Bowling Green State University

The Center for Family and Demographic Research

http://www.bgsu.edu/organizations/cfdr

Phone: (419) 372-7279 cfdr@bgsu.edu

2017 Working Paper Series

PARENTAL INCARCERATION AND CHILD WELL-BEING: CONCEPTUAL AND PRACTICAL CONCERNS REGARDING THE USE OF PROPENSITY SCORES

Jennifer E. Copp[1]

Peggy C. Giordano[2]

Wendy D. Manning[2]

Monica A. Longmore[2]

[1] College of Criminology and Criminal Justice
Florida State University
Tallahassee, FL 32309

[2] Department of Sociology &
Center for Family and Demographic Research
Bowling Green State University
Bowling Green, OH

*This research was supported by grants from The Eunice Kennedy Shriver National Institute of Child Health and Human Development (HD036223, HD044206, and HD66087), the National Institute of Justice, Office of Justice Programs, U. S. Department of Justice (Award Nos. 2009-IJ-CX-0503 and 2010-MU-MU-0031), and in part by the Center for Family and Demographic Research, Bowling Green State University, which has core funding from The Eunice Kennedy Shriver National Institute of Child Health and Human Development (R24HD050959). The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the Department of Justice or the official views of the National Institutes of Health. Direct correspondence to Jennifer E. Copp, College of Criminology and Criminal Justice, Florida State University, Tallahassee, FL 32306 jcopp@fsu.edu.

PARENTAL INCARCERATION AND CHILD WELL-BEING: CONCEPTUAL AND PRACTICAL CONCERNS REGARDING THE USE OF PROPENSITY SCORES

Jennifer E. Copp[1]

Peggy C. Giordano[2]

Wendy D. Manning[2]

Monica A. Longmore[2]

[1] College of Criminology and Criminal Justice

Florida State University

Tallahassee, FL 32309

[2] Department of Sociology

and Center for Family and Demographic Research

Bowling Green State University

Bowling Green, OH 43403

*This research was supported by grants from The Eunice Kennedy Shriver National Institute of Child Health and Human Development (HD036223, HD044206, and HD66087), the National Institute of Justice, Office of Justice Programs, U. S. Department of Justice (Award Nos. 2009-IJ-CX-0503 and 2010-MU-MU-0031), and in part by the Center for Family and Demographic Research, Bowling Green State University, which has core funding from The Eunice Kennedy Shriver National Institute of Child Health and Human Development (R24HD050959). The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the Department of Justice or the official views of the National Institutes of Health. Direct correspondence to Jennifer E. Copp, College of Criminology and Criminal Justice, Florida State University, Tallahassee, FL 32306 jcopp@fsu.edu.

ABSTRACT

Objectives: The aim of the current investigation was to examine the appropriateness of propensity score methods for the study of incarceration effects on children by directing attention to a range of conceptual and practical concerns, including the exclusion of theoretically meaningful covariates, the comparability of treatment and control groups, and potential ambiguities resulting from researcher-driven analytic decisions.

Methods: Using data from the Toledo Adolescent Relationships Study and the Fragile Families and Child Wellbeing Study we examined offspring outcomes in the context of parental incarceration. We identified a roster of potential confounders, including factors motivated by prior research and additional selection forces. We examined the propensity score distributions and treatment effects, and conducted a number of sensitivity checks. We also estimated multilevel propensity score models to further address the issue of bias under unconfoundedness.

Results: Propensity score analyses of non-experimental data can provide results similar to those of randomized experiments. Valid estimates, however, rely on key assumptions which have been largely overlooked in existing research. In addition to these conceptual concerns, we found that propensity scores and treatment effect estimates are highly sensitive to a number of decisions made by the researcher, including aspects where little consensus exists.

Conclusions: Researchers should carefully consider the suitability of propensity score methods to estimate the effect of parental incarceration on children's outcomes in light of the conceptual underpinnings of propensity score analysis and existing data limitations. We discuss the utility of different identification methods and specialized data collection efforts.

Keywords: *parental incarceration; child wellbeing; propensity score matching; selection*

1. INTRODUCTION

The rapid increase in the U.S. incarceration rate over the last several decades has moved researchers and policymakers to consider its effect on families and communities, and in particular, the impact on children of incarcerated parents (see Travis, Western, & Redburn, 2014 for a review). According to recent estimates, more than half of U.S. prisoners—including roughly 52% of state inmates and 63% of federal inmates—are parents of minor children (Glaze & Maruschak, 2008; The Pew Charitable Trusts, 2010). This corresponds to 2.3% of all children in the U.S. under age 18; however, the number of children touched by parental incarceration is much larger as these figures capture only those with a parent currently serving time. Thus, whether the incarceration of a parent has collateral consequences for the next generation has become a particularly important empirical question, and one that has received considerable research attention over the past several years.

Parental incarceration is associated with children's problem outcomes across a broad range of domains, including mental and physical health, behavioral problems, academic achievement, material hardship, and involvement with the criminal justice system (Cho, 2009a, 2009b; Foster & Hagan, 2015; Haskins, 2014; Murray, Farrington, & Sekol, 2012; Murray, Loeber, & Pardini, 2012; Turney; 2017; Turney & Wildeman, 2015; Wildeman, 2010; Wildeman & Andersen, 2016). Beyond effects for the individual child, scholars suggest that parental incarceration has contributed to racial inequality in child well-being given the disproportionate impact of incarceration on individuals and families of color (Lee, McCormick, Hicken, & Wildeman, 2015; Wakefield & Wildeman, 2011, 2013). Yet despite the apparent convergence of research evidence indicating that parental incarceration has consequences for children's health and development, methodological and conceptual concerns related to selection bias continue to call into question the causal nature of incarceration effects on children (Giordano & Copp, 2015; Hagan & Dinovitzer, 1999; Johnson & Easterling, 2012; Murray, Farrington, Sekol, & Olsen, 2009; Sampson, 2011).

In light of the field's growing sensitivity to concerns about selection bias, scholars have looked to an array of advanced statistical techniques. Given the complexities involved in isolating the effect of incarceration while simultaneously accounting for other sources of adversity (Johnson & Easterling, 2012; Wildeman, Wakefield, & Turney, 2013), propensity score analysis has emerged as a leading methodological approach within the incarceration effects tradition. Yet the proliferation of propensity score analyses has occurred without full consideration of its appropriateness for the estimation of incarceration effects on children, and often with little attention to its underlying assumptions (Loughran, Wilson, Nagin, & Piquero, 2015). This is potentially problematic as the use of such methods has become the standard upon which the scientific rigor of incarceration effects research is assessed (Wildeman et al., 2013), and moreover, findings from this work have made their way into important policy discussions.

The current study builds on existing work emphasizing the need for researchers to exercise caution when using propensity score methods (Loughran et al., 2015; Shadish, 2013) by providing an empirical demonstration of the use of propensity scores to estimate the effect of parental incarceration on child well-being. We begin with a discussion of the counterfactual framework at the core of propensity scores as well as important assumptions of this methodological approach. Next, we consider the suitability of propensity score analysis for the estimation of incarceration effects in particular, focusing attention on the conceptual underpinnings of propensity score methods. Finally, we provide an empirical demonstration of propensity scores using data from the Toledo Adolescent Relationships Study (TARS) and the

Fragile Families and Child Wellbeing Study (FFCWB) to underscore the potential ambiguities present in propensity score analysis. Our findings question the use of propensity score models to address the effect of parental incarceration on child wellbeing, and provide future directions for scholars committed to further developing our understanding of the mechanisms driving problem outcomes among the children of incarcerated parents.

2. THE COUNTERFACTUAL FRAMEWORK AND THE STRONG IGNORABILITY ASSUMPTION OF PROPENSITY SCORE ANALYSIS

Given the practical, and obvious ethical, constraints to implementing randomized clinical trials in social science research—and incarceration effects research in particular—propensity score modeling has emerged as a common approach used to remove selection bias from observational data. In addition, propensity score analysis is used to estimate causal effects. This is accomplished by implementing a counterfactual approach to consider unobserved outcomes or what *would have happened* in the absence of a particular cause (Guo & Fraser, 2010; Morgan & Winship, 2014; Shadish, 2002). For example, using the case of parental incarceration, researchers often consider what would have happened to the children of incarcerated parents had they not experienced the incarceration event. Within this framework, individuals have two potential outcomes, including one for each value of the treatment (i.e., potential outcome if parent *is* incarcerated and potential outcome if parent *is not* incarcerated). However, only one of these outcomes is observed for each individual, and the unobserved outcome is referred to as the "counterfactual" outcome. Estimation of the treatment effect is obtained following examination of the difference between the average outcomes of the treatment and control groups, all else being equal. This difference represents the extent to which the observed differences between the treatment and control groups can be attributed to some intervention (e.g., parental incarceration). Given that we can never observe both potential outcomes for any given individual, the so-called "fundamental problem of causal inference" (Holland, 1986) comes down to one of missing data. As a result, researchers produce estimates based on potential outcomes, and therefore, must rely on certain assumptions in order to infer causality.

