Developmental Psychology
2005, Vol. 41, No. 6, 833–850

0012-1649/05/$12.00 DOI: 10.1037/0012-1649.41.6.833

Maternal Employment and Child Development: A Fresh Look Using Newer Methods

Jennifer L. Hill, Jane Waldfogel, Jeanne Brooks-Gunn, and Wen-Jui Han
Columbia University

The employment rate for mothers with young children has increased dramatically over the past 25 years. Estimating the effects of maternal employment on children's development is challenged by selection bias and the missing data endemic to most policy research. To address these issues, this study uses propensity score matching and multiple imputation. The authors compare outcomes across 4 maternal employment patterns: no work in first 3 years postbirth, work only after 1st year, part-time work in 1st year, and full-time work in 1st year. Our results demonstrate small but significant negative effects of maternal employment on children's cognitive outcomes for full-time employment in the 1st year postbirth as compared with employment postponed until after the 1st year. Multiple imputation yields noticeably different estimates as compared with a complete case approach for many measures. Differences between results from propensity score approaches and regression modeling are often minimal.

Keywords: maternal employment, development, matching, imputation

The world is changing. In 1968, only 21% of mothers with a child less than 1 year old were in the labor force; in 2002, this number was close to 60%.[1] A number of federal policies, the Family and Medical Leave Act (FMLA) and Temporary Assistance to Needy Families (TANF) being two of the most visible, directly address, and thus have the potential to affect, the rates of employment for women with young children. Interest is also growing at the state level in maternal employment and child care policies in the first year of life. Under TANF, individual states now have the authority to specify how quickly mothers must begin working after childbirth in order to retain welfare benefits. Some states continue to exempt mothers from work requirements for 1 year after a birth, whereas others require work as early as 3 months. States also have increased flexibility as to how to use childcare dollars, and two states now have programs to use child care funds to reimburse low-income mothers who stay home for up to 1 year postbirth. States also may set policies around maternity leave for women who were employed prior to the birth; in some instances, these policies go beyond the terms of the FMLA to provide a longer period of leave or paid leave.[2]

To best inform the evolution of these existing policies, in addition to the possible creation of new policies, researchers must discern the effects of maternal employment in the first year of life on outcomes for children. This article concentrates primarily on this issue. Developmental psychological theory suggests the importance of early experiences for children's later cognitive and socioemotional development (see recent review in Shonkoff & Phillips, 2000). To the extent that first-year maternal employment affects the care children receive from their parents, and the care received by other caregivers, theory suggests that such employment could affect child outcomes. However, the direction and magnitude of likely effects is not clear a priori, and prior research has not provided consistent answers. Thus, our goal in this article is to provide evidence regarding how maternal employment in early childhood affects children's outcomes through the use of newer and potentially more robust methods.

Previous Research

Child policy research relies on experimental and observational data to estimate impacts on children's well-being. Most experiments in this area focus on specific interventions such as early childhood education, class size, interventions to enhance parenting practices, literacy, or parent–child relationships (Fulgini & Brooks-Gunn, 2000; Karoly et al., 1998; Krueger, 1999; Mosteller, 1995). Unfortunately, experiments do not always address the key questions of interest because of ethical and logistical constraints.

Jennifer L. Hill, School of International and Public Affairs, Columbia University; Jane Waldfogel and Wen-Jui Han, School of Social Work, Columbia University; Jeanne Brooks-Gunn, Teachers College, Columbia University, and College of Physicians and Surgeons, Columbia University.

This study was supported by a Scholars Award from the William T. Grant Foundation to Jane Waldfogel and National Institute of Child Health and Human Development Grant 5R29 HD 35150-04 to Jane Waldfogel. We also thank the NICHD Family and Child Well-Being Research Network.

Correspondence concerning this article should be addressed to Jennifer L. Hill, School of International and Public Affairs, Columbia University, 740 International Affairs Building, 420 West 118th Street, New York, NY 10027. E-mail: jh1030@columbia.edu

[1] Authors' tabulations are from the March 1968 and the March 2002 Current Population Survey.

[2] See Smolensky and Gootman (2003) for a good overview of these policies.

They may also involve select populations leading to inferences that do not generalize well to the population of interest.[3]

We know of no experiments that directly address the impact of maternal employment in the early years on children's development. The example that comes closest, welfare-to-work experiments studied by Manpower Demonstration Research Corporation, was limited to low-income families and included few children under the age of 3 (Morris, 2002; Morris, Gennetian, & Duncan, 2005).

A wealth of previous literature uses observational data such as the National Longitudinal Survey of Youth (NLSY) to estimate the effects of early maternal employment on child outcomes (see, e.g., Han, Waldfogel, & Brooks-Gunn, 2001; Ruhm, 2003). One of the most difficult methodological issues in studying this causal process is the fact that there are substantial differences between women who work soon after their child is born and women who do not. For example, those who work in the first year after their child is born tend to be more educated, to have higher incomes, and to be more likely to be married than those who do not go back to work in the first year. These characteristics are positively related to child outcomes as well. It is impossible, therefore, to know for certain whether differences in outcomes between these groups are related to these initial differences in background characteristics—often called *confounders* or *confounding covariates*—or, instead, to the differences in employment patterns. This phenomenon is referred to as *selection bias* (Heckman, 1979) in the economics literature and *nonignorable treatment assignment* (Rubin, 1978) in the statistics literature if estimates do not properly control for these confounding covariates.

