	—	—	—	
M: Child has been cared for by person other than mother more than 5 hrs/wk	627	75%	—	—	—	
M: Person in home had trouble with law in past year	627	32%	—	—	—	
M: Person in home reported for child abuse in past year	629	7%	—	—	—	
M: Person in home treated by a mental health professional in past year	629	39%	—	—	—	
M: Support from extended family is a family strength	629	64%	—	—	—	
M: Employment situation is a family strength	629	29%	—	—	—	
M: Church, religion, or spirituality is a family strength	629	33%	—	—	—	
M: Conflict or violence has impacted family	629	12%	—	—	—	
M: Drug use by parent has impacted family	629	11%	—	—	—	
M: High crime neighborhood has impacted family	629	6%	—	—	—	
M: Parent being absent has impacted family	629	20%	—	—	—	
M: Stress between home and school has impacted family	629	13%	—	—	—	
M: Unstable home situation has impacted family	629	5%	—	—	—	
M: Death in family has impacted family	629	13%	—	—	—	
M: Past traumatic experience has impacted family	629	10%	—	—	—	
M: No organized groups are a source of support for mother	628	50%	—	—	—	*
M: Mother speaks with friends or family on phone 7 + times per week	629	45%	—	—	—	*
M: Mother has not visited friends in past week	629	26%	—	—	—	*
M: Mother ever drinks alcohol	629	76%	—	—	—	
M: Mother drinks alcohol at least monthly	629	35%	—	—	—	
M: Mother drinks alcohol at least weekly	629	11%	—	—	—	
M: Mother ever stopped drinking due to problems with use	617	5%	—	—	—	
M: Mother currently smokes cigarettes	626	44%	—	—	—	
M: Mother ever uses marijuana	626	12%	—	—	—	
M: Mother uses marijuana at least monthly	626	5%	—	—	—	
M: Mother ever stopped using marijuana due to problems with use	617	6%	—	—	—	
M: Mother ever uses hard drugs	629	6%	—	—	—	
LIP: Live-in partner ever drinks alcohol	216	81%	—	—	—	
LIP: Live-in partner drinks alcohol at least monthly	216	49%	—	—	—	
LIP: Live-in partner drinks alcohol at least weekly	216	20%	—	—	—	
LIP: Live-in partner ever stopped drinking due to problems with use	213	12%	—	—	—	
LIP: Live-in partner currently smokes cigarettes	216	44%	—	—	—	
LIP: Live-in partner ever uses marijuana	213	15%	—	—	—	
LIP: Live-in partner uses marijuana at least monthly	213	8%	—	—	—	
LIP: Live-in partner ever stopped using marijuana due to problems with use	210	10%	—	—	—	
Metric variables						
M: Child age in months	480	41.73	3.24	34	50	
M: Mother age in years at study entry	625	26.64	5.91	16	46	
M: Mother's level of education	629	5.25	1.15	2	9	
M: Household's gross monthly income from all sources	625	4.21	2.25	1	13	
M: Household's gross monthly income from all sources per person living in home	625	1.01	0.63	0.08	5.50	
M: Number of persons living in home	629	4.58	1.62	2	16	
M: Number of adults living in home	629	2.04	0.85	1	6	
M: Number of children living in home	629	2.55	1.26	0	10	

(*table continues*)

Table 1 (*continued*)

Variable	*N*	*M*	*SD*	Min	Max	SA#5
M: Rating of total chaos in home environment (CHAOS)	629	5.30	3.65	0	15	
M: Count of child's total number of behavior problems (ECBI)	622	14.47	7.81	0	36	
M: Rating of child total intensity of behavior problems (ECBI)	625	127.81	32.74	51	226	
M: Rating of child's externalizing behavior (CBCL)	629	0.74	0.33	0.00	1.75	
M: Rating of child's internalizing behavior (CBCL)	629	0.32	0.20	0.00	1.39	
M: Rating of child's attention problems (CBCL)	629	0.71	0.41	0	2	
M: Rating of child's sleep problems (CBCL)	629	4.21	2.84	0	14	
M: Rating of child's somatic complaints (CBCL)	629	0.20	0.19	0.00	1.18	
M: Rating of child's 'other' behavior problems (CBCL)	629	11.88	6.18	0	32	
M: Rating of child's inhibitory control (CBQ)	629	4.23	0.78	1.46	6.62	
M: Rating of positivity in mother-child relationship (ACRS)	629	8.63	3.17	5	23	*
M: Rating of conflict in mother-child relationship (ACRS)	629	26.63	7.89	10	48	*
M: Frequency of total daily hassles (HASSL)	629	44.77	8.76	23	77	*
M: Perception of total daily hassles (HASSL)	628	46.97	13.14	20	92	*
M: Rating of mother's parenting competency (BEPAR)	629	64.98	12.40	23	92	*
M: Rating of mother's parenting laxness (PARTS)	629	2.96	0.98	1	7	
M: Rating of mother's parenting overreactivity (PARTS)	629	2.74	0.80	1.00	5.40	
M: Rating of neighborhood cohesion (MMNQ)	627	14.81	7.83	5	35	
M: Rating of neighborhood danger (MMNQ)	628	7.33	7.24	0	37	
M: Rating of mother's relationship with live-in partner (LOCKE)	388	57.94	9.33	27	74	
M: Count of number of words understood by child (MACDI)	626	60.24	25.49	1	100	
LIP: Live-in partner's total depressive symptoms (CESD)	217	9.65	7.81	0	33	
LIP: Rating of child's externalizing behavior (CBCL)	215	0.57	0.32	0.00	1.54	
LIP: Rating of child's internalizing behavior (CBCL)	215	0.26	0.19	0.00	1.03	
HV: Rating of parent involvement during in-home observation (HOME)	615	2.13	0.97	0	3	
Eagle Consortium GWAS score for early childhood aggression	468	−0.01	3.85	−13.32	10.81	*
Eagle Consortium GWAS score for early childhood internalizing problems	468	0.06	4.25	−10.80	14.24	*

Note. GWAS = Genome Wide Association Study. Prefix of "M:" indicates mother reported the variable; prefix of "LIP:" indicates live-in partner reported the variable; prefix of "HV:" indicates home visitor reported the variable. Asterisk in column "SA5" indicates the variable was excluded from covariate pool in Sensitivity Analysis #5. Tags in parentheses (e.g., "ACRS") are acronyms indicating questionnaire from which score was computed (see Table S1 for key). Table S6 reports the measurement scale of each metric variable.

cent death in the family), child behavior (e.g., aggression, noncompliance, anxiety, sleep), neighborhood factors (e.g., danger, cohesion), parent functioning (e.g., substance use, frequency of contact with friends), and factors related to live-in partners (e.g., relationship satisfaction of mother, live-in partner's substance use). Following Rubin (2007, 2008), all decisions about which covariates to include were finalized before proceeding to analysis of child outcomes in order to protect against bias in the investigators' choices.

Measurement of Child Behavior Outcomes at Child Ages 4 to 14 Years

Child outcomes from ages 4 to 14 years were assessed using multiple measures via mother, secondary caregiver, and teacher report. Table S2 reports descriptive statistics for all outcome variables. Mothers completed the Child Behavior Checklist (CBCL; Achenbach & Rescorla, 2001) at child ages 4, 5, 7.5, 8.5, 9.5, 10.5, and 14 years. Secondary caregivers (when available) also completed the CBCL at child ages 4, 5, 7.5, 8.5, 9.5, 10.5, and 14 years. Most secondary caregivers (79% to 84%, across waves) lived with the child. Secondary caregivers included biological father (44%), grandmother (14%), mother's male boyfriend (13%), stepfather (9%), or aunt (5%). All remaining categories each comprised fewer than 2% of reports. Teachers completed the Teacher Report Form (TRF; Achenbach & Rescorla, 2001) at child ages 7.5, 8.5, 9.5, 10.5 years. Mean raw item responses on the externalizing and internalizing composites of the CBCL and TRF were analyzed as outcomes. On the preschool form administered at child age 4 years, the externalizing composite consisted of items measuring (a) aggressive behavior and (b) attention problems; the internalizing composite consisted of items measuring (a) anxious/depressed behavior, (b) emotionally reactive behavior, and (c) somatic complaints. On the school-age form administered at subsequent waves, the externalizing composite consisted of items measuring (a) aggressive behavior and (b) rule-breaking behavior; the internalizing composite consisted of items measuring (a) anxious/depressed behavior, (b) withdrawn/depressed behavior, and (c) somatic complaints.

Step 1: Equating Families With Depressed and Nondepressed Mothers at Child Age 3 Years

All analyses were completed in the R statistical software environment (v4.0; R Core Team, 2020). Step 1 required that we equate families with depressed versus nondepressed mothers at child age 3 years on all covariates. Step 1 involved three substeps: (a) estimating the propensity score for each case, (b) creating the inverse probability of treatment (IPT) weight for each case, and (c) checking whether the depressed and nondepressed groups were similar on all covariates after weighting the sample.

Estimating the propensity score. We fit a logistic regression model that predicted the log-odds of the mother being depressed at child age 3 years as a function of the first order effects of all 89 covariates. We used the "imputation with constant plus missingness indicators" method to address missing values in the covariates

(Cham & West, 2016). For each family, the propensity score was taken to be the predicted probability of the mother being depressed conditional on that family's values on the covariates.