In observational studies, unbiased estimates hinge upon meeting the strong ignorability assumption. Treatment assignment is said to be strongly ignorable when the potential outcomes are jointly independent of treatment assignment, after adjusting for covariates (i.e., conditional independence) (Rosenbaum & Rubin, 1983). Researchers are commonly of the understanding that tests of balance are, in part, tests of strong ignorability. Yet the conditional independence component of strong ignorability imposes a much more stringent requirement, as it calls for all variables that influence treatment and outcomes to be included in the matching (Smith, 2000; Smith & Todd, 2005). Implicit in this assumption is the idea that treatment selection is based exclusively on observable characteristics and, therefore, strong ignorability requires the availability of rich, high-quality data. A second, and commonly overlooked, condition of strong ignorability is nonzero probability (Rosenbaum & Rubin, 1983). The basic idea of nonzero probability is that all individuals (including members of both the treatment and control conditions) have at least *some* chance of receiving the treatment. This suggests that propensity score methods are incompatible with certain research questions, such as the measurement of effects where the potential outcome is zero in "virtually all practical situations" (Shadish, 2013: p. 134).

Causal inference requires that treatment assignment be strongly ignorable in order to produce unbiased estimates using propensity scores. In practice, however, the ignorable treatment assignment assumption is often violated in research using observational data (Guo & Fraser, 2014). Research produced on the consequences of parental incarceration has devoted particularly little attention to this critical assumption. This is perhaps unsurprising as we rarely encounter elaborate discussions of the assumptions underlying more traditional modeling strategies. A key difference, however, is that propensity score methods are used to detect causal effects, and not merely to describe associations. Additionally, propensity score analysis is considered among the more rigorous approaches within the incarceration effects literature, and thus is increasingly being adopted by researchers examining the consequences of parental incarceration for child well-being. Propensity score analysis can be a powerful analytic tool, however, it is important that researchers carefully consider its suitability given the research question and available data. Furthermore, it is imperative that scholars direct attention to the issue of strong ignorability, as this assumption is routinely violated and/or given short shrift in existing empirical work. In the following sections, we continue our discussion of key assumptions and other important considerations of propensity score analysis, directing attention to existing research applications within the incarceration effects tradition.

3. THE APPLICATION OF PROPENSITY SCORE METHODS TO INCARCERATION EFFECTS RESEARCH

In the incarceration effects literature, scholars have long recognized that the men and women who go to prison differ in ways that not only affect their likelihood of experiencing incarceration, but also influence their familial relationships and family socioeconomic well-being (see Murray & Farrington, 2008). It is therefore quite difficult to determine whether the problem outcomes observed among the children of incarcerated parents are due to the incarceration itself, or these other sources of adversity (e.g., Giordano, 2010; Johnson & Easterling, 2012). Arguably the best way to address the selection bias endemic to research on incarceration effects would be to conduct an experiment; however, as this is generally not feasible, the use of statistical techniques that approximate experimental designs and produce estimates of incarceration effects have proliferated. Among these, propensity score methods are by far the most common.

So how do propensity score methods overcome the problem of selection bias (and satisfy the strong ignorability assumption)? In theory, researchers account for strong ignorability by identifying a set of covariates that characterize the selection process. With respect to incarceration effects research, these include factors associated with incarceration and/or child well-being outcomes. Individuals are then matched based on this vector of covariates with members of the treatment and control groups set to differ only based on their exposure to the treatment (i.e., parental incarceration). Although there exist no tests to confirm whether strong ignorability holds, it is reasonable to expect some discussion of the basis upon which one can assume strong ignorability given the observed covariates. With few exceptions, this assumption is rarely discussed in the literature, and where it is mentioned, scholars often sidestep much deliberation of the issue by indicating that propensity score models do not account/correct/adjust for unobserved heterogeneity.

The problem of unobserved heterogeneity is universal to propensity scores, and consequently, there is almost always some uncertainty as to whether the selection bias has been eliminated from the estimation of the treatment effect. Nevertheless, this uncertainty is typically due to the complex nature of treatment assignment, and therefore, the difficulty of identifying the covariates involved in the selection process (Steiner et al., 2010). In incarceration effects research, however, scholars are typically aware of the exclusion of at least some constructs that are germane to selection, as data limitations preclude the examination of certain key social selection forces (Sampson, 2011). More specifically, and as elaborated in more detail in other work (Giordano & Copp, 2015), research on incarceration effects has seldom included controls for parental criminality despite the well-documented association between criminal offending and sentencing decisions. Instead, propensity score models rely predominantly on a roster of family background characteristics which, although correlated with the selection process and the outcomes of interest, are not inherent to decisions to incarcerate. Because offending is a necessary condition for incarceration and has also been linked to an array of child well-being outcomes it seems that failure to account for the parent's offending behaviors may pose a rather considerable threat to strong ignorability, and consequently, causal estimates.

Some scholars have attempted to address this issue indirectly by conducting sensitivity analyses post-hoc and determining how substantial the unobserved effects would have to be to render findings nonsignificant (e.g., Turney, 2017; Turney & Wildeman, 2015). Yet findings from these analyses are somewhat difficult to interpret substantively. For example, Turney and Wildeman (2015) estimate the gamma statistic (Γ) for hidden biases (Rosenbaum, 2002) and find that in order to reverse the conclusion that incarceration causes detriments to child well-being, unobserved characteristics would have to increase the odds of incarceration by 70%, 130%, and 150% for internalizing-, externalizing-, and delinquent behaviors, respectively. Given that offending variables (e.g., offending history, offense severity) have been identified as among the strongest predictors of sentencing outcomes (e.g., Kramer & Steffensmeier, 1993; Spohn, 1994; Steffensmeier & Demuth, 2001)—including decisions to incarcerate—it seems plausible that parental (offending) behaviors may increase the odds of being incarcerated by a factor as large as those presented above (see also Giordano & Copp, 2015). Rosenbaum (2002) discusses the sensitivity of certain types of research questions to large versus small hidden biases. In our view, even large Γ values in incarceration effects research do not confer the same sense of security as they may in other areas of research given the inability of existing data to balance over selection forces that are intrinsically tied to the treatment. In addition to raising questions about the validity of causal estimates, the potential omission of key selection forces raises additional questions related to the notion of nonzero probability and the determination of regions of common support.

As mentioned above, nonzero probability refers to the assumption that an individual has some chance of being in either the treatment or control condition (Rosenbaum & Rubin, 1983). In the case of incarceration effects research, this implies that all children have at least some chance of experiencing parental incarceration. Although propensity score applications have demonstrated that children belonging to the treatment and control groups are similar across a range of sociodemographic, family, neighborhood, and parental characteristics, the omission of constructs tapping parents' actual offending likely overlooks the fact that they differ in ways that affect their likelihood of experiencing parental incarceration.[1] In practical applications, the nonzero probability assumption is typically upheld as members of the treatment and control groups are identified at all levels of the propensity score. This is based, however, on applications that have been unable to eliminate the effects of confounding due to the omission of parental behaviors. Importantly, except for cases of the wrongfully accused, a child whose parent has never engaged in illegal activity has *no chance* of experiencing a parent's imprisonment.

[1] Some may argue that recent shifts in the composition of the prison population have made this point less relevant, as the current prison population is less dangerous and consists of more low-level, non-violent offenders, and thus poses less of a threat to children (Wakefield & Wildeman, 2014). Yet it is unclear whether sustained contact with low-level and non-violent offenders is less consequential for child well-being than exposure to more serious or violent offenders. Furthermore, many of the low-level and non-violent offenders are serving time for drug offenses, and as recent research has emphasized, parental substance abuse poses substantial risk to children (Wakefield & Powell, 2016).

Although this assumption is rarely assessed in criminological research, it bears further consideration—particularly in future analyses, which include constructs that more precisely model the selection process.

The extent to which propensity scores capture information about the mechanisms driving selection also determines the accuracy of the comparisons made between the treatment and control groups. Thus, if a key selection construct is omitted from the estimation of the propensity score, it may appear that comparisons are being made between two comparable groups when, in fact, the two groups differ in ways that drive selection. In the case of improper comparisons between treatment and control groups, the estimated treatment effect would be biased as it would require extrapolation beyond the available data (Loughran & Mulvey, 2010). In theory, if the unconfoundedness assumption cannot be invoked due to the unavailability of key covariates in the data, the researcher must rely on a different identification method, such as strategies that allow for selection on unobservables (i.e., difference-in-differences estimators, instrumental variable approaches) (Caliendo & Kopeinig, 2008). However, because this assumption is untestable, there is often little consideration of the comparability of the treatment and control groups beyond basic tests of covariate balance. In addition, although researchers frequently invoke common support in their analyses, there is limited attention to the amount of overlap, differences between "off-support" and "on-support" cases, and the meaning of a lack of common support. This is important as different matching algorithms impose different restrictions in terms of the distance between potential matches, as well as the use of off-support cases as counterfactuals.