The three most commonly used approaches to selection bias are (a) regression with controls or covariates, (b) fixed effects models, and (c) instrumental variables approaches. We briefly discuss the limitations of these approaches and then review a different approach, *propensity score matching* (P. R. Rosenbaum & Rubin, 1983; P. R. Rosenbaum & Rubin, 1984; P. R. Rosenbaum & Rubin, 1985), which has been demonstrated to perform better than some competing econometric techniques in certain settings (Dehejia & Wahba, 1999; J. L. Hill, Reiter, & Zanutto, 2004). This approach is gaining in popularity for causal inference across disciplines, including economics (Imbens, Rubin, & Sacerdote, 2001; Lechner, 1999; Sianesi, 2004), public health (Fiebach, 1990), psychiatry (Lavori, Keller, & Endicott, 1995), and health and public policy (Foster, 2003; J. Hill, Waldfogel, & Brooks-Gunn, 2002) and has been recently recommended as a sensible approach by leading developmental researchers (Shonkoff & Phillips, 2000). To our knowledge it has yet to be applied to the question of the effect of maternal employment on children's development.

A second contribution distinguishes this article. Missing data are a problem in most policy research, for both background information and outcomes. Child policy research outcome measures often rely on child assessments, which necessitate in-home visits that depend on the state and fatigue of the child in question. Consequently, there are higher missingness rates for outcomes in child policy than there might be for research areas where this information can be collected via telephone or mail-in surveys. More often than not in the literature this missingness has been dealt with by simply throwing away observations with missing values. Such complete case analyses are only representative of children who completed the assessments (often only those who completed them at all time points). We relied on a more principled approach for handling missing data than has been used in previous studies.

Our results shed light on the potential causal effects of maternal employment on child outcomes and illustrate an approach to causal inference and missing data in the context of observational studies that could be applied in many other topics in child policy. Comparing the results across methods reveal their sensitivity to the choice of selection and missing data approaches.

Selection Bias

This study targets one of the most potentially damaging shortcomings in the previous literature on maternal employment and child outcomes: failure to adequately address selection bias. Few studies have been able to fully correct for this problem. Coupled with variability in samples used (which is also a consequence in some cases of various missing data strategies) and the types of variables included, this problem has resulted in a literature with conflicting conclusions even when the data analyzed come from the same dataset (i.e., the NLSY; Han et al., 2001; Harvey, 1999).

Regression analysis, the first method we review, attempts to address selection bias by including potential confounding covariates in a linear model. For example, a recent article (Ruhm, 2003) includes a large set of family background variables in regression analyses using NLSY data (see also analyses that use data from the National Institute of Child Health and Human Development [NICHD] Study of Early Child Care; Brooks-Gunn, Han, & Waldfogel, 2002).

If regression methods are to yield valid causal estimates, not only must we measure all confounding covariates (an untestable assumption), but we must also specify the regression model correctly. In practice, however, it can be difficult to assess whether required linearity and additivity assumptions are appropriate when there are many covariates. Moreover, regression models calculate causal estimates implicitly by creating comparisons in predicted outcomes that act as counterfactuals across similar types of people. If there are certain kinds of people in the control group (e.g., those who are poor and less educated) who are not represented in the treatment group, regression models still make predictions about the outcomes for the treatment group even for these kinds of people. These estimates are based purely on model extrapolation, and subsequent comparisons across treatment and comparison groups may severely misstate the true treatment effects.

More recently, researchers in this area have used more sophisticated techniques to address the selection bias issue. A few authors have used family fixed effects (Gamoran, Mare, Bethke, 1999; James-Burdumy, 1999; Waldfogel, Han, & Brooks-Gunn, 2002). These models implicitly control for any unobserved characteristics specific to the family that do not vary across children. In essence, fixed effects models use the outcomes for children whose mothers did not go back to work as counterfactuals for children whose (same) mothers did go back to work. However, family fixed effects models, like regression models, rely on the parametric

[3] For a lively discussion of the relative merits of experimental and non-experimental studies in program evaluation and social policy research, see McCall and Green (2004) with discussions by Cook (2004), Cottingham (2004), and Brooks-Gunn (2004).

specification of the particular models used. In addition, they limit the sample to only the portion of the dataset for which there are multiple observations and varied maternal employment patterns per family. Beyond this, in order to trust resulting causal estimates, one must believe for instance that if there are two children in the family and the mother had different employment patterns for each, these employment decisions are independent of unmeasured aspects of the children or time-varying characteristics of the family that might also influence their outcomes.[4]

Instrumental variables methods (Angrist & Krueger, 1999) have been popular in economics for many years but are just starting to make their way into fields such as psychology (see, for instance, Foster & McLanahan, 1996; Morris, Duncan, & Rodriguez, 2004). This approach relies on the properties of a variable assumed to be exogenous, labeled the *instrument*. The instrument must satisfy several key properties for this approach to yield valid causal estimates. The instrument must be sufficiently predictive of the treatment. The instrument must be, in effect, randomly assigned with respect to the outcomes. And, crucially, the instrument must affect the outcomes only through its effect on the treatment—this property is referred to as the *exclusion restriction*. If these primary (and two more secondary) assumptions hold, then the random variation that the instrument creates in the treatment variable can be used to identify an unbiased estimate of the causal effect of the treatment for a particular subpopulation of the sample (Angrist, Imbens, & Rubin, 1996). Those treatment and control individuals whose behavior with regard to the treatment potentially was changed as a result of the instrument act as counterfactuals to each other. One study (James-Burdumy, 1999) used an instrumental variables (IV) approach with NLSY data (with percentage of county civilian labor force employed in services and county-level employment rates and per-capita income as the instruments). Even though the author used the strongest instruments we have seen in this literature, the author herself noted that her instruments were not particularly strong. In addition, it is unclear whether they satisfied the crucial exclusion restriction.