Creating the inverse probability of treatment (IPT) weights. Each family received an IPT weight that was a function of the estimated propensity score. Before creating the weights, we verified that there was overlap in the estimated propensity scores of cases exposed versus not exposed to mothers' depression (a "common region of support"). If the groups do not overlap in estimated propensity score, a causal effect cannot be defined—if all cases in a region were assigned to be either exposed or unexposed, then there are no counterfactual outcomes to compare (Stuart, 2010; West et al., 2014). Families in the depressed group were weighted by [1/*pscore*]; families in the nondepressed group were weighted by [1/(1 - *pscore*)], where *pscore* is the propensity score. This weighting approach weights the sample to be representative of a population in which mothers' depression were randomly assigned (Rosenbaum, 1987).

Checking whether the depressed and nondepressed groups were similar after weighting the sample. If weighting were successful, then the families with depressed and nondepressed mothers would exhibit similar distributions on all 89 covariates after weighting. Following guidelines suggested by Rubin (2001), for metric variables we verified there were (a) no standardized mean differences (SMDs) greater than 0.20 standard deviation and (b) no variance ratios outside of 0.5 to 2. For binary variables, we verified that the rates of endorsement differed by no more than 5%. To calculate balance after weighting, we used weighted means, variances, and proportions (Austin & Stuart, 2015).

Step 2: Comparing Child Behavior Problems in the Equated Groups From Child Age 4 to 14 Years

Step 2 required that we compare the depressed and nondepressed groups on child behavior problems from child ages 4 to 14 years after IPT weighting.

Missing data. Mother-reported outcomes ranged from 9% to 25% missing, secondary-caregiver-reported outcomes ranged from 38% to 56% missing, and teacher-reported outcomes ranged from 45% to 59% missing (Table S2). Missing data were addressed using multiple imputation by chained equations (MICE; Raghunathan, Lepowski, van Hoewyk, & Solenberger, 2001) with the *mice* package (v3.9.0; van Buuren & Groothuis-Oudshoorn, 2011). MICE produces unbiased parameter estimates assuming the data are Missing at Random (MAR), conditional on all variables included in the imputation model (Rubin, 1976). To better satisfy the MAR assumption, our imputation model included both the variables in the subsequent analyses and a comprehensive set of auxiliary variables. 500 imputed data sets were created to minimize between-imputation error (Graham, 2009). See online supplemental material for more detail.

Estimating the effect of mothers' depression at child age 3 years. For each child outcome variable, we estimated two effects. First, the *prima facie* (i.e. unadjusted) effect was estimated simply by comparing families with depressed versus nondepressed mothers at child age 3 years. This comparison does not account for potential confounding variables. Second, the adjusted effect of mothers' depression was estimated by comparing families with depressed versus nondepressed mothers at child age 3 years after weighting the sample with the IPT weights. This comparison accounts for baseline differences on the potential confounding variables on which the groups were equated (see Table 1). In both cases, we regressed the outcome variable on a dummy variable indicating that the mother was depressed (vs. not depressed) at child age 3 years. The coefficient on the dummy variable indicated either the (a) *prima facie* effect or (b) adjusted effect of mothers' depression. Unweighted regressions were fit using the base R *glm* function; weighted regressions were fit in the *survey* package (Lumley, 2003). Estimated effects were converted to the metric of Cohen's *d* for ease of interpretation.

Sensitivity Analyses

Six sensitivity analyses probed whether the primary results were robust to changes in analytical approach. Five of the six sensitivity analyses yielded the same conclusions as the primary analysis; the sixth did not. These first five analyses are described only briefly here; the online supplemental material provides details on the rationale, method, and findings of each analysis.

In sensitivity analysis 1 [SA1], results were similar when equating the depressed and nondepressed groups using propensity score matching, an alternative method for adjusting for confounding variables. In SA2, results were similar when analyzing child outcomes using a repeated-measures model, which was expected to improve statistical power to detect an effect. In SA3, results were similar when measuring child outcomes using DSM symptoms of attention-deficit/hyperactivity disorder, oppositional defiant disorder, conduct disorder, separation anxiety disorder, generalized anxiety disorder, and major depressive disorder, rather than broadband rating scales. In SA4, results were similar when using more stringent thresholds on CES-D total score to define the depressed ($\geq$18) and nondepressed ($\leq$13) groups, increasing the separation between groups and thus the severity of exposure. In SA5, results were similar when excluding 10 covariates (see rightmost column of Table 1) upon which it might be controversial to equate the depressed and nondepressed groups at child age 3 years. Thus, the first five sensitivity analyses (SA1-SA5) all indicated that the primary findings were robust to changes in the analytical approach.

In contrast, the sixth sensitivity analysis did not yield the same conclusions as the primary result and so is described here in more detail (also see online supplemental material). In SA6, we equated the depressed and nondepressed groups on the same 89 covariates, now measured at child age 2 years instead of age 3 years. The purpose of this sensitivity analysis was to address concerns that covariates measured concurrently with mothers' depression at child age 3 years might themselves potentially be mediators of the causal effect on child behavior problems, such that adjusting for these variables "blocks" the mediation pathway and produces biased estimates of the causal effect. Nearly all theories of causation emphasize temporal precedence of causes (Reichardt, 2019; Shadish et al., 2002), but many investigators have treated concomitant relationships as if they were potentially causal. To address this alternative viewpoint, SA6 used only measurements of the covariates at child age 2 years. Because the covariates were measured 1 year prior to mothers' depression (at child age 3 years), this analysis has the advantage of ruling out the possibility that the measured covariates mediated the effect of mothers' depression at child age 3 years on later child behavior problems. However, equating the groups on covariates measured at child age 2 years was expected to leave substantial differences between the depressed and nondepressed groups on other factors at child age 3 years because the covariates do not have perfect temporal stability (i.e. differences between the two groups can arise in the interim year). Thus, SA6 was expected to remove only a portion of the potential confounding by covariates, place an upper limit on the magnitude of the causal effect in this sample, and appeal to readers who disagree with our thinking about when covariates should be measured.

Results

Step 1: Equating Families With Depressed and Nondepressed Mothers at Child Age 3 Years

IPT weights were created using the estimated propensity scores. The common region of support spanned from 0.07 and to 0.95 in the propensity score metric. There were no unexposed cases with estimated propensity scores exceeding 0.95, and there were no exposed cases with estimated propensity scores below 0.07. After discarding cases outside the common region of support, 463 families remained. IPT weights ranged from 1.06 to 13.43, with a median of 1.44 and an interquartile range of 1.20 to 2.08. Within the weighted sample, mean CES-D total score was 24.9 ($SD = 7.8$) in the depressed group and mean CES-D total score was 8.5 ($SD = 3.9$) in the nondepressed group. As described below, in the weighted sample, the depressed and nondepressed groups were successfully equated on all 89 covariates at child age 3 years.

Metric variables. Figure 1 shows the SMD between the depressed and nondepressed groups on the 39 metric covariates before weighting (open white circles) and after weighting (filled black circles). Prior to weighting, 24 of the 35 metric covariates exhibited SMDs greater than 0.20 *SD*. For example, the depressed group exhibited more home chaos (SMD = 0.59), more child internalizing problems (SMD = 0.69), and poorer relationships between mothers and their live-in partners (SMD = −0.44). After weighting, all 35 metric covariates exhibited SMDs of less than 0.12 *SD*. In addition, all metric covariates exhibited variance ratios between 0.5 and 2.0, with most falling between 0.8 and 1.4. After weighting, balance on squared covariate terms paralleled that obtained on main effects.

Binary variables. Figure 2 shows the proportions of respondents endorsing each of the 54 binary covariates in the depressed and nondepressed groups before weighting (white circles) and after weighting (black circles). Prior to weighting, the rates of endorsement differed by more than 5% for 33 of 54 binary covariates. For example, depressed mothers were more likely to be below the poverty line (76% vs. 60%) and were less likely to be married (34% vs. 44%). After weighting, the rates of endorsement differed by less than 5% for all 54 covariates.

Step 2: Comparing Child Behavior Problems in the Equated Groups From Child Age 4 to 14 Years

In Step 2, we compared the depressed versus nondepressed mothers at child age 3 years on later child behavior problems. Comparisons were made both before and after weighting to equate the groups on the 89 covariates. “Unadjusted” *prima facie* effects refer to simple comparisons of families with depressed versus nondepressed mothers, without consideration of potential confounding variables. “Adjusted” effects refer to comparisons that adjust for confounding variables using IPT weighting. Table S5 reports all estimated effect sizes and associated 95% confidence intervals.

Figure 3 depicts the overall pattern of results. Each panel depicts the mean effect size across ages for a specific combination of measure (child externalizing or internalizing problems and informant being mother, secondary caregiver, or teacher). Equating the depressed and nondepressed groups at baseline (i.e. weighting) reduced the mean effect size in all cases. Below, we consider effects first for externalizing behavior then for internalizing behavior.

Externalizing behavior. Figure 4 shows the effects of mothers’ depression at child age 3 years on child externalizing behavior at child ages 4 through 14 years. Based on mother report of externalizing behavior (Panels A and B), the unadjusted *prima facie* effects ranged from $d = 0.28$ to 0.56, with a mean value across ages of $d = 0.36$. The *prima facie* effect was statistically significant at all ages ($p < .05$). After weighting, effects ranged from $d = 0.01$ to 0.16, with a mean value of $d = 0.06$. The weighted effect was not statistically significant at any age (*ns*).