There are a number of helpful sensitivity analyses including the use of different matching algorithms and other modifications to reduce the region of common support and assess the stability of the estimates. Yet fundamentally, if there is insufficient overlap in the kinds of people who go to prison and the kinds of people who do not, propensity score methods are not an ideal solution to the problem. Recognizing the often striking variability in the probability of experiencing parental incarceration, both within and across samples, a more recent approach employed by scholars of incarceration effects has been to consider the effect of parental incarceration on child wellbeing by the propensity for experiencing this event (Turney, 2017; Turney & Wildeman, 2015). This approach is referred to as the stratification-multilevel method of estimating heterogeneous treatment effects (see Xie, Brand, & Jann, 2012 for an overview), and has been identified as a potential strategy for removing most of the selection bias between the treatment and control groups (Rosenbaum & Rubin, 1984). Researchers employing this approach first estimate propensity scores for each unit, and then construct balanced strata based on those scores (Xie et al., 2012). Treatment effects are estimated within stratum and trends are identified across strata-specific effects. This approach has a number of advantages, including the ability to examine variability in incarceration effects as a function of the factors that shape individuals' chances of experiencing parental incarceration. Moreover, with respect to selection bias, this approach purportedly limits bias by comparing units that are more similar across observed covariates and the likelihood of receiving treatment. It is unclear, however, whether unobserved heterogeneity is really less problematic in heterogeneous treatment effect models, as current applications have resorted to relying on a more limited roster of covariates in order to comply with the within-stratum balance requirement (Turney & Wildeman, 2015; see also Turney, 2015). Thus, if one of the advantages of heterogeneous treatment effect models is bias reduction, then introducing bias by excluding potential confounders seems at odds with the original goals of relying on this methodological strategy.

4. THE CURRENT STUDY

The increase in the use of propensity score models in incarceration effects research is unsurprising given the desire to estimate causal effects using non-experimental data. However, the application of propensity score analyses to research on the consequences of parental incarceration for child well-being is potentially problematic for reasons that limit the ability of researchers to infer causality. This includes the exclusion of theoretically meaningful covariates, resulting in omitted variable bias. Additionally, well-executed propensity score designs require a comparable control group; nevertheless, the comparability of treatment and control groups using existing broadly representative data is unclear. More fundamentally, propensity score analyses implement a counterfactual framework to establish causality which, in the case of incarceration effects research, considers what would have happened to the children of incarcerated parents had they not experienced the incarceration event (and vice versa). One of the underpinnings of propensity score analysis is the idea that individuals have some chance of being in either the treatment or control group (i.e., positivity assumption or nonzero probability). This sets up a somewhat difficult question conceptually as children of non-offending parents have virtually no chance of experiencing the treatment. Until more empirical attention is directed to these issues, it is premature to view the findings from this work as evidence of a causal association. For these reasons alone we encourage researchers to continue to pursue other innovative and advanced methodologies to examine whether and how parental incarceration transmits risk to children.

Yet in addition to these more conceptual concerns, which confront the issue of whether propensity score methods are appropriate for the study of incarceration effects, there exist a number of practical concerns with respect to the implementation of propensity score analyses in incarceration effects research. In the current investigation, we used data from the Toledo Adolescent Relationships Study (TARS) and the Fragile Families and Child Wellbeing Study (FFCWB) to provide an empirical demonstration of propensity score methods to estimate incarceration effects. We draw on these two data sets because each has unique strengths. In particular, the FFCWB study is recognized as a leading source of information on parental incarceration, and includes many indicators of the child's social context. However, the FFCWB data include a limited number of intergenerational indicators (e.g., single-item indicators of parental drug use and violence), and no questions about parental offending behaviors. The TARS data offer a more comprehensive set of covariates, including the parents' and other family members' involvement in crime, violence, and drug use. Further, parental incarceration in the TARS is assessed using both official and self-report data.

The current analyses build on our prior work by reexamining the TARS data to highlight potential ambiguities that may arise during propensity score generation, treatment effect estimation, and post-estimation sensitivity analyses. As such, we focus on the following: (1) identifying covariates and estimating propensity scores; (2) estimating the treatment effect; and (3) performing sensitivity analyses and examining effect heterogeneity. Prior applications of heterogeneous treatment effect models have excluded key variables from the propensity score matching equation (e.g., Turney & Wildeman, 2015). Accordingly, our replication of a prior FFCWB analysis addresses the within-stratum balance requirement, and considers the implication of covariate exclusion for causal estimates.

Our hope is that this effort sparks additional discussion regarding the need for specialized data collection efforts to address existing data limitations, and the potential utility of pursuing other methodological approaches that may be better equipped to address the question of whether/how incarceration confers risk to children.

5. DATA AND METHODS

This research drew on data from the Toledo Adolescent Relationships Study (TARS) and the Fragile Families and Child Well-Being Study (FFCWB). The TARS is based on a stratified random sample of 1,321 adolescents and their parents/guardians. Five waves of TARS data were collected in the years 2001, 2002, 2004, 2006, and 2011. The sampling frame of the TARS study encompassed 62 schools across seven school districts. The initial sample was drawn from enrollment records of the 7th, 9th, and 11th grades, but school attendance was not a requirement for inclusion in the study. The stratified, random sample was devised by the National Opinion Research Center and includes over-samples of Black and Hispanic adolescents. The initial sample included 1,321 respondents and wave 5 retained 1,021 valid respondents, or 77% of wave 1. Respondents' ages ranged from 12 to 19 at wave 1, and 22 to 29 years at wave 5. For the multivariate analyses we draw primarily on waves 1 and 5 of the structured interviews, including the wave 1 parent/caregiver questionnaire. The analytic sample includes all respondents who participated in the structured interviews; however, respondents with missing or invalid responses on our outcome variables were restricted from the analyses. These restrictions resulted in a final analytic sample of 996 respondents.

The FFCWB is a longitudinal birth cohort study of nearly 5,000 children born in 20 large U.S. cities between 1998 and 2000, including 3,700 children born to unmarried parents and 1,200 born to married parents. Baseline in-person interviews were conducted shortly after the birth of the focal child, and follow-up interviews were completed when children were aged one, three, five, and nine. The weighted sample data are representative of nonmarital births to parents residing in cities with populations over 200,000 (Reichman, Teitler, Garfinkel, & MacLanahan, 2001). Study outcomes were assessed at year 9, and additional covariate information was drawn primarily from the baseline and year 1 interviews. The final analytic sample consisted of 3,196 respondents. Construction of the analytic sample, measures, and handling of missing data followed the procedures outlined in Turney and Wildeman (2015).

Measures

Outcome variables. We examined two indicators of young adult well-being including low educational attainment and adult arrest. *Low educational attainment* was taken from the wave 5 questionnaire based on responses to the following: "How far have you gone in school." We created a dichotomous variable to indicate whether the respondent dropped out prior to completing high school (1 = yes and 0 = no). *Adult arrest*, a single item administered at the time of the fifth interview, asked respondents how often they had been arrested since they turned 18, excluding alcohol related offenses and traffic violations. We created a dichotomous variable to indicate any adult arrests (1 = yes and 0 = no).

Explanatory variables. Parental incarceration, was based on administrative and self-report data. Administrative data were compiled through online records searches and physical searches of court records to identify respondents with exposure to maternal and/or paternal incarceration since birth and prior to age 18. In addition, a single item from the wave 1 questionnaire completed by a resident parent or guardian included the following prompt: "Many children experience changes in their living situation. The following are examples of such changes."

Among the changes listed was, "One of your child's parents was sent to prison."[2] To ensure appropriate temporal ordering, we exclude from our analyses those respondents who experienced parental incarceration prior to the wave 1 interview, based on wave 1 self-reports. Thus, our measure of parental incarceration is a dichotomous variable indicating whether the respondent had experienced a parent's incarceration since wave 1 and prior to age 18 (1 = yes).

Parent's age was a continuous variable taken from the wave 1 parent questionnaire. We referred to this variable as *mother's age,* as the vast majority of parents who completed the parent questionnaire were the biological mothers of the focal respondents. *Household instability,* a 8-item summed scale from the parent questionnaire assessed changes in the focal child's living situation including: (1) "Your child and other family members moved to a different house"; (2) "A relative (other than a parent or sibling), friend, or boy/girlfriend moved into your child's home"; (3) "Your child went to live with her/his other parent (if parents not living in the same household) or another relative"; (4) "One of your child's parents spent more than a week in a hospital or treatment facility"; (5) "Child welfare officials took your child away from his/her parents"; (6) "Your child moved in with a friend's (or boy/girlfriend's) family"; (7) "Your child ran away"; and (8) "Your child moved into his or her own apartment." We created eight dichotomous variables to indicate whether the respondent had experienced each change, and summed scores to create a composite scale. *Poverty,* from U.S. census data at the time of the first interview, indicated the "percent of population living below the poverty level" in the respondent's census tract while growing up. *Receipt of public assistance,* a single item asked the parent at wave 1: "Are you now receiving (any kind of governmental or public) assistance?" (1 =

[2] At the start of the questionnaire parents were informed that the interviewer would frequently refer to the focal child as "your child," and that while they may have other children, they should keep in mind the focal child who was the subject of the interview.

yes). *Parent's perception of neighborhood quality,* a ten-item summed scale, asked parents whether the following posed problems in the neighborhood: "high unemployment"; "litter or trash on the sidewalks and streets"; "run down and poorly kept buildings and yards"; "quarrels in which someone is badly hurt"; "drug use or drug dealing in the open"; "youth gangs"; "vacant or abandoned houses or storefronts"; "prostitution"; "abandoned cars"; and "graffiti"($\alpha = .91$).