We use propensity score matching as an alternative to the previous four approaches to selection bias because it relies on weaker (more plausible) parametric assumptions and is more robust to model misspecification (Drake, 1993) than is regression. In comparison to fixed-effect models, although there is a trade-off between weaker parametric assumptions and stronger structural assumptions, the propensity score inferences can be made for a broader subpopulation. Propensity score matching also appears to be more appropriate than an IV approach, given the absence of a useful and valid instrument. Our approach did not, however, allow us to avoid the assumption of selection on observables. We further discuss our approach and how we implemented it in the Method section.

Missing Data

Missing data plague the vast majority of empirical studies, and the literature on the effects of maternal employment on children's development is no exception. Researchers use a variety of strategies to side step the problem, including complete case analyses (also called *listwise deletion*), complete variables analyses, dummy variable indicators for missing data, and nonresponse weighting. Several of these strategies are often combined. The methods vary in how difficult they are to implement and the extent to which they might create bias or large standard errors in subsequent estimates (see Little & Rubin, 2002, and Jones, 1996, for helpful discussions). However, even an amalgam of these techniques may not be enough to allow for analyses with adequate samples sizes for a given set of desired variables.

Complete case analysis is generally the most popular approach. This can produce biased estimates of the effects for the full sample if the people who are excluded from analysis are different in important ways from those who are included. It might provide unbiased results for the reduced sample, but this sample may not be representative of any population of interest. Moreover, a complete case analysis throws away portions of the sample that might lead to substantial losses of precision. Performing complete case analyses involving both the background variables and the cognitive outcomes would reduce the sample size from 6,114 to 1,543.

As detailed in the Method section, we used an approach to missing data called *multiple imputation* (MI) that fills in missing values with predictions based on the observed data. MI typically relies on weaker (more plausible) assumptions than the methods discussed above and appropriately reflects our uncertainty about these missing values.

Hypotheses

We have several hypotheses about how our estimates of the effects of maternal employment on children's development might change as we move from regression analyses of cases with no missing data (complete case analyses) to regression and propensity score analyses performed on multiply imputed data.

> *Hypothesis 1*: Because MI allows us to keep a sample size of 6,114, and because the types of people represented are different across multiply imputed and complete case samples, we expected that average treatment effects might vary for many of our outcome measures across these two missing data strategies. However, we could not predict a priori how much, or in what direction, the estimates would change. Although we could observe what types of people had missing data, we did not know how treatment effects might have differed across these missing data groups.

> *Hypothesis 2*: Because propensity score approaches are a more robust alternative to standard regression approaches in dealing with potential selection bias, we expected that there might have been differences in estimates across these two sets of analyses. The extent and direction of the differences would depend on the nature of the selection bias.

---

[4] A special case of the fixed effects model is a differences-in-differences (DID) model. This approach could be used in this context if an exogenous shock had occurred that induced changes in employment patterns between time periods for one group of people (treatment group) but not for another (comparison group). The change over time in the comparison group can be used as a counterfactual to the change over time that occurs in the treatment group. This assumes, however, that the change for the control group represents what would have happened to the treatment group in the absence of being exposed to the treatment. To our knowledge, no one has used DID to address this research question.

Method

The NLSY Sample

We used data from the NLSY (see Chase-Lansdale, Mott, Brooks-Gunn, & Philips, 1991). The NLSY is a longitudinal survey that began in 1979 with a cohort of approximately 12,600 young men and women aged 14 to 21 and continued annually until 1994 and biannually thereafter. Starting in 1986, the NLSY also collected information on female respondents' children. These children have also participated in cognitive and behavioral assessments biannually from 1986–2000 (at the time these analyses were performed, child assessment data had only been released through 2000). Our sample comprises 6,114 children of the NLSY born from 1982 to 1993, who were consequently old enough to have been tested for the child assessments for ages 3 or 4, ages 5 or 6, and ages 7 or 8 by 2000. The NLSY has been used extensively in prior research on this topic. Thus, our analyses of whether and how results with these data would change as we used more robust methods should shed light on the results of a number of prior studies as well as the policy implications that should be drawn from them.

Measures

To highlight the fact that our goal is to estimate a causal effect of maternal employment on children's development and to elucidate the related statistical issues, we borrow terminology from randomized experiments. Therefore, we divide our measures conceptually into three types: *background* (or *pretreatment*) *measures*, *treatments*, and *outcomes*.

Background measures or confounding covariates. Confounding covariates represent the variables that might influence both selection into a treatment group as well as outcomes. These must be conceived of as either occurring pretreatment (e.g., income in year before child was born) or being unchanged by the treatment (e.g., ethnicity). As others have noted (Harvey, 1999; P. R. Rosenbaum, 1984), it is generally inappropriate to attempt to control for posttreatment variables because they may have been influenced by the treatment.

Confounding covariates include child's ethnicity (e.g., Hispanic, Black, White), child's gender, child's age (in months as of 2000, as children were tested anywhere within a 2-year age range), whether the child was the first born in the family, whether the mother was married when she gave birth, whether the household was in poverty when the mother gave birth, whether the mother worked in the year before giving birth, the mother's age at birth, the mother's (age-adjusted) Armed Forces Qualification Test (AFQT) score, whether the child was low-birth weight (less than 2,500 g), whether the child was born at least 4 weeks early, household income (logged) and family size in the year before the child was born, mother's educational attainment by the time the child was born (less than high school, high school completed, college started but not completed, college completed), and whether the mother's mother worked when she was first surveyed (1979). All of these variables have been postulated to affect both maternal employment decisions and child development outcomes.