Based on secondary caregiver report of externalizing behavior (Panels C and D), the unadjusted *prima facie* effects ranged from $d = 0.17$ to 0.35, with a mean value across ages of $d = 0.24$. The *prima facie* effect was statistically significant at all ages ($p < .05$). After weighting, effects ranged from $d = -0.08$ to 0.12, with a mean value of $d = 0.01$. The weighted effect was not statistically significant at any age (*ns*).

Based on teacher report of externalizing behavior (Panels E and F), the unadjusted *prima facie* effects ranged from $d = 0.06$ to 0.17, with a mean value across ages of $d = 0.12$. The *prima facie* effect was not statistically significant at any age (*ns*). After weighting, effects ranged from $d = -0.07$ to 0.20, with a mean value of $d = 0.05$. The weighted effect was not statistically significant at any age (*ns*).

Internalizing behavior. Figure 5 shows the effects of mothers’ depression at child age 3 years on child internalizing behavior at ages 4 through 14 years. Based on mother report of internalizing behavior (Panels A and B), the unadjusted *prima facie* effects of mothers’ depression at child age 3 years ranged from $d = 0.41$ to 0.65, with a mean value across ages of $d = 0.53$. The *prima facie* effect was statistically significant at all ages ($p < .05$) except age 5 and 14 years (*ns*). After weighting, effects ranged from $d = -0.01$ to 0.20, with a mean value of $d = 0.07$. The weighted effect was not statistically significant at any age (*ns*).

Based on secondary caregiver report of internalizing behavior (Panels C and D), the unadjusted *prima facie* effects of mothers’ depression at child age 3 years ranged from $d = 0.07$ to 0.43, with a mean value across ages of $d = 0.28$. The *prima facie* effect was statistically significant ($p < .05$) at all child ages except ages 5 and 14 years (*ns*). After weighting, effects ranged from $d = -0.20$ to 0.24, with a mean value of $d = 0$. The weighted effect was not statistically significant at any age (*ns*).

Based on teacher report of internalizing behavior (Panels E and F), the unadjusted *prima facie* effects of mothers’ depression at child age 3 years ranged from $d = 0.07$ to 0.22, with a mean value across ages of $d = 0.14$. The *prima facie* effect was not statistically

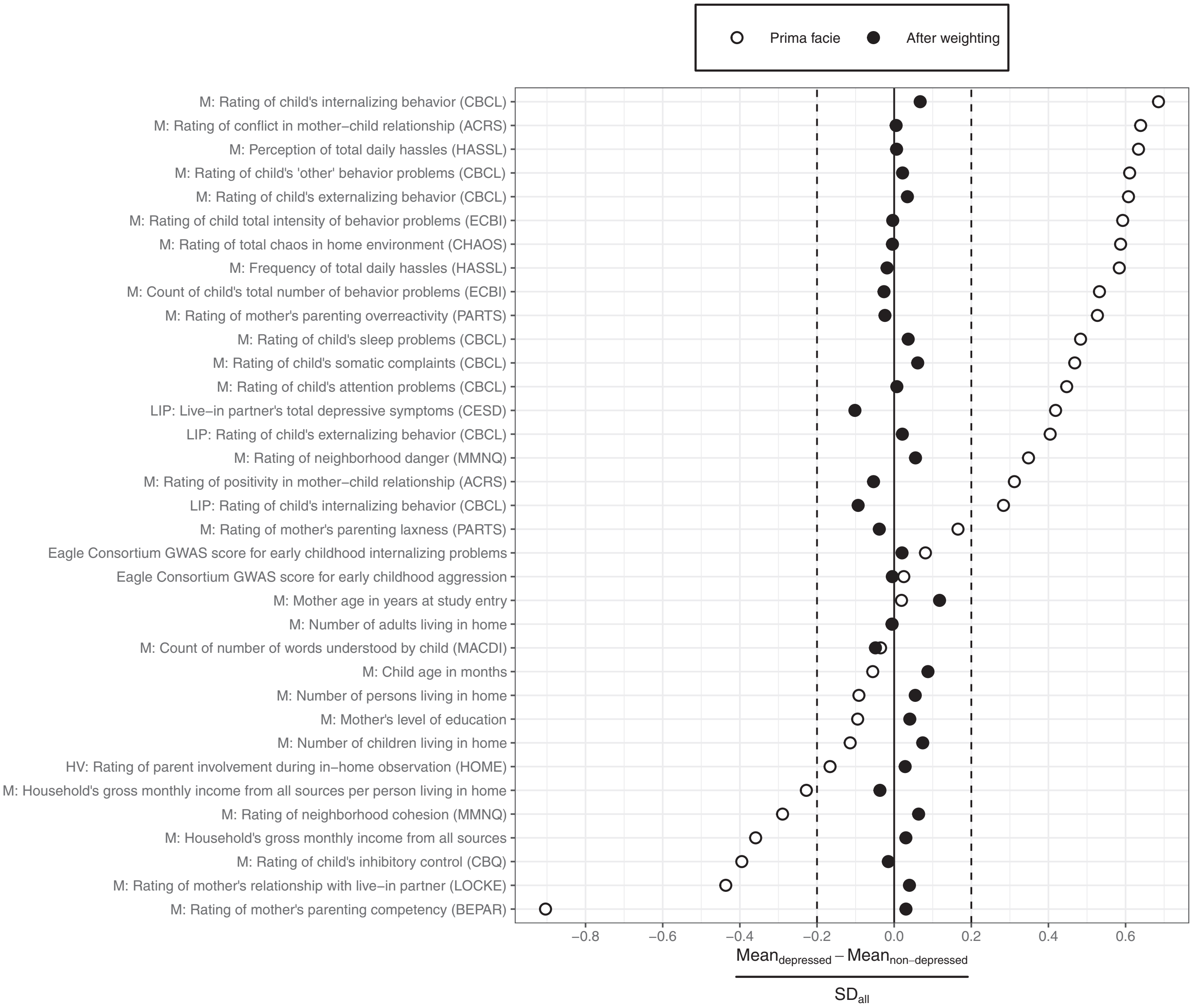

Figure 1. Balance on means of metric covariates at child age 3 years before and after weighting. Prefix of "M:" indicates mother reported the variable; prefix of "LIP:" indicates live-in partner reported the variable; prefix of "HV:" indicates home visitor reported the variable. Variables are sorted from highest to lowest standardized mean difference (SMD) prior to weighting. Open circles to the left or right of the band indicated by vertical dashed lines indicate variables with substantial (|SMD| > 0.20) imbalance between depressed and nondepressed groups before weighting. Filled circles are all within the band indicated by vertical dashed lines, confirming that depressed and nondepressed groups have similar means on these covariates after weighting. Exact values are reported in Table S3. Acronyms in parentheses after variable names indicate measure from which score was calculated (see Table S1).

significant at any age (*ns*). After weighting, effects ranged from $d = -0.11$ to 0.11, with a mean value of $d = 0$. The weighted effect was not statistically significant at any age (*ns*).

Sensitivity Analysis #6

(See online supplemental material for sensitivity analyses 1 through 5.) As expected, adjusted effects were larger when we adjusted only for covariates measured at child age 2 years (i.e. 1 year prior to exposure to mothers' depression at child age 3). Figure S4 depicts the overall pattern of results, which were similar for child externalizing and internalizing behavior. Based on mother report, the mean effect after weighting was $d = 0.27$ versus the mean *prima facie* effect of $d = 0.45$ (attenuation of 40%). Based on secondary caregiver report, the mean effect after weighting was $d = 0.17$ versus the mean *prima facie* effect of $d = 0.26$ (attenuation of 36%). Based on teacher report, the mean effect after weighting was $d = 0.127$ versus the mean *prima facie* effect of $d = 0.129$ (attenuation of 1%). Adjusted effects based on mother and secondary caregiver report were statistically significant ($p <$