We accounted for parent/family antisocial lifestyle using 5 separate scales, including family conflict, parent intimate partner violence (IPV), coercive parenting, parent's early problem behavior, and parent's adult alcohol/substance abuse. *Family conflict* items included the focal (child) respondent's retrospective reports of the following: "Family members fought a lot"; "Family members often criticized one another"; "Family members sometimes got so angry they threw things"; and "Family members sometimes hit each other." *Parent IPV,* a dichotomous variable based on 4 items asking how often either one of respondent's parents: "threw something at the other"; "pushed, shoved, or grabbed the other"; "slapped the other in the face or head with an open hand"; and "hit the other." Additionally, at the time of the first interview, we asked the custodial parent/caregiver whether they "threatened to hit your child" and "pushed, grabbed, slapped, or hit your child." *Coercive parenting,* a dichotomous variable, indicated whether the parent/caregiver ever engaged in such behaviors. *Parent's early problem behavior,* a 3 item scale based from the wave 1 parent questionnaire asked whether the following happened during their own teen years: "I was suspended or expelled from school"; "I got (someone) pregnant"; "I was arrested by the police." Finally, *parent's adult alcohol/substance abuse,* a 3-item composite scale, assessed whether the parent had done the following in the past year (wave 1): "used alcohol to get drunk"; "gone out to party with friends"; and "used drugs to get high (not because you were sick)" ($\alpha = .80$).

Reside with grandmother indicated whether the focal child's grandmother resided in the household. *Family structure* (wave 1) included the following categories: two biological parents (contrast category), step-family, single-parent family, and any "other" family type. To control for socioeconomic status, we used the highest level of education reported in the wave 1 parent questionnaire. This measure, referred to as "*mother's education*," is represented by a series of dichotomous variables indicating less than high school, high school (contrast category), some college, and college or more. We included *mother's employment* to indicate whether the parent completing the questionnaire was employed at the time of the interview. We measured *mother's depressive symptoms* using a revised 6-item version of the CES-D (Radloff, 1977) ($\alpha = .87$), asking respondents how often each of the following statements was true during the past seven days: "you felt you just couldn't get going"; "you felt that you could not shake off the blues"; "you had trouble keeping your mind on what you were doing"; "you felt lonely"; "you felt sad"; and "you had trouble getting to sleep or staying asleep." Responses ranged from (1) never to (8) every day, and we created a scale based on the mean of the responses. We used a 7-item scale to assess *parenting stress* (wave 1 parent questionnaire), which was the mean of the following: "Raising my child can be a nerve wracking job"; "Some days I feel unsure about the best way to handle a situation involving my child"; "I'd like to be able to do a better job of communicating with my child"; "I sometimes feel overwhelmed by my responsibilities as a mother(father)"; "I worry I don't give my child enough attention"; "I feel I am faced with more problems as a parent now than when my child was younger"; and "I feel on edge or tense when I'm with my child" (responses ranged from 1 "strongly disagree" to 5 "strongly agree"). We included a series of sociodemographic indicators: *gender*, *age*, measured in years at wave 5, as well as four dichotomous variables to measure *race/ethnicity* including non-Hispanic White (contrast category), non-Hispanic Black, Hispanic, and other. Finally, we accounted for early *delinquency* (wave 1) of the focal child using a 10-item variety score version of Elliott and Ageton's (1980) self-report instrument.

6. ANALYTIC STRATEGY

We began by selecting a group of covariates to construct the propensity scores using the TARS data. Covariate selection is a critical step in propensity score analyses, yet given differences in data quality and the availability of covariates, researchers must choose from available indicators. Accordingly, we estimated propensity scores for each observation using a set of covariates that, based on prior research, represent factors associated with parental incarceration and/or child well-being. In addition, we included a set of constructs tapping the family climate (e.g., household instability and parent/family antisocial lifestyle)—an important yet undertheorized source of social selection. We examined the distribution of the propensity scores across the treatment and control groups. Next, we presented descriptive analyses for the full set of covariates prior to- and post-matching, and examined covariate balance to confirm that the only observable difference across treatment and control groups is the experience of parental incarceration. We then estimated the average treatment effect on the treated based on the propensity scores using kernel matching, a non-parametric matching estimator that potentially uses all members of the control group to create a counterfactual observation for a treatment group member (Caliendo & Kopeinig, 2008; Guo & Fraser, 2010). To examine the robustness of the results, we conducted a series of sensitivity analyses, comparing our findings across a range of matching algorithms and specifications. We examined heterogeneous treatment effects using multilevel propensity score models to further address the issue of bias under unconfoundedness.

We also drew on data from the FFCWB study to examine an existing application of heterogeneous treatment effect models. The FFCWB data are one of the most widely used sources of data for studying incarceration effects, and moreover, these data have been used to study effect heterogeneity. Using prior research as a guide (Turney & Wildeman, 2015), we examined the heterogeneous effect of maternal incarceration on internalizing and externalizing problem behaviors, PPVT-III scores, and juvenile delinquency, focusing particular attention on whether the decision to exclude covariates from the matching algorithm in order to comply with the within-stratum balance requirement influenced propensity scores and effect estimates.

7. RESULTS

Identifying covariates and estimating propensity scores. In Figure 1 we presented the distribution of the propensity scores across treatment and control groups in the TARS data. The average propensity score was 0.17 (0.26 for treatment group and 0.15 for control group), and scores ranged from 0.02-0.67. Although researchers seldom present these distributions, they may be a helpful tool for determining where common support exists, and whether we should be concerned with matches at the upper end of the propensity score range where scores for the nontreated cases are especially sparse. Much of the distribution of the control group overlapped with that of the treatment group; however, we see a sizeable concentration of scores at the lower end of the propensity score distribution. That is, 75% of the propensity scores of the control group were less than 0.20, whereas only one-third of the treatment cases fell within this range. Conversely, 50% of observations in the treatment group had propensity scores greater than 0.24, as compared to less than 20% of control group observations. Examination of these distributions helps determine the sensitivity of propensity score estimates to covariate selection. Furthermore, it shows that fewer appropriate comparisons may become available in the data as we come closer to approximating the life circumstances of children who have experienced parental incarceration.

We presented descriptive statistics (Table 1), including sample means for all study variables by the experience of parental incarceration. We also provided the *t*-statistic for mean differences across groups, as well as correlations between the study variables and parental incarceration. Notably, 18 of the 25 covariates included were significantly different between the subgroups with and without exposure to parental incarceration.

Estimation of the treatment effect. Next, we employed kernel-based matching which, unlike other standard matching algorithms uses weighted averages of all control-group observations to construct the counterfactual outcome. A major advantage of this approach is that it uses more information than matching algorithms that draw on only a subset of control group observations. An obvious limitation of this approach is that precisely because all control group members are used, bad matches are potentially included in the process. In these analyses we used a generalized version of kernel matching (i.e., local linear matching), as this approach better handles data in which the control group observations are not distributed symmetrically around the treatment group observations (Smith & Todd, 2005)—as is the case in the current investigation. Post-matching descriptive information (Table 2) revealed that covariate values of the treatment and control groups were nearly identical. That is, *t*-test comparisons of means across the treatment and control groups suggested that none of the differences were significant post-matching. The percentage reduction in bias similarly indicated that matching substantially reduced covariate imbalance. This suggests that, independent of treatment status, observations with similar propensity scores should have the same distribution on observable characteristics (Becker & Ichino, 2002). Accordingly, any observed differences between the outcomes of treatment and control group members can be attributed to the treatment. As other scholars have noted, however, it cannot be inferred that exposure to treatment is random based on balance alone, and additional attention must be directed to the strong ignorability assumption. In the current investigation we moved forward with our analyses in light of these findings, and return to the issue of strong ignorability in the conclusion. It is also important to note that although matching did appear to account for the differences between treatment and control groups along the propensity score, a number of standardized differences exceeded 10% in absolute value post-matching, and a handful (family structure, mother's education, race/ethnicity) approached 20%, suggesting that a large amount of bias remained for these variables.

Table 3 presented estimates for the average effect of parental incarceration on child well-being across two outcomes including adult arrest and low educational attainment. We provided differences based on the unmatched sample in the first column, which indicated that as compared to their peers with no history of parental incarceration, children of currently/previously incarcerated parents are more likely to be arrested as adults ($b = 0.082$, $p < .01$), and are more likely to not complete high school ($b = 0.081$, $p < .001$). We presented the matched differences in the next column. These matched differences indicate that there were no statistically significant differences in adult arrest or educational attainment for the treatment and control groups.

Sensitivity analyses and effect heterogeneity. Given different substantive conclusions based on the unmatched and matched samples, we employed a number of sensitivity analyses (not shown) to determine the robustness of the findings. We began by "trimming" observations, which effectively imposes common support by dropping a specified percent of treatment observations (Guo & Fraser, 2010). We reestimated our models, trimming 2%, 5%, and 10% of observations (Heckman et al., 1997). This approach helps identify whether the treatment effects are sensitive to the distributional properties of the estimated propensity scores. The substantive findings of the matched sample remained unchanged. More specifically, the matched differences indicated that there were no significant differences between the treatment and control groups across the outcomes included in this investigation, and that this conclusion remained unchanged after trimming 2%, 5%, and 10% of observations.