Treatment variables. Prior literature suggests that the effects of maternal employment vary by both intensity and timing of return to work (Morris, 2002; Morris et al., 2005). Consequently, we defined treatment groups on the basis of the following categorizations of maternal employment: (a) never worked in the first 3 years (*never*), (b) did not work in the first year but worked sometime in the following 2 years (*no first*), (c) worked part time in the first year (*part time*), or, (d) worked full time in the first year (*full time*).[5] The cut-off between part-time and full-time employment is 24 hr per week and was derived through exploratory analyses regarding the sensitivity of regression estimates to cut-off choice. Also, 24 hr per week can be equated to working less than or more than a 3-day week, which could easily have implications for the type of child care needed. These categorizations allowed us to investigate the effects of various gradations in work levels as well as the effects of the timing of the return to work.

Outcome measures. Cognitive outcomes include the Peabody Picture Vocabulary Test—Revised (PPVT–R; Dunn & Dunn, 1981), administered at ages 3 and 4; the Peabody Individual Achievement Test in Math (PIAT–M), administered at ages 5 or 6 and 7 or 8; and the Peabody Individual Achievement Test in Reading (PIAT–R), also administered at ages 5 or 6 and 7 or 8. All of these are widely recognized tests and the PIAT–R in particular has been linked to other measures of achievement (reading and math) as well as verbal intelligence (Dunn & Dunn, 1981; Phillips, Brooks-Gunn, Duncan, Klebanov, & Crane, 1998).

Behavioral outcomes are derived from children's scores on subscales of the Behavioral Problems Index (BPI), a parental report measure that has been associated with subsequent teacher ratings (Kohen, Brooks-Gunn, McCormick, & Graber, 1997; although prior research has tended to find that early maternal employment or early child care might affect teacher ratings more strongly than might parental ratings). The subscales of the BPI correspond to 10 items reflecting internalizing problems and 18 items reflecting externalizing problems, and assessments were made at ages 4, 5 or 6, and 7 or 8. Because no BPI scores are available for children who were age 3 at their first assessment and as assessments were only performed every other year, we excluded this first assessment from our outcome measures because we felt it was inappropriate to impute data for the children who were not old enough to be assessed at the first time point (and we did not want to lose over half of our sample by ignoring these children). The alphas for these subscales range from .69 to .79 for the Internalizing scale and from .80 to .87 for the Externalizing scale (Han et al., 2001). These subscales are examined separately because prior research has found that maternal employment has differential effects on these two types of behaviors (studies have found some negative effects of early full-time maternal employment on externalizing behaviors but not on internalizing behaviors; see for instance Han et al., 2001).

Missing Data and Multiple Imputation

Missing data are an issue in this study as evidenced by the rates of missingness (see Appendix A). The indicators for the items *respondent's mother worked* and *household income* have some of the highest missingness rates for the covariates. All of the outcome variables have nonnegligible rates of missingness. There are also substantial differences in rates of missingness across maternal employment groups. These differences highlight the importance of choice of missing data assumptions. Complete case analyses assume that missingness is completely random (everyone must have same probability of having missing data), whereas MI allows missingness to vary over different types of people.

To retain as many children as possible who met our eligibility criteria for inclusion in our analyses (i.e., at least age 7 by the time of the 2000 assessment), we used MI (Baer, Kivlahan, Blume, McKnight, & Marlatt, 2001; Rubin, 1987; Schafer, 1997). MI replaces missing values with predictions based on all the other information observed in the study. This creates a "completed" dataset with the original data augmented by imputed data. MI can accommodate many different patterns of missing data. In contrast to single imputation techniques, however, MI properly accounts for our uncertainty about these missing values (leading to appropriate standard errors) by imputing several values for each missing value (with variability due to both sampling error and model uncertainty), creating

[5] In defining the part-time and full-time treatment groups (groups c and d), we do not make a distinction between those who worked in the first year and continued working thereafter, and those who worked in the first year and did not work in the next two years because in our sample, the vast majority of those who worked in the first year also worked in the next two years.

several completed datasets. Standard analyses can be performed on each completed dataset and then results are combined across analyses in straightforward ways (Rubin, 1987) yielding one final answer.[6]

It is important to note that MI relies on more plausible assumptions about the process that creates the missing data than do standard approaches. For instance, rather than assuming that any two people are equally likely to have missing data on a given variable (as complete case analysis does), MI methods typically assume that two people have the same probability of having missing data on a given variable if they have the same values observed for all other variables. This is referred to as the *missing at random* assumption (Little & Rubin, 2002). MI is a more robust missing data technique because it allows researchers to retain a much larger sample size (compared with complete cases), which helps us to believe inferences that generalize beyond our data and to retain a much larger covariate set (compared with complete variables), which helps us to believe our assumptions regarding the selection process necessary for causal inferences.

We implemented MI using the freely available MICE software that runs in Splus/R.[7] Appendix B describes our imputation strategy in more detail. Table B1 displays mean values for the background data for MI as compared with two types of complete case analyses (dropping those participants who were missing outcomes or covariates, or just those missing outcomes). Some notable differences are evident across the three samples. Consider, for example, the fact that the percentage of mothers with less than a high school education rises sharply when the cases that had missing data are included. Table B2 displays similar information for outcome variables. The cognitive outcome means change noticeably across missing data methods and differentially across groups, with the biggest changes occurring for the never group and the smallest for the full-time group. The mean values for the behavioral outcomes change when MI is used, with mean scores rising more often than not.

Differences Across Groups

What are the differences across maternal employment groups once MI has replaced missing values? To answer this question, Table 1 displays the means and standard deviations after MI for all the background and outcome variables used in the analyses, broken down by treatment group. There are substantial differences between these groups. In general, the covariate means differed in ways that might be expected. Those mothers who worked and worked sooner were more likely to be married, have a first born, and have worked before their first child was born and were less likely to have been in poverty before their child was born. They also had higher AFQT scores, prebirth income, and educational attainment at the time their children were born.