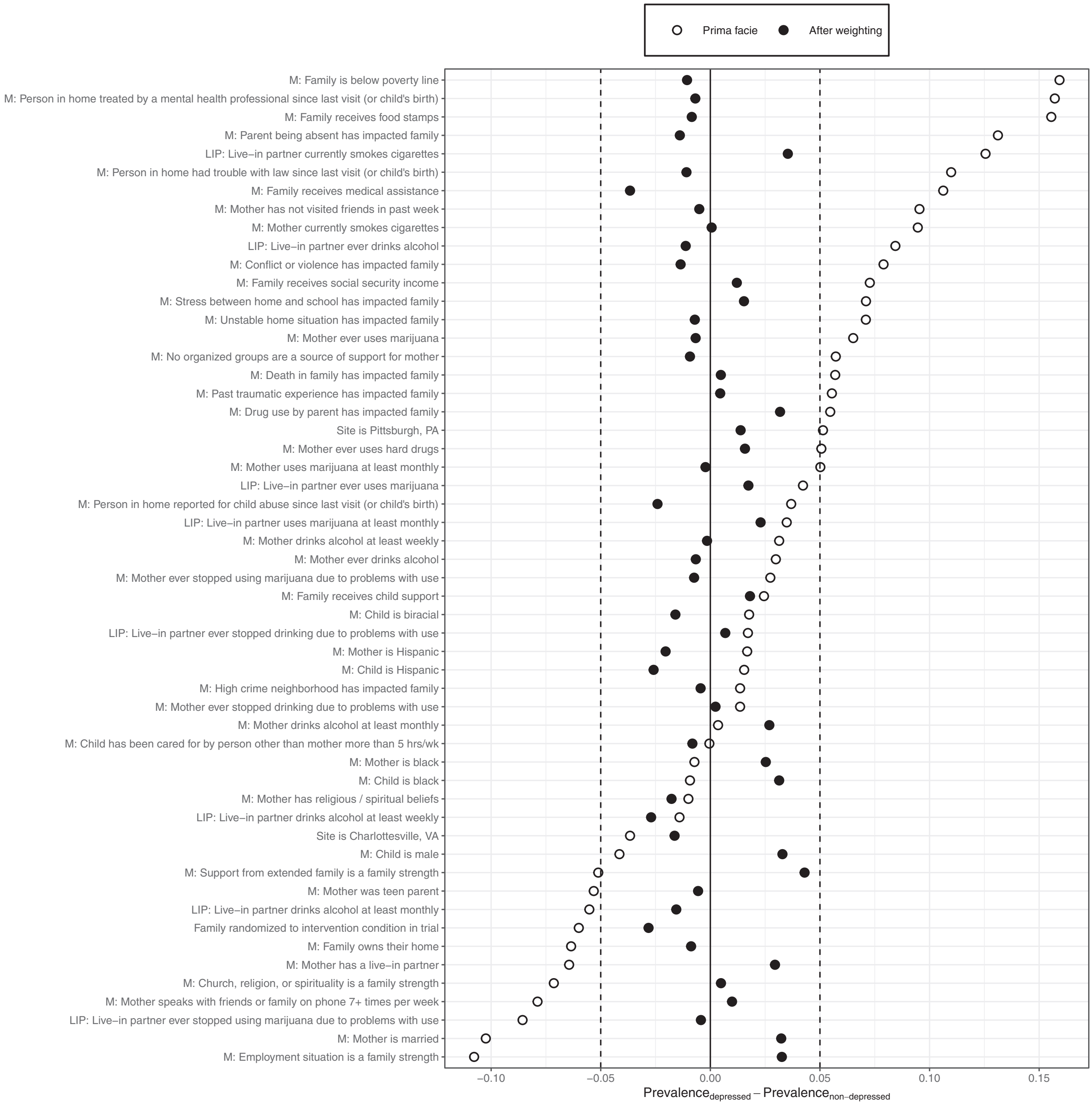

Figure 2. Balance on prevalences of binary covariates at child age 3 years before and after weighting. Prefix of “M:” indicates mother reported the variable; prefix of “LIP:” indicates live-in partner reported the variable; prefix of “HV:” indicates home visitor reported the variable. Variables are sorted from highest to lowest difference in prevalence prior to weighting. Open circles to the left or right of the band indicated by vertical dashed lines indicate variables with more than a 5-percentage-point difference in prevalence in the depressed and nondepressed groups before weighting. Filled circles are all within the band indicated by vertical dashed lines, confirming that depressed and nondepressed groups have similar prevalences of these covariates after weighting. Exact values are reported in Table S3.

.05) in repeated-measures models; adjusted effects based on teacher report were not statistically significant (*ns*).

As expected, even after weighting using propensity scores based on covariates measured at child age 2 years, there remained many substantial differences between the depressed and nondepressed groups on potential confounders measured at child age 3 years. In other words, the depressed and nondepressed groups were only partially equated. Of 35 metric covariates (Figure S5), there re-

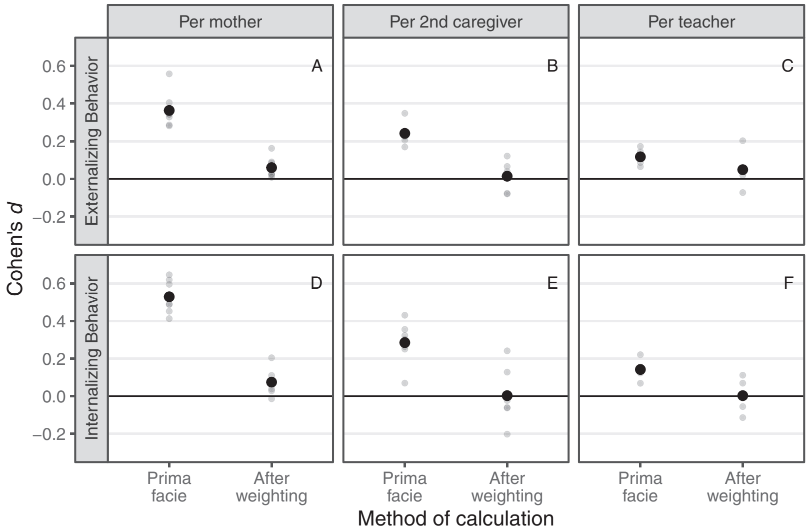

Figure 3. Mean effects of mothers' depression at child age 3 years on later child behavior problems. Large black dots indicate mean effect size across age for each combination of outcome, informant, and method of calculation. Lighter, smaller black dots indicate individual effect sizes contributing to each mean and indicate the variability across the age range of the study.

mained 21 with SMDs of more than 0.20 after weighting (e.g., $d = 0.50$ for mother's rating of child externalizing problems). Of 54 binary covariates (Figure S6), there remained 14 with differences in prevalence greater than 5% after weighting (e.g., 76% of depressed mothers were below poverty line vs. 65% of nondepressed mothers). Thus, the adjusted effects estimated in SA6 reduced but did not eliminate the potential for covariates to be alternative explanations for later differences in child behavior problems, substantially weakening an interpretation of the estimates as causal effects of mothers' depression.

Discussion

Hundreds of published studies have documented that the children of depressed mothers exhibit more externalizing and internalizing behavior problems. Yet this work has failed to isolate the effect of mothers' depression from the effects of the many co-occurring risk factors, leaving the true cause of children's negative outcomes unclear. We attempted to estimate the specific causal effect of mothers' depression at child age 3 years by applying propensity score methods to data from the Early Steps Multisite Trial. Mothers with clinically-elevated versus non-clinically-elevated symptoms of depression (as measured by the CES-D) at child age 3 years were equated on 89 other factors, including mother, child, and family characteristics. Prior to equating, the *prima facie* effect of mothers' depression at child age 3 years on child behavior problems from ages 4 to 14 years was in the small to medium range (mean $d = 0.30$) and most effects were statistically significant (26 of 36 measures). After equating, the adjusted effect was in the very small to small range (mean $d = 0.04$) and never statistically significant (0 of 36 measures). Results were robust to sensitivity analyses varying the extremity of mothers' depression, the measurement of child outcomes, the covariates included, and aspects of the analytic approach. Findings suggest that a substantial portion of the *prima facie* association between mothers' depression and later child behavior problems is accounted for by confounding variables rather than a causal effect of the mothers' depressive symptoms per se.

Prima Facie Effects of Mothers' Depression

There were robust *prima facie* effects of mothers' depression on subsequent child behavior problems. The magnitude of *prima facie* effects paralleled those obtained in the most recent meta-analysis (Goodman et al., 2011). Consider the first available follow-up for each outcome, which is most comparable to the follow-up interval of published studies. For mother-reported externalizing behavior, we estimated $d = 0.56$ compared to $d = 0.47$ in Goodman et al. (2011). For mother-reported internalizing behavior, we estimated $d = 0.62$ compared to $d = 0.52$ in Goodman et al. (2011). For teacher-reported externalizing behavior, we estimated $d = 0.14$ compared to $d = 0.28$ in Goodman et al. (2011). For teacher-reported internalizing behavior, we estimated $d = 0.12$ compared to $d = 0.30$ in Goodman et al. (2011). Thus, before accounting for confounding variables, the Early Steps sample exhibited the typical association between mothers' depression and child behavior problems.

Adjusted Effects of Mothers' Depression

The *prima facie* effects do not account for the many other differences between families with depressed versus nondepressed mothers (i.e. potential confounding variables). We estimated adjusted effects after equating the families on 89 baseline covariates (see Table 1) using IPT weighting. Adjusted effects were smaller than *prima facie* effects in nearly all cases—the mean effect size was reduced by 85% for mother report, 97% for secondary caregiver report, and 80% for teacher report. Although no adjusted

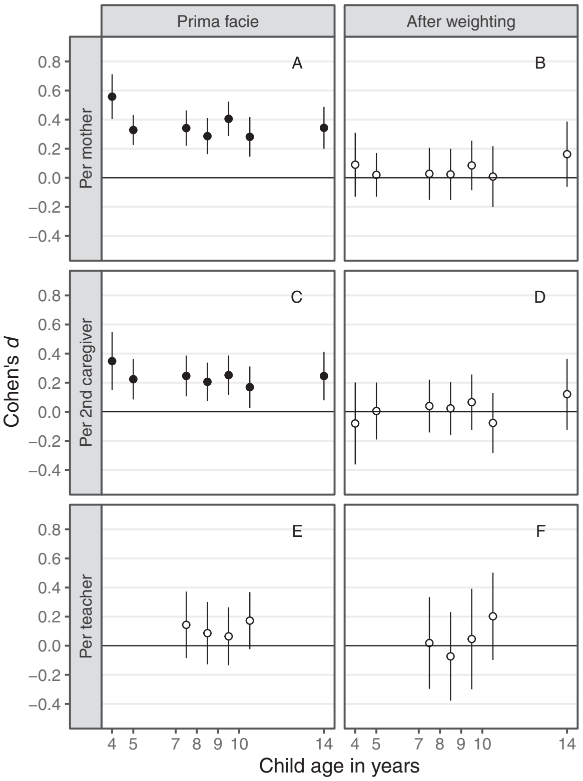

Figure 4. Effects of mothers' depression at child age 3 years on later child externalizing behavior. Filled circles indicate effects that are statistically significant ($p < .05$); open circles indicate effects that are not statistically significant (*ns*). Error bars indicate 95% confidence intervals.

effect was statistically significant, most (25 of 36) indicated more behavior problems in the children of depressed mothers than in the children of nondepressed mothers.