Additional sensitivity analyses used different bandwidth sizes. That is, whereas the average effect models in Table 3 relied on the default bandwidth (0.06), we examined the average effects based on four additional models using bandwidths ranging from 0.005 to 0.8. Similar to the findings described above, findings were substantively similar across the different bandwidth specifications, providing additional support for the null findings based on the average effect findings presented in Table 3. Supplemental models also examined results using nearest neighbor matching, including nearest neighbor without replacement (caliper = .25 * SD of logit of propensity score) and nearest 5 neighbors (caliper = 0.005), and standard regression models, and the null findings were robust to these variations. In a recent critique of propensity score methods, Loughran and colleagues (2015) indicated that under conditions of unobserved heterogeneity and other sources of hidden bias, propensity scores are no better at producing causal estimates than more traditional regression approaches. In the current analyses, substantive findings based on the standard regression models were similar to those obtained from the models using propensity score estimates. More specifically, the effect of parental incarceration on the outcomes (adult arrest and low educational attainment) was not significant, controlling for the full range of covariates included in this investigation (see also authors, 2016).

A final set of analyses examined heterogeneous treatment effect models. These models have been identified as an effective approach for understanding variation in the effect of parental incarceration on child wellbeing by allowing researchers to "simultaneously consider the possibilities of negative, positive, or null effects" (Turney, 2015: p. 469). Similar to the balance requirements of traditional propensity score approaches, multilevel approaches require that within each stratum, the treatment and control groups do not significantly differ across covariates. This can be especially difficult to achieve in incarceration effects research given the often sizeable differences between members of the treatment and control groups on key selection constructs. The sample selection criteria utilized in the current investigation excluded respondents with exposure to parental incarceration prior to the wave 1 interview, as their inclusion would have raised concerns regarding appropriate temporal ordering. Although necessary, their exclusion results in a sample that is more comparable across the matching covariates than is often the case in incarceration effects research employing propensity score techniques. Accordingly, using the full roster of study variables presented in Table 1, we were able to achieve balance within-stratum. Based on the propensity scores obtained, we grouped observations into 4 strata such that those with the lowest propensities of experiencing the treatment were in stratum 1 and those with the highest propensities were in stratum 4. Table 4 presented the descriptive statistics for the matched sample across the covariates included in this portion of the investigation by stratum.

Based on the propensity scores and stratum presented in Table 4, we estimated propensity score stratum-specific effects (level-1), and trends across strata using variance-weighted least squares regression (Xie et al., 2012). Table 5 presented the results from the multilevel models estimating heterogeneity in the effect of parental incarceration. Similar to the findings from the average effect models, the findings suggested a null effect of parental incarceration on wellbeing across stratum. That is, the level-1 coefficients indicated that the effect of parental incarceration was nonsignificant across strata. Further, the level-2 slopes suggested that the between-stratum differences in the effect of parental incarceration were nonsignificant.

Within-stratum balance requirement. Given that we were able to achieve balance within stratum using the full roster of study covariates in the TARS data analyses presented above, we drew on data from the FFCWB for this final stage of our analyses. We replicated the Turney and Wildeman (2015) analyses by first identifying a set of matching covariates, including factors associated with incarceration and/or child well-being, to generate a propensity score for each observation (see Appendix A1 for descriptive statistics). Using the full list of controls, the average propensity score for the FFCWB sample was 0.09, and scores ranged from 0.00-0.80.[3] Following the steps outlined by Turney (2015), we restricted the number of covariates included in the matching equation to ensure that covariate values for the treatment and control groups were similar within these subgroupings (see also Turney & Wildeman, 2015). Based on the propensity score estimates obtained, we grouped observations into 3 strata such that those with the lowest propensities of experiencing the treatment were in stratum 1 and those with the highest propensities were in stratum 3. The range of the propensity scores was narrower as a result of constraining the model estimating the propensity scores to a smaller set of covariates. That is, the upper limit was reduced from 0.80 to 0.33. This is consistent with findings from Turney and Wildeman's (2015) recent examination of heterogeneous effects where propensity scores generated during the estimation of average effects ranged from 0.00 – 0.79 and those obtained using a more limited set of covariates in the multilevel findings ranged from 0.00 – 0.30.

[3] To confirm that the Turney and Wildeman (2015) analyses were appropriately replicated, we estimated a series of average effect models. Similar to the findings reported by the authors, we found no evidence of an average effect of maternal incarceration on internalizing problem behaviors, externalizing problem behaviors, PPVT-III scores, or early juvenile delinquency.

We examined the sensitivity of these findings by first considering how the exclusion of covariates influenced the identification of stratum. We found that the designation of individuals to stratum would have changed entirely with the inclusion of the full roster of covariates. That is, the narrow propensity range identified above is not merely indicative of scores being constrained, but rather changes in individuals' relative propensities of experiencing the treatment. For example, whereas members of stratum 1, 2, and 3 had propensity scores ranging from [0.01 – 0.05], [0.05 – 0.10), and [0.10 – 0.33], respectively, according to propensity score estimates using all study variables, scores for stratum 1 ranged from [0.00 – 0.53], stratum 2 [0.01 – 0.50], and stratum 3 [0.02 – 0.80]. While this method is lauded for reducing bias by comparing more homogeneous groups within-stratum, our analyses suggested that the decision to exclude covariates in order to achieve balance may introduce more bias and result in making comparisons between groups that substantively differ in ways that are integral to the selection process.

Furthermore, in addition to achieving balance within-stratum, another consideration of multilevel propensity score methods is that the average propensity scores of treatment and control members were statistically similar within-stratum. In the FFCWB data, it was not possible to group individuals such that the covariates were balanced and the propensity scores were similar. The greatest difference in scores were observed in stratum 1 ($p < .05$) and stratum 3 ($p < .001$), which includes both the highest and lowest risk groups. These findings suggest that the omission of known confounds can significantly influence our assessment of the impact of parental incarceration on child well-being.

8. CONCLUSIONS

In a recent article, Wakefield and colleagues (2016) commented on our current understanding of the consequences of the American criminal justice system for families and children, concluding that "our view remains obscured by serious data limitations" (p. 10). We agree with this view, and suggested that in order to further develop our understanding of the mechanisms underlying the link between parental incarceration and problem outcomes, we need to continue to think creatively about future data collection and research efforts. Yet despite the general consensus that we have much to uncover with respect to the consequences of criminal justice policy for family life, scholars largely contend that we have strong evidence of a causal effect of parental incarceration on children. This evidence derives in part from a number of recent studies using a single dataset—a sizeable percentage of which rely on propensity score matching (e.g., Haskins, 2014; 2015; 2016; Turney, 2014; Turney & Haskins, 2014; Turney & Wildeman, 2015; Wakefield & Wildeman, 2011; Wildeman, 2009; 2010; 2014; Wildeman & Turney, 2014). The increase in the use of propensity score methods to address incarceration effects on children is intuitive given the potential for selection bias. This paper addresses several of the assumptions implicit in the strategy and questions the extent to which propensity scores actually helped solve the selection problem in this particular context.

In the current investigation, we addressed the potential limitations of propensity scores for the study of incarceration effects by directing attention to a series of methodological and practical concerns. Our aim was to raise a healthy level of skepticism with respect to existing causal estimates of incarceration effects, and to encourage researchers to either provide more thoughtful discussion of the potential limitations of their findings or to consider alternative estimation strategies. We focused in particular on the assumption of strong ignorability, and argued that existing applications have often provided limited attention to this key tenant of propensity score techniques. As noted by Loughran and Mulvey (2010), "…when attempting to make causal inference, we must make sure there is nothing unobservable which is potentially biasing our estimates despite our best efforts to control for observables" (p. 178). Although it is impossible to completely rule out unobserved heterogeneity, a number of scholars have expressed concern regarding the exclusion of parental behaviors (Giordano, 2010; Giordano & Copp, 2015; Johnson & Easterling, 2012; Sampson, 2011)—a particularly important source of selection bias. Thus, it is potentially problematic that researchers have proceeded ahead, assuming basic comparability across groups when, due to data limitations, they have excluded perhaps the strongest predictor of the treatment from the matching algorithm. Lacking data to appropriately model the selection process, the use of propensity score methods (and resulting causal estimates) may be more problematic for investigating incarceration effects than is typically assumed.

Yet even if we were able to provide good measurement of the selection process, whether propensity scores are an appropriate method for providing estimates of incarceration effects using broadly representative data is unclear. That is, the second component of strong ignorability is nonzero probability, which implies that individuals have a nonzero chance of being in either the treatment or control condition. This idea is central to the counterfactual framework underlying propensity scores, as the potential outcome under the unobserved condition must be a plausible one. Using large, representative samples, we must confront the fact that a substantial portion of the sample has no chance of experiencing the treatment as the parents of the studies' focal children have not engaged in behaviors that would put them at risk of being incarcerated. This issue is further complicated by the lack of adequate controls for parental behaviors, and brings into question the quality of the matches obtained. While scholars often focus on the issue of unobserved heterogeneity, an equally important consideration is the design of a good comparison group. It is difficult to imagine a scenario where children with non-offending parents would have a similar propensity to experience a parent's confinement as their peers who have in fact experienced parental incarceration. However, this is effectively the type of comparison that is being made, and thus despite a number of causal claims, it is unclear to what we may attribute the observed effect.