It is interesting to note that those who worked part time in the first year appear to have been in a slightly better position than those who worked full time in terms of some key resources: income, health of child, and intelligence of the mother. Arguably, these differences between groups should lead to positive bias in our treatment effect estimates for cognitive outcomes and negative bias for behavioral outcomes (where positive-valued outcomes correspond to undesirable behaviors) because those who were relatively more advantaged would be expected to have had children with better outcomes. There were also differences in ethnic breakdown between maternal employment groups, most prominently between those who worked part time in the first year and the other groups.

There were also differences across groups for the outcome variables. Cognitive outcomes were clearly highest for those children whose mothers worked part time in the first year. Similarly, we see the least problematic behavioral measurements on average for these same children. However, the differences in means of background variables across treatment groups shown in Table 1 make us wary of interpreting the differences in means of outcome variables as causal effects.

Propensity Score Matching

If it is assumed that all variables that affect both selection into treatment groups and outcomes (confounding covariates) have been measured and that the problem of missing data has been appropriately dealt with, the issue becomes how best to control for the differences in these measures across treatment groups. Linear regression ignores the fact that groups may be too different with respect to these covariates (i.e., there is no overlap) to merit comparison and imposes strict linearity assumptions. Propensity score matching, which also requires the assumption of selection on observables, acts as a remedy for these last two problems.

Propensity score matching estimates the effect of maternal employment by creating matched groups based on background characteristics. First, the *propensity score*, or the probability of being treated (i.e., the probability that the child's mother worked postbirth), is estimated for each child. This is straightforward with logistic regression (or probit, or similar) with the maternal employment variable as the dependent variable and the background variables as predictors. This estimated propensity score is used as a one-number summary of all of the covariates for each person. Matching on the estimated propensity score tends to balance all of the covariates across matched maternal employment groups. The remaining imbalance across groups with respect to the confounding covariates is used as a diagnostic for the adequacy of the model used to estimate propensity scores and can lead to revised estimation and better balance. Comparisons of average outcomes across maternal employment groups (or regression-adjusted versions of this estimator) can be used as estimates of the average effect of maternal employment for the group for which matches were found (the treatment group).[8]

Propensity score matching allows us to focus on targeted treatment effect estimates for either the treatment or the control group (e.g., either the children of mothers who went back to work in the first year or the children of mothers who did not go back to work in the first year). Consider the so-called *effect of the treatment on the treated*. This is the quantity estimated in order to draw conclusions about the treatment group. Suppose the treatment group is defined as the children of mothers who go back to work within the first year after birth. To estimate the effect of the treatment on the treated, we would then find matches for these children from among those children who represent the counterfactual—children whose mothers did not go back to work in the first year. Comparisons between the resulting matched groups should yield estimates of going back to work in the first year for those whose mothers engaged in this behavior (but not necessarily for the other children). The effect of the treatment on the controls requires finding matches for the children of mothers who did not work in the first year from among children with mothers who did. If these two groups of children represent distinct mixes of the characteristics that moderate treatment effects, different average treatment effects might apply.

These distinctions are important because when a given policy is implemented, typically it affects a group of people currently exhibiting one

[6] The only caveat regarding the analyses performed on imputed datasets is that the model used to create the imputation always must be at least as general as any subsequent analyses. So, for instance, we were careful to ensure that any variables or interactions that we included in our analyses were also included in the imputation model.

[7] For more on MICE software that runs in Splus/R, see www.multiple-imputation.com

[8] Although we do have to posit a model for estimating propensity scores, this is merely a means to an end for creating better balance across treatment groups. This balance can be diagnosed directly and without knowing what the associated treatment effects will be, thereby creating a more "honest" assessment. Propensity score inferences, therefore, tend to be much more robust to misspecification of the model used to estimate the scores than regression influences are (Drake, 1993).

Table 1
Means and Standard Deviations of Background and Outcome Variables Across Treatment Groups