Residual confounding. Although our baseline measurement of covariates was comprehensive, we did not measure all potential confounders. There may remain important determinants of child behavior upon which the depressed and nondepressed groups differed even after equating (i.e. weighting). For example, we had no information about mothers' depression earlier in the child's life (e.g., at child age 1 year) or when the child was in utero, yet both perinatal and postnatal exposure to depression are associated with increased risk of behavioral disturbance (Goodman & Gotlib, 1999; Goodman & Halperin, 2020; Stein et al., 2014). Thus, even after weighting, children of the depressed mothers may have been exposed to more prenatal or postnatal mothers' depression than were children of nondepressed mothers. If such unmeasured covariates affect children's behavior, then properly accounting for residual confounding on these measures might be expected to further attenuate the adjusted effect of mothers' depression at child age 3 years.[3]

Families excluded from estimation of the adjusted effect. Approximately one-quarter of the sample was excluded from the adjusted analysis because their estimated propensity scores fell outside the region of common support between those exposed versus not exposed to mothers' depression. There were no depressed mothers with very low ($<$0.07) estimated probability of being depressed; there were no nondepressed mothers with very high ($>$0.95) estimated probability of being depressed. Conceptually, mothers falling outside the region of common support are mothers for whom no counterfactual comparison can be made—the sample did not contain mothers with similar values across the covariates yet with opposite levels of depression. Thus, the estimated adjusted effects apply to mothers with reasonable probability (i.e., $0.07 < \textit{probability} < 0.95$) of being either depressed or nondepressed.[4]

Nature of the Adjusted Effects and Cautions Against Misinterpretation and Overgeneralization

Hundreds of studies have shown that the children of depressed mothers have more behavior problems. The primary contribution of this study was its attempt to move beyond association. Our design asked the following question: Do later behavior problems differ in children with depressed versus nondepressed mothers at child age 3 years if the families are initially similar on current income, child aggression, neighborhood danger, and the rest of the 89 covariates listed in Table 1?[5]

Overall, results were consistent with the belief there is a very small to small causal effect of exposure to a mother with clinically-elevated symptoms of depression at child age 3 years on children's later broadband externalizing and internalizing behavior problems. Yet, no single study can "prove" or "disprove" causality (Hill, 1965). Data were not truly experimental, and causal inference from nonrandomized data is a difficult task that merits cautious interpretation. While our goal was to estimate causal effects, this does not guarantee that we succeeded in doing so. Giving our adjusted effects a causal interpretation depends on the soundness of our analytic approach and the extent to which we met the causal assumptions. Readers must judge for themselves the extent to which we have successfully produced a statistical comparison that merits a causal interpretation. To aid in the reader's judgment, we made a transparent comparison of the depressed and nondepressed groups after equating (Figure 1, Figure 2, Table S3, Table S4) and conducted six different sensitivity analyses that probed how our analytical choices might have affected results. The reader may choose to reject our causal framing and interpret the findings simply as comparing unadjusted versus adjusted associations between mothers' depression and child behavior problems.

[3] Sensitivity analysis for the effects of unobserved confounders (which we refer to here as residual confounding) is an important component of a nonrandomized study when a substantial and/or statistically significant effect is found (Liu, Kuramoto, & Stuart, 2013; Rosenbaum, 1986). We do not report a sensitivity analysis for unobserved confounding in this article because the adjusted effect sizes for mothers' depression were very small and not statistically significant. Because residual confounding would be expected to further attenuate the effect, a sensitivity analysis was not indicated.

[4] The exclusion of cases outside the common region of support does not explain the overall pattern of results. The difference in mothers' depression remained large ($d = 1.40$) even after excluding these cases for the weighted analysis. In addition, the matched analysis yielded slightly larger causal effects despite excluding a considerably larger portion of the sample.

[5] Each *prima facie* or adjusted effect represents the total effect of the mothers' depression at child age 3 years on a child behavior outcome, including both the direct effect and potential indirect effects through all mediators.

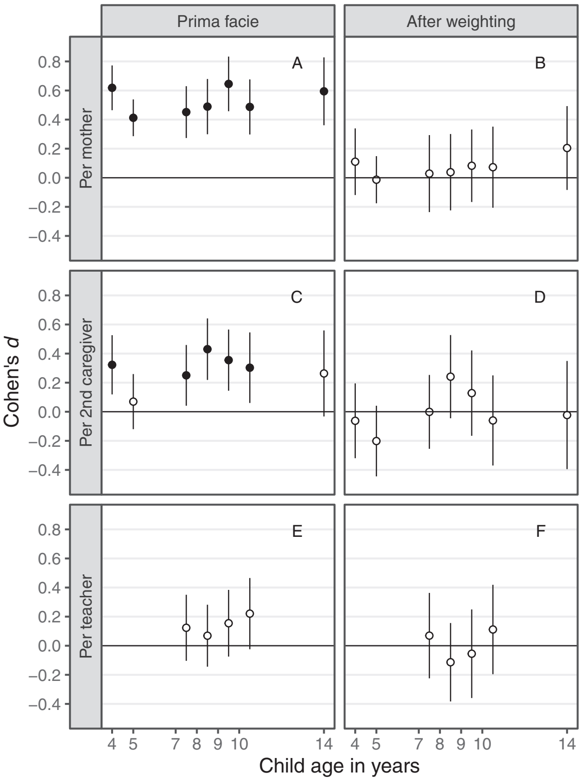

Figure 5. Effects of mothers' depression at child age 3 years on later child internalizing behavior. Filled circles indicate effects that are statistically significant ($p < .05$); open circles indicate effects that are not statistically significant (*ns*). Error bars indicate 95% confidence intervals.

The conceptual framework used in this study—emulating a hypothetical randomized trial to estimate the causal effect of a specific exposure—has been widely applied in statistics, economics, political science, medicine, and epidemiology, but has a limited history of applications in developmental psychopathology. Thus, it is unclear whether the attenuation of effect we observed after adjusting for confounding variables (a) is unique to mothers' depression or (b) would also be observed with other commonly studied single risk factors (e.g., poverty, marital conflict). Applications of propensity score techniques in other areas of developmental psychopathology have at times found substantial attenuation of the *prima facie* effect (e.g., Steinberg & Monahan, 2011) and at times found little attenuation (e.g., Odgers et al., 2008). As more applications and theoretical analyses accumulate, they may reveal weaknesses of the present analysis or reveal unique difficulties or limitations in these methods' application to developmental psychopathology. We caution readers to avoid the simplistic and premature takeaway that "maternal depression does not matter for child behavior problems."

In addition, note that our findings do not contradict (and in fact replicate) the fact that children of depressed mothers exhibit more externalizing and internalizing behavior problems. The children of depressed mothers displayed substantially more behavior problems on nearly all measurements in this large, ethnically diverse sample (see panels A, C, and E in Figures 4 and 5), with effect sizes ranging up to a mean difference of 0.65 standard deviations. Rather, our findings suggest (but not prove) that mothers' depression is not the primary cause of this discrepancy. Replications and extensions are needed to verify this conclusion in larger samples; with different measurement and timing of mothers' depression, confounders, and child outcomes; across different developmental epochs; and with different quasi-experimental designs. Only with additional work on this topic will we be able to fully evaluate the validity of the current findings and the limits on their generalizability.

Developmental Timing

Mothers' depression was measured when the child was approximately 3 years old ($M = 41.7$ months, $SD = 3.2$ months). Goodman and Gotlib (1999) outlined several reasons to expect that exposure to mothers' depression during early childhood will produce larger effects than exposure when children are older. Mothers play a larger role in regulating their children's emotions when they are younger, negative effects incurred at a younger age may have more time to "snowball" into cumulative deficits, and toddlers have less ability than older children to escape exposure to depressed mothers' affect and behavior and seek out alternative sources of support. Meta-analyses confirm that the association of mothers' depression with child behavior is stronger in younger children (Connell & Goodman, 2002; Goodman et al., 2011). Cognitive–behavioral therapy for mothers' depression during pregnancy has stronger impact on children's behavior when outcome variables are measured at a younger age (Goodman, Cullum, Dimidjian, River, & Kim, 2018). Thus, the developmental timing of exposure to mothers' depression in this study (i.e. child age 3 years) would be expected to produce larger observed effect sizes than had children been exposed at later ages.