In addition, propensity scores are highly sensitive to a number of decisions made by the researcher, including aspects where little consensus exists. In the current investigation, we provided an empirical demonstration of propensity score techniques, focusing in particular on some of the more critical decision-making points, and highlighting the variable nature of the findings as a result of the researcher's analytic choices. Given that there is no clear guide on what to include in the initial model estimating the propensity score, we began by identifying a set of selection constructs. We found that despite achieving balance, the distribution of the treatment and control groups were quite distinct such that few treatment cases were found in the lower end of the distribution and few control cases in the upper end of the distribution, raising concerns about the appropriateness of available comparisons. In the next step, we estimated the treatment effect and found that parental incarceration did not appear to influence the odds of adult arrest or low educational attainment. We conducted a number of sensitivity analyses to determine the robustness of the findings to a range of model specifications. We found that null findings remained largely unchanged across the supplemental analyses conducted in this investigation.

The finding of a null average effect, however, does not rule out the possibility that individuals may respond differently to treatment (i.e., posttreatment heterogeneity). The potential for heterogeneity in incarceration effects has been explored in some of the more recent incarceration effects research, indicating that the detrimental effects of parental incarceration on child well-being are most strongly felt among those least likely to experience this event (Turney, 2017; Turney & Wildeman, 2015). Understanding variation has become an important focus of incarceration effects research (National Research Council, 2014), and thus it is quite likely that scholars will continue to use this method in future investigations. Thus, in the current investigation we wanted to further explore this analytic technique and determine its utility for future research in the incarceration effects tradition. Findings from our multilevel treatment effect models using the TARS data were consistent with the average effect results, and suggested a null effect of parental incarceration on children's outcomes across stratum. However, as the sample restriction criteria used in the TARS portion of the analyses resulted in treatment and control groups that were more comparable than is typically the case using broadly representative data, we were unable to appropriately examine one of our primary concerns regarding multilevel treatment effect models—the exclusion of covariates to satisfy the within-stratum balance requirement. Thus, we turned to the FFCWB and an existing application of multilevel treatment effect models recently published by Turney and Wildeman (2015).

As was somewhat anticipated, we found that it was incredibly difficult to balance covariates within stratum in the FFCWB data. Accordingly, we followed the steps outlined by Turney (2015), and excluded a subset of covariates in order to comply with the within-stratum balance requirement. Despite taking this step, we found ourselves somewhat concerned with the potential implications of the loss of covariates—particularly since that decision is not theoretically based and appears arbitrary.

Similar to the results presented by Turney and Wildeman (2015; see also Turney, 2017), we found that parental incarceration appeared to take the biggest toll on children least likely to confront it. However, supplemental analyses revealed significant cause for concern with these preliminary results. In particular, the exclusion of covariates resulted in individuals being shifted across strata. That is, individuals who were placed in the high risk stratum based on the full roster of covariates were potentially recategorized as low-risk when using the more limited set of covariates. This was well captured in an exercise where we examined the propensity scores of individuals assigned to stratum 1, 2, and 3 (where assignment was based on fewer covariates), using the scores generated from the full list of study covariates. We found that the propensity scores ranged from roughly 0 to .80 for all three strata, suggesting that the loss of covariates resulted in a significant loss of information which altered the substantive meaning of low, medium, and high risk in this context. Furthermore, in addition to balancing covariates within-stratum, scholars must ensure that propensity scores are statistically similar within-stratum. In the FFCWB data, we struggled to balance covariates and maintain similar propensity scores across treatment and control groups within-stratum. Notably, the most significant differences between the propensity scores of individuals who did and did not receive the treatment was in strata 1 and 3, which includes the subgroup where the effect of parental incarceration was deemed most deleterious according to the models. Based on our exploration of this particular methodological approach, we suggest that although heterogeneous treatment effect models are a useful strategy for examining differential responses to treatment, they may not be equally adept at estimating posttreatment heterogeneity across treatment types. In addition to the series of concerns we have outlined with respect to the use of propensity scores to estimate average incarceration effects, we suggest the need for additional scrutiny of heterogeneous treatment effect estimates.

Why should any of this matter for the field of incarceration effects research? We argue that this information should be of vital interest to incarceration effects scholars, as recent research employing propensity score methods has made a number of causal claims and recommendations for criminal justice policy. In particular, scholars suggest that reducing our reliance on incarceration will improve children's outcomes (e.g., Wakefield & Wildeman, 2011; 2014; Turney & Wildeman, 2015). In our view it is premature to suggest with any degree of certainty that parental incarceration effects are the key underpinning of the lower levels of well-being observed in subgroups of children who have experienced this event. Thus, the idea of reducing incarceration to improve child well-being is a policy that most can enthusiastically get behind. Yet working to reduce the incarceration option without simultaneously addressing other disadvantages such children are very likely to face may not provide the immediate benefits that all agree are desired.

It is also important to develop alternative methodological strategies that bypass the either-or assumptions of much of this line of research (i.e., is it incarceration or the other disadvantages that drives the detriments to child well-being?). Although the above considerations suggest the need for caution in making causal claims about incarceration effects net of these co-existing adversities, strategies that capture synergistic and reciprocally related effects may hold the most promise. For example, data sets that contain repeated measures of a broader portfolio of factors, including criminal justice experience, parental characteristics (e.g., parents' antisocial behavior) and family dynamics (e.g. parents' use of coercive parenting, financial circumstances) can be leveraged to actively model ways in which parental incarceration and these other dimensions of the child's experience operate together and upon one another to affect key well-being outcomes.

REFERENCES

Becker, S.O., & Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *The Stata Journal, 2*(4), 358-377.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31-72.

Cho, R. M. (2009a). Impact of Maternal Imprisonment on Children's Probability of Grade Retention. *Journal of Urban Economics, 65*, 11-23.

Cho, R. M. (2009b). The Impact of Maternal Imprisonment on Children's Educational Achievement. *Journal of Human Resources, 44(3),* 772-797.

Elliott, D.S., & Ageton, S.S. (1980). Reconciling race and class differences in self-reported and official estimates of delinquency. *American Sociological Review*, 95-110.

Foster, H., & Hagan, J. (2015). Punishment Regimes and the Multilevel Effects of Parental Incarceration: Intergenerational, Intersectional, and Interinstitutional Models of Social Inequality and Systemic Exclusion. *Annual Review of Sociology, 41*, 135-158.

Giordano, P.C. (2010). *Legacies of crime: A follow-up of the children of highly delinquent girls and boys.* Cambridge University Press.

Giordano, P.C., & Copp, J.E. (2015). "Packages" of risk: Implications for determining the effect of maternal incarceration on child wellbeing. *Criminology & Public Policy*, *14*(1), 157-168.

Glaze, L.E., & Maruschak, L.M. (2008). *Parents in prison and their minor children.* Washington, DC: US Department of Justice, Office of Justice Programs.

Guo, S., & Fraser, M.W. (2010). *Propensity score analysis: Statistical methods and analysis*. Thousand Oaks, CA: Sage.

Hagan, J., & Dinovitzer, R. (1999). Collateral Consequences of Imprisonment for Children, Communities, and Prisoners. *Crime and Justice, 26*, 121-162.

Haskins, A. R. (2014). Unintended Consequences: Effects of Paternal Incarceration on Child School Readiness and Later Special Education Placement. *Sociological Science, 1*, 141-158.

Haskins, A.R. (2015). Paternal incarceration and child-reported behavioral functioning at age 9. *Social Science Research, 52*, 18-33.

Haskins, A.R. (2016). Beyond boys' bad behaviors: Paternal incarceration and cognitive development in middle childhood. *Social Forces, 95*(2), 861-892.

Heckman, J.J., Ichimura, H., & Todd, P.E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies, 64*(4), 605-654.

Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945-970.

Johnson, E.I., & Easterling, B. (2012). Understanding unique effects of parental incarceration on children: Challenges, progress, and recommendations. *Journal of Marriage and Family, 74*(2), 342-356.

Kramer, J., & Steffensmeier, D. (1993). Race and imprisonment decisions. *The Sociological Quarterly, 34*(2), 357-376.

Lee, H., McCormick, T., Hicken, M.T., & Wildeman, C. (2015). Racial inequalities in connectedness to imprisoned individuals in the United States. *Du Bois Review: Social Science Research on Race, 12*(2), 269-282.

Loughran, T.A., & Mulvey, E.P. (2009). Estimating treatment effects: Matching quantification to the question. In A.R. Piquero, D. & D. Weisburd (Eds.), *Handbook of Quantitative Criminology* (pp. 163-180). New York: Springer.

Loughran, T.A., Wilson, T., Nagin, D.S., & Piquero, A.R. (2015). Evolutionary regression? Assessing the problem of hidden biases in criminal justice applications using propensity scores. *Journal of Experimental Criminology, 11*(4), 631-652.

Morgan, S.L., & Winship, C. (2014). *Counterfactuals and causal inference: Methods and principles for social research.* New York: Cambridge University Press.

Murray, J., & Farrington, D.P. (2008). The effects of parental imprisonment on children. *Crime and Justice, 37*(1), 133-206.

Murray, J., Farrington, D.P., & Sekol, I. (2012). Children's antisocial behavior, mental health, drug use, and education performance after parental incarceration: A systematic review and meta-analysis. *Psychological Bulletin, 138*(2), 175-210.