	Overall (n = 6,114)		Never (n = 1,494)		No first (n = 1,092)		Part-time (n = 725)		Full-time (n = 2,803)	
Variable and data set	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Work before first born	0.69	0.46	0.32	0.47	0.53	0.50	0.85	0.35	0.91	0.28
Mom's mom worked	0.36	0.48	0.37	0.48	0.38	0.48	0.38	0.49	0.36	0.48
Hispanic	0.20	0.40	0.23	0.42	0.22	0.41	0.14	0.35	0.20	0.40
Black	0.28	0.45	0.31	0.46	0.30	0.46	0.17	0.37	0.29	0.45
White	0.52	0.50	0.46	0.50	0.48	0.50	0.69	0.46	0.51	0.50
Female	0.49	0.50	0.49	0.50	0.47	0.50	0.50	0.50	0.50	0.50
Child is first born	0.38	0.49	0.26	0.44	0.37	0.48	0.40	0.49	0.45	0.50
Married	0.66	0.48	0.55	0.50	0.64	0.48	0.80	0.40	0.68	0.47
Household in poverty	0.26	0.44	0.42	0.49	0.34	0.47	0.15	0.36	0.17	0.38
<High school	0.31	0.46	0.47	0.50	0.37	0.48	0.21	0.40	0.22	0.42
High school	0.38	0.49	0.33	0.47	0.42	0.49	0.37	0.48	0.40	0.49
<College	0.20	0.40	0.13	0.34	0.16	0.37	0.25	0.43	0.23	0.42
College	0.11	0.32	0.06	0.24	0.06	0.23	0.18	0.38	0.14	0.35
Family size	3.44	1.87	3.69	1.95	3.69	1.97	3.21	1.61	3.27	1.82
Log scaled income	9.89	1.41	9.52	1.75	9.60	1.63	10.22	1.05	10.12	1.09
Child's age in 2000	155.18	44.53	157.69	44.47	165.96	42.62	153.26	44.89	150.15	44.35
Mom's age at birth	25.37	4.06	25.23	4.12	24.32	4.07	25.88	4.01	25.73	3.96
Mom's IQ test	47.06	26.17	38.86	25.91	43.09	23.98	58.19	26.60	50.09	25.42
Birth weight	117.70	21.06	116.10	22.45	117.46	21.42	120.07	20.87	118.03	20.11
Preterm	0.69	1.68	0.72	1.73	0.65	1.59	0.67	1.80	0.70	1.65
PPVT, ages 3–4	86.72	20.25	81.78	21.56	85.31	20.06	94.42	18.01	87.91	19.37
PIAT–M, ages 5–6	99.03	13.85	95.99	14.19	98.63	13.75	102.84	13.49	99.81	13.46
PIAT–M, ages 7–8	100.43	12.83	97.97	13.21	99.79	13.09	103.32	12.13	101.25	12.46
PIAT–R, ages 5–6	104.35	13.65	101.67	14.02	103.51	12.91	107.20	13.43	105.37	13.52
PIAT–R, ages 7–8	104.07	13.59	101.06	14.70	103.17	13.58	106.67	12.83	105.36	12.83
BPI										
Internalizing, age 5–6	2.61	2.24	2.75	2.37	2.69	2.31	2.34	2.09	2.56	2.18
Internalizing, age 7–8	2.64	2.33	2.68	2.37	2.82	2.42	2.46	2.26	2.59	2.28
Externalizing, age 5–6	5.65	3.75	5.76	3.85	5.87	3.92	5.18	3.52	5.62	3.69
Externalizing, age 7–8	5.80	3.88	5.98	4.03	5.96	3.97	5.42	3.65	5.75	3.81

Note. PPVT–R = Peabody Picture Vocabulary Test—Revised; PIAT–M = Peabody Individual Achievement Test—Math; PIAT–R = Peabody Individual Achievement Test—Reading; BPI = Behavioral Problems Index.

behavior and induces them to switch to another type of behavior, which is exactly the type of phenomenon that these effects (treatment on treated, treatment on controls) address. Regression and other model-based estimates tend to assume implicitly that everyone in the sample has the same treatment effect. [9]

Once propensity scores have been estimated, there are several different ways of creating a matched comparison group.[10] We implemented two matching methods in this article, each of which has relative strengths and weaknesses, as a form of robustness analysis to demonstrate the potential sensitivity of our results to the choice of matching method.

One-to-one, nearest-neighbor matching with replacement (MWR) is arguably the most popular form of matching used today. Each treated unit is simply matched to the control unit with the closest propensity score. Each control unit can be chosen as a match as many times as it is the closest match. Control units never matched are discarded. In comparison to matching without replacement, where each control unit can be used only once, this method is simpler to program, results in the same matches no matter the order in which matches are found, and has been demonstrated in certain situations to reduce bias (Dehejia & Wahba, 2000). Moreover, regression adjustment (with corresponding robust [Huber, 1967] standard errors) can be performed on the matched samples that might help to further reduce bias and increase efficiency (the latter being particularly important given that MWR can sometimes result in substantial reductions in sample size).[11]

In addition to simple MWR, we also used a variant of an approach called *full matching* (P. Rosenbaum, 1991). The basic idea is that rather than matching each treated unit to just one control unit, we match it to as many control units as fall in a given neighborhood of the treated unit. If a given treatment unit has a propensity score that has only one neighboring control but several neighboring treated units, these units would be matched to-

[9] Even models that relax this assumption generally estimate overall treatment effects for everyone in the sample. This amounts to a weighted average of both the effect of the treatment on the treated and the effect of the treatment on controls, which can be difficult to interpret and map to the relevant policy question.

[10] Propensity scores have also been used to create subclasses, to reweight samples, or as a substitute for many predictors in a regression equation, but these alternatives will not be discussed here (D'Agostino, 1998).

[11] Several studies have found that covariance adjustment combined with matching on covariates produces more reliable results than either technique used on its own (Rubin, 1973, 1979; Rubin & Thomas, 2000). This result may seem at odds with the previous discussion regarding the dangers of the strict parametric assumptions imposed by linear regression. It is driven by the fact that once samples have been reduced to treatment and control groups with sufficient overlap (i.e. where we do not have to rely on extrapolations) the biases caused by model misspecification (i.e. assuming linearity when it is not completely justified) decline dramatically. This is related to the discussion of so-called *doubly robust* estimators (Robins & Ritov, 1997).

gether (several treated to one control). These neighborhoods are nonoverlapping, so each control unit is used only once. Our full matching procedure forms a partition of the data in which each subclass has either one treatment unit and one or more control units or one control unit and one or more treated units. Full matching has the advantage of using all of the data and Rosenbaum (1991) has shown that these types of partitions should yield the closest matches possible overall when using all the data. On the other hand, full matching that uses all of the data[12] might have the disadvantage of some of the matches not being as close as those with MWR, which discards all but the very closest. In addition, a straightforward approach to incorporating regression adjustment is not obvious in this context. Because our algorithm (treatment-directed full matching, described in Appendix C) targets treatment units and forms the partition with respect to their propensity scores, it is easier to program than standard full matching.[13]

Highlighted Treatment Group Comparisons

We have defined four treatment groups on the basis of the categorizations of maternal employment: (a) never worked in first three years (*never*), (b) did not work in first year but worked sometime in following two years (*no first*), (c) worked part time in first year (*part time*), or (d) worked full time in first year (*full time*). We can conceptualize these treatment groups as lying on a continuum from the least amount to the most amount of time worked by mothers. Given this, it may be of primary interest to perform comparisons between contiguous (or, in one case, near-contiguous) categories on this continuum.