Measurement of Mothers' Depression

Findings must be understood in the context of how mothers' depression was measured (Hernán, 2005; VanderWeele & Hernán, 2013). We attempted to estimate the causal effect of being exposed to a mother with CES-D total score ≥16 (vs. ≤15) at a single measurement when the child was 3 years old. Sensitivity analyses yielded similar results when comparing mothers ≤13 versus ≥18 in CES-D total score.

We measured mothers' depression using a self-report rating scale (the CES-D), not a DSM-based diagnostic interview. 33% of the mothers screened for study entry had CES-D total scores ≥16, a prevalence considerably higher than that expected for Major Depressive Disorder (e.g., past-year rate among mothers of 10.2%; Ertel, Rich-Edwards, & Koenen, 2011). Thus, our "depressed" group likely included mothers with milder forms of depression in addition to mothers who would have met diagnostic criteria for Major Depressive Disorder (Coyne, 1994; Joiner, Walker, Pettit, Perez, & Cukrowicz, 2005; Ruscio, 2019). The most severely impairing, debilitating forms of depression (e.g., recurrent, prolonged MDD) may well have larger negative effects on children than the effects observed in this study. In addition, although we measured mothers' depression at a single timepoint, chronic exposure across childhood may have greater negative impact.

However, there were clear reasons to expect that mothers' depression as measured in the present study would show effects on child mental health. First, a meta-analysis indicates that the effect of mothers' depression is similar when measured via symptom rating scales (like the CES-D) versus clinical diagnoses (Goodman et al., 2011). Second, the difference in the mean CES-D total score between the depressed and nondepressed groups after weighting was quite large at child age 3 years (d = 1.40) and the mean difference remained substantial at all later ages (ds ranging from 0.21 to 0.64; M = 0.49; Figure S7). Thus, although mothers' depression was measured at child age 3 years using a symptom rating scale asking about the past week, children in the depressed and nondepressed groups received substantially different levels of exposure over the duration of their childhood. Replicating findings in samples with different timing, severity, and duration of exposure to mothers' depression is a clear priority for future work.

Nature of Sample

The sample was high-risk: families in the WIC Nutritional Supplement program who also possessed multiple risk factors for child conduct problems (Dishion et al., 2008). Mothers are eligible for the WIC program when they are pregnant or have a child up to age 5 years, report gross income at or below 185% of the U.S. Poverty Income Guidelines, and are deemed to exhibit nutritional risk. The effect of mothers' depression is larger in younger children, in low-income families, and in single-parent families (Goodman et al., 2011). Thus, families in the WIC program would be expected to show stronger effects of mothers' depression on child behavior problems than families in the general population. Moreover, enrolled families had multiple risk factors for child conduct problems, increasing risk beyond that of the general WIC population. For example, when compared to families screened-out during recruitment, screened-in families reported lower monthly income (∼ $1,700 vs. ∼ $1,960) and the mothers were more likely to be unmarried (36% vs. 45%). Thus, the published literature suggests that the current sample might be expected to show stronger effects of mothers' depression than the general population.

Strengths and Limitations

Strengths of the current study include its large sample size (N = 629), use of multiple raters of children's behavior problem outcomes (mother, secondary caregiver, and teacher), prospective design, and extended period of follow-up (eight measurements over the span of 12 years). Most importantly, data collection included a comprehensive set of covariates that were related to both the mothers' depression status and the children's later mental health outcomes. This feature enabled us to adjust for 89 variables that might otherwise confound the relation of mothers' depression to subsequent child behavior problems, including rarely available factors such as polygenic scores indexing risk for child behavior problems. Finally, the depressed and nondepressed groups were equated and compared in local conditions (i.e. drawn from same sample, measured at the same time, completed an identical battery of covariates), a feature of nonrandomized studies that produce estimates closer to the true causal effect (Cook, Shadish, & Wong, 2008; Cook, Zhu, Klein, Starkey, & Thomas, 2020).

The greatest limitation of the current study, the measurement of mothers' depression, was previously discussed. Other limitations arise from the measurement of child outcomes. First, secondary caregivers were a heterogeneous group and likely possessed mixed ability to accurately rate child behavior (e.g., 16% to 20% were not living with the child at the time of reporting). Second, there were substantial missing data on the teacher report of child behavior (45% to 59%), limiting statistical power to detect effects. Third, because mothers' depression was measured when the child was 3 years old, some outcomes (e.g., teacher report, *DSM* symptoms) could not be measured until several years later, at which point effects may have been attenuated. Finally, outcomes were limited to the constructs of externalizing and internalizing behavior problems. Although meaningful and commonly studied, these constructs do not capture the full range of vulnerability to the development of psychopathology.

Future Directions

We have already discussed one priority for future work: replicating this design in samples including children with exposure to more severe, more sustained maternal depression. Another priority for future work is to explore moderators and mediators of the causal effect of exposure to mothers' depression on children's behavior problems, which may differ substantially from what has been found to moderate or mediate the association. Perhaps the mean causal effect is very small because only a subset of children experiences negative effects of exposure to mothers' depression. For example, a mean causal effect of d = 0.04 could be produced by 10% of children experiencing an effect of d = 0.40 and 90% of children experiencing an effect of d = 0 (i.e. the weighted average would be d = 0.04). Probing possible diathesis-stress models will require the fusion of data across multiple studies or the use of large national, archival data sets to achieve the necessary statistical power to detect a moderated effect of this magnitude.

Implications

When the causal effect of mothers' depressive symptoms was isolated from those of many co-occurring clinical features, it was much smaller than the *prima facie* association. It appears that other co-occurring factors that are conceptually distinct from depression are responsible for a large part of the observed difference in the outcomes of children with depressed versus nondepressed mothers. To fully understand why children of depressed mothers exhibit more behavior problems, a multicausal theory is needed that jointly considers the cluster of co-occurring clinical features that often accompany mothers' depression (Kendler, 2019).

Failing to account for the many ways in which depressed and nondepressed mothers differ may compromise our understanding of how mothers' depression functions in complex developmental systems. For example, mothers' depression may appear to statistically moderate the effect of another factor on child behavior problems, when in fact it simply co-occurs with the variable that causally moderates the relation (e.g., marital conflict). Similarly, mothers' depression may appear to mediate the effect of another factor on child behavior problems, when in fact it simply serves as a proxy for other methods of transmission.

The finding of very small causal effects suggests that, despite the moderate to strong *prima facie* association, reductions in mothers' depressive symptoms may not in and of themselves lead to reduced child behavior problems. Meta-analyses have found reductions in child behavior problems following randomization to cognitive–behavioral therapy for mothers' depression, but tests of whether these effects are mediated by reductions in mothers' depressive symptoms have yielded inconsistent results (Cuijpers, Weitz, Karyotaki, Garber, & Andersson, 2015; Goodman et al., 2018; Gunlicks & Weissman, 2008). One explanation for that inconsistency is that cognitive–behavioral therapy for mothers' depression affects multiple factors that influence child behavior, only one of which is mothers' depressive symptoms. This hypothesis is consistent with the current findings. Thus, integrated interventions that directly target both mothers' depression and children's behavior problems may be necessary to address their comorbidity (Goodman & Garber, 2017).

Results also suggest that studies seeking to explore the causal effect of mothers' depression on child behavior problems will require very large samples sizes. The mean causal effect in this study was $d = 0.04$. If the true causal effect of mothers' depression were $d = 0.04$, more than 12,700 cases would be required to detect this effect with 80% power using a two-sample t test. Even if the true causal effect were $d = 0.20$, more than 750 cases would be required. As these samples sizes are not common in the field of developmental psychopathology (Reardon, Smack, Herzhoff, & Tackett, 2019), investigators interested in causal pathways that include the link of mothers' depression to child behavior problems should focus on design and analysis techniques that can improve statistical power.

Finally, findings underscore the value of causal inference methods (e.g., IPT weighting, propensity score matching) for addressing questions in developmental psychopathology, a field in which experimentation is often difficult or impossible. The association between mothers' depression and child behavior problems is robust and has been observed consistently across hundreds of studies (Goodman et al., 2011). If replications confirm that such a well-established association is mostly explained by confounding variables, what might we find when applying similar designs to other developmental phenomena? If developmental psychopathologists can increase the correspondence between the estimated statistical effects and the true causal effects, their theories will become more accurate (Ohlsson & Kendler, 2020).

Conclusion

This study found that a substantial portion of the *prima facie* effect of mothers' depression (as measured by clinically-elevated symptoms on the CES-D) on children's behavior problems was explained by co-occurring clinical features rather than mothers' depressive symptoms per se. Nonetheless, findings were consistent with the belief that there is a very small to small mean causal effect of exposure to a mother with clinically-elevated depressive symptoms at child age 3 years on later child externalizing and internalizing behavior. This study ($N = 629$) was underpowered to detect effect sizes of this magnitude, and adjusted effects were generally not statistically significant even when potentially clinically meaningful (e.g., $d = 0.20$). Again, we caution readers to avoid the premature takeaway that "maternal depression does not matter for child behavior problems." This study was a first step—more work is needed to replicate, contextualize, and extend these findings and enhance our understanding the causal effects of exposure to mothers' depression during childhood.

References

Achenbach, T. M., & Rescorla, L. (2001). *Manual for the ASEBA school-age forms and profiles*. Burlington: University of Vermont, Research Center for Children, Youth & Families.