Murray, J., Farrington, D.P., & Sekol, I. (2012). Children's antisocial behavior, mental health, drug use, and education performance after parental incarceration: A systematic review and meta-analysis. *Psychological Bulletin, 138*(2), 175-210.

Murray, J., Loeber, R., & Pardini, D. (2012). Parental involvement in the criminal justice system and the development of youth theft, marijuana use, depression, and poor academic performance. *Criminology, 50*(1), 255-302.

Pew Charitable Trusts. (2010). *Collateral costs: Incarceration's effect on economic mobility.* Washington, DC: The Pew Charitable Trusts.

Radloff, L.S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*(3), 385-401.

Reichman, N., Teitler, J., Garfinkel, I., & McLanahan, S. 2001. The fragile families and child wellbeing study: Sample and design. *Children and Youth Services Review, 23,* 303-326.

Rosenbaum, P.R. (2002). *Observational studies.* New York: Springer.

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

Rosenbaum, P.R., & Rubin, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516-524.

Sampson, R. (2011). The Incarceration Ledger: Toward a New Era in Assessing Societal Consequences. *Criminology & Public Policy, 10(3),* 819-828.

Shadish, W.R. (2002). Revisiting field experimentation: Field notes for the future. *Psychological Methods, 7*(1), 3-18.

Shadish, W.R. (2013). Propensity score analysis: Promise, reality and irrational exuberance. *Journal of Experimental Criminology*, *9*(2), 129-144.

Smith, J.A. (2000). A critical survey of empirical methods for evaluating active labor market policies. *Schweizerische Zeitschrift fuer Volkswirtschaft und Statistik*, *136*(3), 1-22.

Smith, J.A., & Todd, P.E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125*(1), 305-353.

Spohn, C. (1994). Crime and the social control of blacks: Offender/victim race and the sentencing of violent offenders. In G.S. Bridges and M.A. Meyers (Eds.), *Inequality, crime, and social control.* Boulder, CO: Westview Press, Inc.

Steffensmeier, D., & Demuth, S. (2001). Ethnicity and judges' sentencing decisions: Hispanic-Black-White comparisons. *Criminology, 41*(3), 145-178.

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

Travis, Western, & Redburn (Eds.). (2014). *The Growth of Incarceration in the United States: Exploring Causes and Consequences.* Washington, D.C.: The National Academies Press.

Turney, K. (2014). The consequences of paternal incarceration for maternal neglect and harsh parenting. *Social Forces, 92*(4), 1607-1636.

Turney, K. (2015). Beyond average effects: Incorporating heterogeneous treatment effects into family research. *Journal of Family Theory & Review, 7*(4), 468-481.

Turney, K. (2017). The unequal consequences of mass incarceration for children. *Demography.*

Turney, K., & Haskins, A.R. (2014). Falling behind? Children's early grade retention after paternal incarceration. *Sociology of Education, 87*(4), 241-258.

Turney, K., & Wildeman, C. (2015). Detrimental for some? The heterogeneous effects of maternal incarceration on child wellbeing. *Criminology & Public Policy*, *14*(1), 125-156.

Wakefield, S., Lee, H., & Wildeman, C. (2016). Tough on crime, tough on families? Criminal justice and family life in America. *The ANNALS of the American Academy of Political and Social Science, 665*(1), 8-21).

Wakefield, S., & Powell, K. (2016). Distinguishing petty offenders from serious criminals in the estimation of family life effects. *The ANNALS of the American Academy of Political and Social Science*, *665*(1), 195-212.

Wakefield, S. & Wildeman, C. (2014). *Children of the prison boom: Mass Incarceration and the future of American inequality*. New York, NY: Oxford University Press.

Wakefield, S., & Wildeman, C. (2011). Mass Imprisonment and Racial Disparities in Childhood Behavioral Problems. *Criminology & Public Policy, 10(3)*, 793-817.

Wildeman, C. (2010). Paternal Incarceration and Children's Physically Aggressive Behaviors: Evidence from the Fragile Families and Child Wellbeing Study. *Social Forces, 89(1),* 285-310.

Wildeman, C. (2016). Is it better to sit on our hands or just dive in? *Criminology and Public Policy*, *15*(2), 497-502.

Wildeman, C., & Andersen, S.H. (2016). Paternal incarceration and children's risk of being charged by early adulthood: Evidence from a Danish policy shock. *Criminology*, *00*(0), 1-27.

Wildeman, C., Wakefield, S., & Turney, K. (2013). Misidentifying the effects of parental incarceration? A comment on Johnson and Easterling (2012). *Journal of Marriage and Family*, *75*(1), 252-258.

Xie, Y., Brand, J.E., & Jann, B. (2012). Estimating heterogeneous treatment effects with observational data. *Sociological Methodology, 42*(1), 314-347.

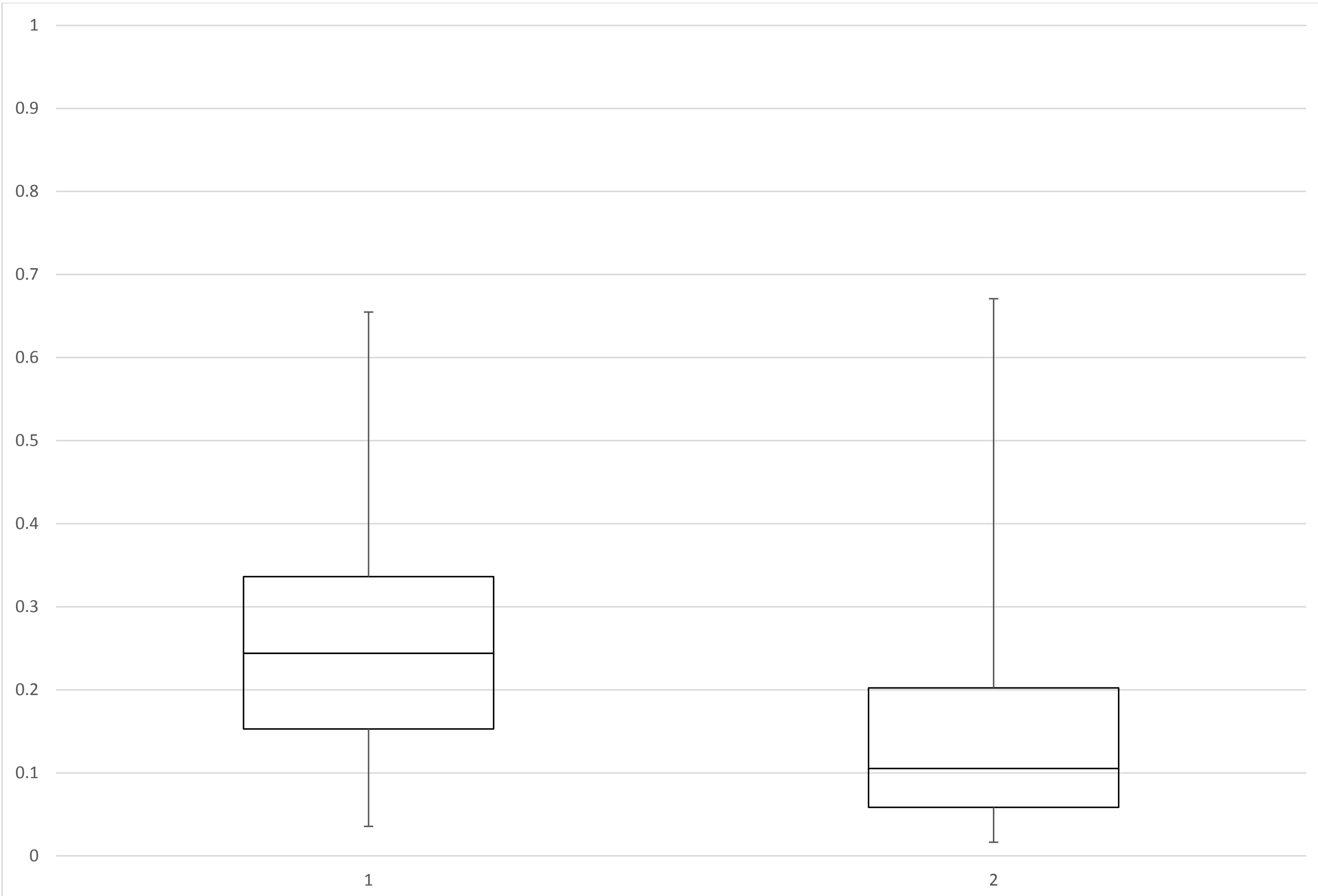

Figure 1. Distribution of the Propensity Score for Parental Incarceration among Treatment and Control Groups, TARS.