In the first four comparisons presented in the Results section, we calculated the effect of the treatment on the treated. In this context, this estimand corresponds to the effect of their mothers' employment pattern on the average outcomes of the children in the treatment group as compared with the average outcomes these children would have had if their mothers had instead had the same employment pattern as those in the comparison group. In these comparisons, the treatment group is always considered the group in the pair whose mothers worked more or sooner. For example, when comparing those children whose mothers worked part time in the first year with those whose mothers held off work until after the first year, the effect of the treatment on the treated represents the effect of the mothers' part-time work in the first year as compared with what the outcomes would have been if the mothers had instead held off work until after the first year.

These comparisons should correspond most closely to employment decisions that mothers might make in the absence of a strong intervention. However, some policies effect dramatic changes in people's behavior. For instance, welfare reform has given states the option to require mothers to participate in work as soon as 3 months postbirth, rather than exempting all mothers for at least 12 months. Therefore, in the Discussion section we also examine comparisons that shed light on the potential implications of such policies.

Results

Overview of Data Analysis

The tables in this section present comparisons between different pairs of maternal employment groups. For each table, we calculated (but do not display) a table of *t* scores corresponding to differences in means (i.e., the group that worked more minus the group that worked less) between the given employment groups, first for the full multiply imputed sample and then for each matched sample (MWR and full matching). The smaller the *t* score, the greater the balance between the groups and the more faith we have in the associated causal estimate. We briefly summarize this distribution of *t* scores to reveal the gains in balance achieved by the matching.

Next, several estimates of the causal effect for each comparison are presented: complete case regression estimates, multiply imputed regression estimates, multiply imputed propensity score estimates, both from MWR (regression adjusted) and full matching. Each can be interpreted as the adjusted difference in mean outcomes for the employment group in the pair working more minus the mean outcomes for the group in the pair working less. Robust standard errors (with Huber estimates) were used for the MWR treatment effect estimates, and the full matching standard errors were calculated using bootstrapping. Sample sizes are provided for each comparison. Sample sizes decreased when we performed MWR because we discarded noncomparable units and because some control units were used more than once (but only counted once in the sample size). Note that detrimental effects appear as negative values for cognitive outcomes and positive values for behavioral outcomes (because the behavioral scales measure problems).

Comparisons Between Worked After First Year and Never Worked

We found (in results not shown but available from Jennifer L. Hill on request) that the balance for the matched samples used in this comparison for the MI propensity score analyses was greatly improved compared with that of the unmatched MI samples. For the unmatched MI samples, 8 of the 19 covariates had significant *t* scores when we examined differences in means across treatment groups, with *t* scores ranging as high as 10.70. For the matched samples, none of the *t* scores for difference in means for either matching method was above 0.90.

Table 2 shows the regression and propensity score model results for this comparison. Moving across the table, columns 1 and 2 show the coefficients and standard errors from regression models that used complete case data, columns 3 and 4 show results from standard (unmatched) regression models estimated on multiply imputed data, columns 5 and 6 show results from propensity score MWR (with additional covariance adjustment) estimated on multiply imputed data, and columns 7 and 8 show results from propensity score full matching (difference in means) estimated on multiply imputed data. Each set of models controlled for the entire set of covariates (listed in the note to the table) either through matching, or covariance adjustment (regression), or both—as in the case of the MWR strategy.

All but one of the estimated effects of the cognitive outcome variables were positive with slightly greater magnitude in point

[12] Theoretically we would discard units (either treatment or control) in the tails of the propensity score distribution if they had scores that were far from the rest of the units; this was not necessary with our data.

[13] In a departure from standard full matching implementation (e.g., Gastwirth, Krieger, & Rosenbaum, 2000; Ming & Rosenbaum, 2000), we did not perform inference with randomization-based tests (which would facilitate regression adjustment) because these rely on an assumption of additive treatment effects which were not supported in our data. Rather, we estimated treatment effects with the weighted mean of the subclass-specific treatment effects, where the weights are equal to the number of treated units in each subclass. We estimated standard errors by using bootstrapping (Efron & Tibshirani, 1998) with 500 iterations (which seemed, upon testing, to provide reasonable stability).

Table 2
Comparison of Treatment Effects

Analysis	Complete case regression[a]		MI regression[b]		MI propensity score–MWR[c]		MI propensity score–FM[b]	
	TE	*SE*	TE	*SE*	TE	*SE*	TE	*SE*
PPVT–R, ages 3–4	1.50	1.44	1.58*	0.79	1.05	1.38	1.86	1.30
PIAT–M, ages 5–6	0.81	1.08	1.62*	0.57	1.72*	0.87	1.78*	0.88
PIAT–M, ages 7–8	0.78	1.01	0.98	0.54	0.61	0.73	0.93	0.82
PIAT–R, ages 5–6	−0.49	0.97	0.82	0.62	0.38	0.69	0.78	0.95
PIAT–R, ages 7–8	0.46	1.06	0.95	0.53	0.76	0.76	1.07	0.77
BPI								
Internalizing, age 5–6	0.10	0.18	−0.01	0.10	0.03	0.12	−0.07	0.17
Internalizing, age 7–8	0.17	0.18	0.19	0.10	0.23	0.13	0.16	0.16
Externalizing, age 5–6	0.12	0.29	0.20	0.17	0.33	0.19	0.23	0.25
Externalizing, age 7–8	−0.04	0.29	0.08	0.19	0.15	0.25	0.07	0.27