Ahern, J. (2018). Start with the "c-word," follow the roadmap for causal inference. *American Journal of Public Health, 108,* 621. http://dx.doi.org/10.2105/AJPH.2018.304358

Arnold, D. S., O'Leary, S. G., Wolff, L. S., & Acker, M. M. (1993). The parenting scale: A measure of dysfunctional parenting in discipline situations. *Psychological Assessment, 5,* 137–144. http://dx.doi.org/10.1037/1040-3590.5.2.137

Arriaga, R. I., Fenson, L., Cronan, T., & Pethick, S. J. (1998). Scores on the MacArthur communicative development inventory of children from low and middle-income families. *Applied Psycholinguistics, 19,* 209–223. http://dx.doi.org/10.1017/S0142716400010043

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46,* 399–424. http://dx.doi.org/10.1080/00273171.2011.568786

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine, 34,* 3661–3679. http://dx.doi.org/10.1002/sim.6607

Bagner, D. M., Pettit, J. W., Lewinsohn, P. M., & Seeley, J. R. (2010). Effect of maternal depression on child behavior: A sensitive period? *Journal of the American Academy of Child & Adolescent Psychiatry, 49,* 699–707. http://dx.doi.org/10.1097/00004583-201007000-00010

Beardslee, W. R., Bemporad, J., Keller, M. B., & Klerman, G. L. (1983). Children of parents with major affective disorder: A review. *The American Journal of Psychiatry, 140,* 825–832. http://dx.doi.org/10.1176/ajp.140.7.825

Boggs, S. R., Eyberg, S., & Reynolds, L. A. (1990). Concurrent validity of the Eyberg Child Behavior Inventory. *Journal of Clinical Child Psychology, 19,* 75–78. http://dx.doi.org/10.1207/s15374424jccp1901_9

Bullock, B., & Dishion, T. J. (2004). *Family Affective Attitude Rating Scale (FAARS)*. Eugene, OR: Child and Family Center.

Caldwell, B. M., & Bradley, R. H. (2003). *HOME inventory administration manual*. Little Rock: University of Arkansas for Medical Sciences and University of Arkansas at Little Rock.

Cham, H., & West, S. G. (2016). Propensity score analysis with missing data. *Psychological Methods, 21,* 427–445. http://dx.doi.org/10.1037/met0000076

Child and Family Center. (2001). *CFC parent questionnaire: Service utilization*. Eugene, OR: Child and Family Center.

Choe, D. E., Shaw, D. S., Brennan, L. M., Dishion, T. J., & Wilson, M. N. (2014). Inhibitory control as a mediator of bidirectional effects between early oppositional behavior and maternal depression. *Development and Psychopathology, 26,* 1129–1147. http://dx.doi.org/10.1017/S0954579414000613

Cicchetti, D. (1984). The emergence of developmental psychopathology. *Child Development, 55,* 1–7. http://dx.doi.org/10.2307/1129830

Cicchetti, D. (1993). Developmental psychopathology: Reactions, reflections, projections. *Developmental Review, 13,* 471–502. http://dx.doi.org/10.1006/drev.1993.1021

Conduct Problems Prevention Research Group. (1999). Initial impact of the fast track prevention trial for conduct problems: I. The high-risk sample. *Journal of Consulting and Clinical Psychology, 67,* 631–647. http://dx.doi.org/10.1037/0022-006X.67.5.631

Connell, A. M., & Goodman, S. H. (2002). The association between psychopathology in fathers versus mothers and children's internalizing and externalizing behavior problems: A meta-analysis. *Psychological Bulletin, 128,* 746–773. http://dx.doi.org/10.1037/0033-2909.128.5.746

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27,* 724–750. http://dx.doi.org/10.1002/pam.20375

Cook, T. D., Zhu, N., Klein, A., Starkey, P., & Thomas, J. (2020). How much bias results if a quasi-experimental design combines local comparison groups, a pretest outcome measure and other covariates? A within study comparison of preschool effects. *Psychological Methods*. Advance online publication. http://dx.doi.org/10.1037/met0000260

Coyne, J. C. (1994). Self-reported distress: Analog or ersatz depression? *Psychological Bulletin, 116,* 29–45. http://dx.doi.org/10.1037/0033-2909.116.1.29

Crnic, K. A., & Greenberg, M. T. (1990). Minor parenting stresses with young children. *Child Development, 61,* 1628–1637. http://dx.doi.org/10.2307/1130770

Crnic, K. A., Greenberg, M. T., Ragozin, A. S., Robinson, N. M., & Basham, R. B. (1983). Effects of stress and social support on mothers and premature and full-term infants. *Child Development, 54,* 209–217. http://dx.doi.org/10.2307/1129878

Cuijpers, P., Weitz, E., Karyotaki, E., Garber, J., & Andersson, G. (2015). The effects of psychological treatment of maternal depression on children and parental functioning: A meta-analysis. *European Child & Adolescent Psychiatry, 24,* 237–245. http://dx.doi.org/10.1007/s00787-014-0660-6

Dishion, T. J., Brennan, L. M., Shaw, D. S., McEachern, A. D., Wilson, M. N., & Jo, B. (2014). Prevention of problem behavior through annual family check-ups in early childhood: Intervention effects from home to early elementary school. *Journal of Abnormal Child Psychology, 42,* 343–354. http://dx.doi.org/10.1007/s10802-013-9768-2

Dishion, T. J., & Kavanagh, K. (1997). *FAST: Family assessment task*. Eugene: Child and Family Center, University of Oregon.

Dishion, T. J., & Kavanagh, K. (2003). *Intervening in adolescent problem behavior: A family-centered approach*. New York, NY: Guilford Press.

Dishion, T. J., Shaw, D., Connell, A., Gardner, F., Weaver, C., & Wilson, M. (2008). The family check-up with high-risk indigent families: Preventing problem behavior by increasing parents' positive behavior support in early childhood. *Child Development, 79,* 1395–1414. http://dx.doi.org/10.1111/j.1467-8624.2008.01195.x

Downey, G., & Coyne, J. C. (1990). Children of depressed parents: An integrative review. *Psychological Bulletin, 108,* 50–76. http://dx.doi.org/10.1037/0033-2909.108.1.50

Ertel, K. A., Rich-Edwards, J. W., & Koenen, K. C. (2011). Maternal depression in the United States: Nationally representative rates and risks. *Journal of Women's Health, 20,* 1609–1617. http://dx.doi.org/10.1089/jwh.2010.2657

Foster, E. M. (2010). Causal inference and developmental psychology. *Developmental Psychology, 46,* 1454–1480. http://dx.doi.org/10.1037/a0020204

Freedman, D. A. (1987). As others see us: A case study in path analysis. *Journal of Educational Statistics, 12,* 101–128. http://dx.doi.org/10.3102/10769986012002101

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7,* 457–472. http://dx.doi.org/10.1214/ss/1177011136

Goodman, S. H. (2020). Intergenerational transmission of depression. *Annual Review of Clinical Psychology, 16,* 213–238. http://dx.doi.org/10.1146/annurev-clinpsy-071519-113915

Goodman, S. H., Cullum, K. A., Dimidjian, S., River, L. M., & Kim, C. Y. (2018). Opening windows of opportunities: Evidence for interventions to prevent or treat depression in pregnant women being associated with changes in offspring's developmental trajectories of psychopathology risk. *Development and Psychopathology, 30,* 1179–1196. http://dx.doi.org/10.1017/S0954579418000536

Goodman, S. H., & Garber, J. (2017). Evidence-based interventions for depressed mothers and their young children. *Child Development, 88,* 368–377. http://dx.doi.org/10.1111/cdev.12732

Goodman, S. H., & Gotlib, I. H. (1999). Risk for psychopathology in the children of depressed mothers: A developmental model for understanding mechanisms of transmission. *Psychological Review, 106,* 458–490. http://dx.doi.org/10.1037/0033-295X.106.3.458

Goodman, S. H., & Halperin, M. S. (2020). Perinatal depression as an early stress: Risk for the development of psychopathology in children. In K. Harkness & E. P. Hayden (Eds.), *The Oxford Handbook of Stress and Mental Health* (pp. 287–312). London, UK: Oxford University Press.

Goodman, S. H., Rouse, M. H., Connell, A. M., Broth, M. R., Hall, C. M., & Heyward, D. (2011). Maternal depression and child psychopathology: A meta-analytic review. *Clinical Child and Family Psychology Review, 14,* 1–27. http://dx.doi.org/10.1007/s10567-010-0080-1

Gotlib, I. H., Goodman, S. H., & Humphreys, K. L. (2020). Studying the intergenerational transmission of risk for depression: Current status and future directions. *Current Directions in Psychological Science, 29,* 174–179. http://dx.doi.org/10.1177/0963721420901590

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576. http://dx.doi.org/10.1146/annurev.psych.58.110405.085530

Greenland, S., & Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology, 15,* 413–419. http://dx.doi.org/10.1093/ije/15.3.413

Gunlicks, M. L., & Weissman, M. M. (2008). Change in child psychopathology with improvement in parental depression: A systematic review. *Journal of the American Academy of Child & Adolescent Psychiatry, 47,* 379–389. http://dx.doi.org/10.1097/CHI.0b013e3181640805

Hernán, M. A. (2005). Invited commentary: Hypothetical interventions to define causal effects—Afterthought or prerequisite? *American Journal of Epidemiology, 162,* 618–620. http://dx.doi.org/10.1093/aje/kwi255

Hernán, M. A. (2018). The c-word: Scientific euphemisms do not improve causal inference from observational data. *American Journal of Public Health, 108,* 616–619. http://dx.doi.org/10.2105/AJPH.2018.304337

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if?* Boca Raton, FL: Chapman & Hall/CRC.

Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine, 58,* 295–300. http://dx.doi.org/10.1177/003591576505800503

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, 42,* 1–28. http://dx.doi.org/10.18637/jss.v042.i08

Højsgaard, S., Halekoh, U., & Yan, J. (2005). The R package geepack for generalized estimating equations. *Journal of Statistical Software, 15,* 1–11.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139025751

Ingoldsby, E. M., & Shaw, D. S. (2002). Neighborhood contextual factors and early-starting antisocial pathways. *Clinical Child and Family Psychology Review, 5,* 21–55. http://dx.doi.org/10.1023/A:1014521724498

Johnston, C., & Mash, E. J. (1989). A measure of parenting satisfaction and efficacy. *Journal of Clinical Child Psychology, 18,* 167–175. http://dx.doi.org/10.1207/s15374424jccp1802_8

Joiner, T. E., Walker, R. L., Pettit, J. W., Perez, M., & Cukrowicz, K. C. (2005). Evidence-based assessment of depression in adults. *Psychological Assessment, 17,* 267–277. http://dx.doi.org/10.1037/1040-3590.17.3.267

Kendler, K. S. (2019). From many to one to many—The search for causes of psychiatric illness. *Journal of the American Medical Association*

Psychiatry, 76, 1085. http://dx.doi.org/10.1001/jamapsychiatry.2019.1200

Liu, W., Kuramoto, S. J., & Stuart, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention Science, 14,* 570–580. http://dx.doi.org/10.1007/s11121-012-0339-5

Locke, H. J., & Wallace, K. M. (1959). Short marital-adjustment and prediction tests: Their reliability and validity. *Marriage & Family Living, 21,* 251–255. http://dx.doi.org/10.2307/348022

Lumley, T. (2003). Analysis of complex survey samples. *Journal of Statistical Software, 9,* 1–19.

Matheny, A. P., Jr., Wachs, T. D., Ludwig, J. L., & Phillips, K. (1995). Bringing order out of chaos: Psychometric characteristics of the confusion, hubbub, and order scale. *Journal of Applied Developmental Psychology, 16,* 429–444. http://dx.doi.org/10.1016/0193-3973(95)90028-4

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology, 110,* 40–48. http://dx.doi.org/10.1037/0021-843X.110.1.40

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Morgan, S. L., & Winship, C. (2014). *Counterfactuals and causal inference* (2nd ed.). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781107587991

O'Brien Caughy, M. O., Randolph, S. M., & O'Campo, P. J. (2002). The Africentric home environment inventory: An observational measure of the racial socialization features of the home environment for African American preschool children. *Journal of Black Psychology, 28,* 37–52. http://dx.doi.org/10.1177/0095798402028001003

Odgers, C. L., Caspi, A., Nagin, D. S., Piquero, A. R., Slutske, W. S., Milne, B. J., . . . Moffitt, T. E. (2008). Is it important to prevent early exposure to drugs and alcohol among adolescents? *Psychological Science, 19,* 1037–1044. http://dx.doi.org/10.1111/j.1467-9280.2008.02196.x

Ohlsson, H., & Kendler, K. S. (2020). Applying causal inference methods in psychiatric epidemiology: A review. *Journal of the American Medical Association Psychiatry, 77,* 637. http://dx.doi.org/10.1001/jamapsychiatry.2019.3758

Pappa, I., St Pourcain, B., Benke, K., Cavadino, A., Hakulinen, C., Nivard, M. G., . . . Tiemeier, H. (2016). A genome-wide approach to children's aggressive behavior: The EAGLE consortium. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 171,* 562–572. http://dx.doi.org/10.1002/ajmg.b.32333

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. New York, NY: Basic Books.

Pelham, W. E., Gnagy, E. M., Greenslade, K., & Milich, R. (1992). Teacher ratings of *DSM–III–R* symptoms for the disruptive behavior disorders. *Journal of the American Academy of Child & Adolescent Psychiatry, 31,* 210–218. http://dx.doi.org/10.1097/00004583-199203000-00006

Phinney, J. S. (1992). The multigroup ethnic identity measure: A new scale for use with diverse groups. *Journal of Adolescent Research, 7,* 156–176. http://dx.doi.org/10.1177/074355489272003

Pianta, R. C. (1995). *Child-parent relationship scale*. Charlottesville: University of Virginia.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1,* 385–401. http://dx.doi.org/10.1177/014662167700100306

Raghunathan, T. E., Lepowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27,* 85–95.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reardon, K. W., Smack, A. J., Herzhoff, K., & Tackett, J. L. (2019). An N-pact factor for clinical psychological research. *Journal of Abnormal Psychology, 128,* 493–499. http://dx.doi.org/10.1037/abn0000435

Reichardt, C. S. (2019). *Quasi-experimentation: A guide to design and analysis*. New York, NY: Guilford Press Publications.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society Series A (General), 147,* 656–666. http://dx.doi.org/10.2307/2981697

Rosenbaum, P. R. (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics, 11,* 207–224. http://dx.doi.org/10.3102/10769986011003207

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association, 82,* 387–394. http://dx.doi.org/10.1080/01621459.1987.10478441

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

Rothbart, R. M., Ahadi, S. A., Hershey, K. L., & Fisher, P. (2001). Investigations of temperament at three to seven years: The children's behavior questionnaire. *Child Development, 72,* 1394–1408. http://dx.doi.org/10.1111/1467-8624.00355

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–592. http://dx.doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley. http://dx.doi.org/10.1002/9780470316696

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2,* 169–188. http://dx.doi.org/10.1023/A:1020363010465

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics, 31,* 161–170. http://dx.doi.org/10.1111/j.1467-9469.2004.02-123.x

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association, 100,* 322–331. http://dx.doi.org/10.1198/016214504000001880

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine, 26,* 20–36. http://dx.doi.org/10.1002/sim.2739

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics, 2,* 808–840. http://dx.doi.org/10.1214/08-AOAS187

Ruscio, A. M. (2019). Normal versus pathological mood: Implications for diagnosis. *Annual Review of Clinical Psychology, 15,* 179–205. http://dx.doi.org/10.1146/annurev-clinpsy-050718-095644

Rutter, M. (2007). Proceeding from observed correlation to causal inference: The use of natural experiments. *Perspectives on Psychological Science, 2,* 377–395. http://dx.doi.org/10.1111/j.1745-6916.2007.00050.x

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shaffer, D., Fisher, P., Lucas, C. P., Dulcan, M. K., & Schwab-Stone, M. E. (2000). NIMH diagnostic interview schedule for children version IV (NIMH DISC-IV): Description, differences from previous versions, and reliability of some common diagnoses. *Journal of the American Academy of Child & Adolescent Psychiatry, 39,* 28–38. http://dx.doi.org/10.1097/00004583-200001000-00014

Shaw, D. S., Connell, A., Dishion, T. J., Wilson, M. N., & Gardner, F. (2009). Improvements in maternal depression as a mediator of intervention effects on early childhood problem behavior. *Development and Psychopathology, 21,* 417–439. http://dx.doi.org/10.1017/S0954579409000236

Sroufe, L. A., & Rutter, M. (1984). The domain of developmental psychopathology. *Child Development, 55,* 17–29. http://dx.doi.org/10.2307/1129832

Stein, A., Pearson, R. M., Goodman, S. H., Rapa, E., Rahman, A., McCallum, M., . . . Pariante, C. M. (2014). Effects of perinatal mental disorders on the fetus and child. *The Lancet, 384,* 1800–1819. http://dx.doi.org/10.1016/S0140-6736(14)61277-0

Steinberg, L., & Monahan, K. C. (2011). Adolescents' exposure to sexy media does not hasten the initiation of sexual intercourse. *Developmental Psychology, 47,* 562–576. http://dx.doi.org/10.1037/a0020613

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25,* 1–21. http://dx.doi.org/10.1214/09-STS313

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45,* 1–67. http://dx.doi.org/10.18637/jss.v045.i03

VanderWeele, T. J., & Hernán, M. A. (2013). Causal inference under multiple versions of treatment. *Journal of Causal Inference, 1,* 1–20. http://dx.doi.org/10.1515/jci-2012-0002

Vilagut, G., Forero, C. G., Barbaglia, G., & Alonso, J. (2016). Screening for depression in the general population with the center for epidemiologic studies depression (CES-D): A systematic review with meta-analysis. *PLoS ONE, 11,* e0155431. http://dx.doi.org/10.1371/journal.pone.0155431

West, S. G., Cham, H., Thoemmes, F., Renneberg, B., Schulze, J., & Weiler, M. (2014). Propensity scores as a basis for equating groups: Basic principles and application in clinical treatment outcome research. *Journal of Consulting and Clinical Psychology, 82,* 906–919. http://dx.doi.org/10.1037/a0036387

Received May 11, 2019
Revision received July 27, 2020
Accepted July 29, 2020 ■