Table 1. Descriptive Statistics for the Unmatched Sample (n = 996)

	Parent Ever Incarcerated Mean	Parent Never Incarcerated Mean	*t* for Difference in Means	Correlation with Parental Incarceration
Mother's age	39.11	42.06	-6.33***	-0.20***
Household instability	1.81	1.39	4.27***	0.13***
Poverty	20.12	12.72	6.30***	0.20***
Receipt of public assistance	0.17	0.10	2.46*	0.08*
Parent's perception of neighborhood quality	2.71	1.62	4.68***	0.15***
Parent/family antisocial lifestyle				
Family conflict	2.29	1.97	4.71***	0.15***
Parent IPV	0.51	0.28	5.93***	0.18***
Coercive parenting	0.33	0.22	2.99**	0.09**
Parent's early problem behavior	0.56	0.32	4.45***	0.14***
Parent's adult alcohol/substance abuse	0.45	0.33	2.99**	0.09**
Reside with grandmother	0.02	0.03	-0.99	-0.03
Family structure (bio parents)				
Step-parent	0.17	0.13	1.46	0.05
Single parent	0.28	0.20	2.27*	0.07*
Other family	0.17	0.10	2.52*	0.08*
Parent's Education (high school)				
Less than high school	0.20	0.08	4.68***	0.15***
Some college	0.33	0.34	-0.39	-0.01
College or more	0.10	0.25	-4.44***	-0.14***
Mother's employment (unemployed)				
Employed	0.71	0.80	-2.53*	-0.08*
Mother's depressive symptoms	2.39	2.00	3.54***	0.11***
Parenting stress	2.89	2.83	0.94	0.03
Child gender (male)				
Female	0.49	0.55	-1.37	-0.04
Child race/ethnicity (White)				
Black	0.34	0.19	4.54***	0.14***
Hispanic	0.17	0.10	3.02**	0.10**
Other	0.02	0.02	-0.39	-0.01
Child (w1) delinquency	1.52	1.23	1.72	0.05

* p < .05; ** p < .01; *** p < .001

Source: Toledo Adolescent Relationships Study

Table 2. Descriptive Statistics for the Matched Sample (n = 996)

	Parent Ever Incarcerated Mean	Parent Never Incarcerated Mean	*t* for Difference in Means	Percent Bias Reduction
Mother's age	39.11	38.72	0.69	86.9
Household instability	1.81	1.63	1.30	56.9
Poverty	20.12	18.29	1.08	75.2
Receipt of public assistance	0.17	0.14	0.76	54.5
Parent's perception of neighborhood quality	2.71	2.51	0.60	81.5
Parent/family antisocial lifestyle				
Family conflict	2.29	2.25	0.60	88.7
Parent IPV	0.51	0.57	-1.10	73.8
Coercive parenting	0.33	0.33	0.00	100.0
Parent's early problem behavior	0.56	0.54	0.28	90.1
Parent's adult alcohol/substance abuse	0.45	0.46	-0.33	85.0
Reside with grandmother	0.02	0.01	1.00	16.7
Family structure (bio parents)				
Step-parent	0.17	0.10	1.77	-57.1
Single parent	0.28	0.33	-0.83	47
Other family	0.17	0.15	0.45	73.2
Parent's Education (high school)				
Less than high school	0.20	0.14	1.60	45.0
Some college	0.33	0.41	-1.60	-438.5
College or more	0.10	0.08	0.58	88.5
Mother's employment (unemployed)				
Employed	0.71	0.73	-0.42	76.3
Mother's depressive symptoms	2.39	2.17	1.28	43.3
Parenting stress	2.89	2.84	0.56	25.8
Child gender (male)				
Female	0.49	0.48	0.33	68.8
Child race/ethnicity (White)				
Black	0.34	0.42	-1.47	50.0
Hispanic	0.17	0.16	0.44	77.3
Other	0.02	0.01	0.45	-25.0
Child (w1) delinquency	1.52	1.40	0.47	58.5

* $p < .05$; ** $p < .01$; *** $p < .001$

Source: Toledo Adolescent Relationships Study

Table 3. Propensity Score Matching Estimates of the Average Effect of Parental Incarceration on Child Wellbeing (n = 996)

	Unmatched	Matched
Adult Arrest	0.082**	0.036
	(0.026)	(0.050)
Low Educational Attainment	0.081***	0.011
	(0.011)	(0.047)
Treatment *N*	166	166
Control *N*	830	830

* p < .05; ** p < .01; *** p < .001

Source: Toledo Adolescent Relationships Study

Table 4. Descriptive Statistics for the Matched Sample, by Stratum (n = 904)

	Stratum 1 p = [0.03 – 0.10]	Stratum 2 p = [0.10 – 0.20]	Stratum 3 p = [0.20 – 0.40]	Stratum 4 p = [0.40 – 0.63]
Mother’s age	44.30	40.81	38.13	36.25
Household instability	0.97	1.62	1.88	2.30
Poverty	6.18	13.77	21.67	33.31
Receipt of public assistance	0.05	0.13	0.17	0.27
Parent’s perception of neighborhood quality	0.60	1.74	3.12	4.45
Parent/family antisocial lifestyle				
Family conflict	1.73	2.00	2.36	2.77
Parent IPV	0.08	0.29	0.60	0.77
Coercive parenting	0.14	0.27	0.34	0.47
Parent’s early problem behavior	0.13	0.29	0.66	0.88
Parent’s adult alcohol/substance abuse	0.23	0.37	0.48	0.58
Reside with grandmother	0.02	0.03	0.02	0.00
Family structure (bio parents)				
Step-parent	0.09	0.15	0.19	0.23
Single parent	0.15	0.19	0.32	0.41
Other family	0.02	0.15	0.19	0.21
Parent’s Education (high school)				
Less than high school	0.01	0.07	0.19	0.47
Some college	0.36	0.40	0.36	0.21
College or more	0.35	0.13	0.04	0.00
Mother’s employment (unemployed)				
Employed	0.86	0.77	0.70	0.61
Mother’s depressive symptoms	1.82	1.96	2.35	3.24
Parenting stress	2.78	2.86	2.89	3.02
Child gender (male)				
Female	0.57	0.51	0.54	0.33
Child race/ethnicity (White)				
Black	0.03	0.21	0.40	0.56
Hispanic	0.02	0.11	0.21	0.27
Other	0.02	0.04	0.02	0.00
Child (w1) delinquency	1.07	1.33	1.38	2.08
Treatment *N*	21	33	87	25
Control *N*	304	213	180	41

* p < .05; ** p < .01; *** p < .001

Source: Toledo Adolescent Relationships Study

Table 5. Propensity Score Matching Estimates of the Heterogeneous Effects of Parental Incarceration on Child Wellbeing with Region of Common Support (n = 904)

	Level 1				Level 2
	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Trend
Adult Arrest	0.084	-0.036	0.078	0.036	0.001
	(0.055)	(0.061)	(0.044)	(0.113)	(0.031)
Low Educational Attainment	0.031	-0.010	0.057	-0.002	-0.000
	(0.030)	(0.048)	(0.049)	(0.084)	(0.022)
Treatment *N*	21	33	87	25	
Control *N*	304	213	180	41	

* p < .05; ** p < .01; *** p < .001
Source: Toledo Adolescent Relationships Study

Appendix A1. Descriptive Statistics of Study Variables Fragile Families and Child Well-Being Data (n = 3,196)

Variable	Mean	(SD)	Minimum	Maximum
Dependent Variables				
Internalizing problem behaviors (y9)	0.159	(0.16)	0	2
Externalizing problem behaviors (y9)	0.179	(0.20)	0	2
PPVT-III (y9)	92.898	(14.85)	37	159
Early juvenile delinquency (y9)	1.249	(1.78)	0	17
Independent Variable				
Mother incarceration (y3, y5, y9)	0.089			
Control Variables				
Mother race (b)				
Non-Hispanic White	0.206			
Non-Hispanic Black	0.503			
Hispanic	0.258			
Non-Hispanic other race	0.033			
Mother and father a mixed-race couple (b)	0.152			
Mother foreign-born (b)	0.133			
Mother age (b)	24.995	(5.97)	14	43
Mother lived with both biological parents at age 15 (b)	0.410			
Mother education (b)				
Less than high school	0.331			
High school diploma or GED	0.318			
Postsecondary education	0.352			
Father education (b)				
Less than high school	0.319			
High school diploma or GED	0.380			
Postsecondary education	0.301			
Mother in poverty (y1)	0.416			
Mother material hardship (y1)	1.791	(2.97)	0	12
Mother employment (y1)	0.550			
Father employment (y1)	0.773			
Mother lives with child's grandparent (y1)	0.189			
Mother relationship with child's father (y1)				
Married	0.281			
Cohabiting	0.312			
Nonresidential romantic	0.061			
Separated	0.346			
Mother has new partner (y1)	0.120			
Mother relationship quality (y1)	3.342	(3.34)	1	5
Mother number of children in household (y1)	2.302	(1.32)	0	9
Mother parenting stress (y1)	2.081	0.71	0	4
Mother depression (y1)	0.146			
Child male (b)	0.521			
Child born low birth weight (b)	0.094			
Child temperament (y1)	2.991	0.68	1	5
Mother smoked during pregnancy (b)	0.191			
Mother used drugs or drank alcohol during pregnancy (b)	0.125			
Mother has substance abuse problem (y1)	0.035			
Father has substance abuse problem (b, y1)	0.263			
Mother impulsivity (y5)	1.675	(0.65)	1	4
Father impulsivity (y1)	2.038	(0.67)	1	4
Mother reports domestic violence (b, y1)	0.044			
Mother previously incarcerated (b, y1)	0.008			
Father previously incarcerated (b, y1)	0.326			

Source: Fragile Families and Child Well-Being Study (FFCWB)