Note. Table shows treatment effect (TE) estimates for the comparison of children of mothers who went back to work after the first year of their child's life and children of mothers who did not work at all in the first 3 years. Effect is estimated for children of mothers who went back to work after the first year of their child's life. MI = multiple imputation; MWR = matching with replacement; FM = full matching; PPVT–R = Peabody Picture Vocabulary Test—Revised; PIAT–M = Peabody Individual Achievement Test—Math; PIAT–R = Peabody Individual Achievement Test—Reading; BPI = Behavioral Problems Index.
[a] Complete case data for cognitive outcomes ($n = 557$); complete case data for behavioral outcomes ($n = 351$).
[b] $n = 2{,}586$.
[c] $n = 1{,}738$.
* $p < .05$.

estimates overall for multiply imputed estimates, particularly for full matching on propensity scores. For the effect on PIAT–M at age 5 or 6 the complete case estimate was positive, but it was small and not statistically significant. In contrast, all three of these estimates on multiply imputed data were larger and statistically significant (with effect sizes ranging from .07 to .13). The regression estimates also yielded a positive result for PPVT–R scores, though this did not hold up in the propensity score analyses.

All of the results for the behavioral outcomes were positive (indicating higher levels of behavior problems), with three exceptions. However, the point estimates were small and none was statistically significant. There was no consistent pattern in the point estimates change as we adjusted differentially for missing data and selection bias.

Comparisons Between Worked Part Time in the First Year and Did Not Work Until After the First Year

The matched samples had substantially better balance than did the unmatched MI samples for this comparison as well. For the unmatched MI samples, 15 of the 19 covariates had significant *t* scores when we examined differences in means across treatment groups, with *t* scores ranging as high as 16.20. None of the balance *t* scores for difference in means for either matching method was above .90.

Table 3 displays the treatment effect estimates for this comparison. We found a different pattern of results than we did for the first comparison. In the background data, we had seen some suggestion that the children whose mothers worked part time in the first year might have been positively selected as they were more advantaged on a number of characteristics. Table 3 indicates that as we moved from complete case regression analyses to potentially more robust strategies for controlling for these differences (the positive coefficients of part-time employment in the first year compared with not working until after the first year), the cognitive outcomes diminished and became negative (with small gaps associated with both the switch from complete cases to MI as well as with the switch from MI regression to MI and propensity scores). However, none of these coefficients were statistically significant, so the overall finding was one of no effect on cognitive outcomes and the same is true for the behavioral outcomes. Thus, the main conclusion to be drawn from this table is that overall there were no substantial differences in outcomes for children whose mothers worked part time in the first year compared with what would have happened if those mothers had delayed work until after the first year. This finding is robust across specifications.

Comparisons Between Worked Full Time in the First Year Versus Did Not Work Until After the First Year

The balance for the matched samples improved even more substantially compared with that of the unmatched MI samples for this comparison. For the unmatched MI samples, 11 of the 19 covariates had significant *t* scores when we examined differences in means across treatment groups, with *t* scores ranging as high as 23.80. For the MWR samples, none of the *t* scores for difference in means was significant, and for the full matching none was above 1.30.

Table 4 shows the results for this comparison. These results are of particular interest because this is the comparison in which the strongest negative effects on cognitive outcomes have been found in prior research. In common with that research, our complete case regression results (shown in columns 1 and 2) also show significant negative effects of full-time employment in the first year on cognitive outcomes for all but one of the measures (PIAT–R at ages 5 or 6, which is nearly significant). However, these effects changed as we applied more robust methods. A comparison of columns 3 and 4 with columns 1 and 2 shows that expanding the analysis to include the cases that were missing data reduced the magnitude of the negative effects, in

Table 3
Comparison of Treatment Effects

Analysis	Complete case regression[a]		MI regression[b]		MI propensity score–MWR[c]		MI propensity score–FM[b]	
	TE	*SE*	TE	*SE*	TE	*SE*	TE	*SE*
PPVT–R, ages 3–4	2.02	1.52	1.45	0.93	0.59	1.02	0.33	1.24
PIAT–M, ages 5–6	0.59	1.20	−0.20	0.68	−0.49	1.06	−0.68	1.15
PIAT–M, ages 7–8	−0.42	1.10	−0.37	0.68	−0.87	1.13	−0.85	1.14
PIAT–R, ages 5–6	1.00	1.09	−0.36	0.70	−0.93	0.91	−1.49	0.94
PIAT–R, ages 7–8	0.28	1.15	−0.80	0.82	−1.47	1.40	−1.26	1.01
BPI								
Internalizing, age 5–6	−0.08	0.19	−0.10	0.13	−0.06	0.12	−0.04	0.20
Internalizing, age 7–8	−0.04	0.19	−0.09	0.14	0.01	0.19	−0.02	0.20
Externalizing, age 5–6	−0.41	0.32	−0.20	0.20	−0.23	0.24	−0.09	0.31
Externalizing, age 7–8	−0.21	0.33	−0.09	0.21	0.00	0.19	0.09	0.31

Note. Table shows treatment effect (TE) estimates for the comparison of children of mothers who went back to work part-time in the first year of their child's life and children of mothers who did not work at all in the first 3 years. Effect is estimated for children of mothers who went back to work part-time in the first year of their child's life. MI = multiple imputation; MWR = matching with replacement; FM = full matching; PPVT–R = Peabody Picture Vocabulary Test—Revised; PIAT–M = Peabody Individual Achievement Test—Math; PIAT–R = Peabody Individual Achievement Test—Reading; BPI = Behavioral Problems Index.
[a] Complete case data for cognitive outcomes (n = 483); complete case data for behavioral outcomes (n = 290).
[b] n = 1,817.
[c] n = 1,096.

some cases substantially. For instance, the effect of first-year full-time employment on the PPVT–R at age 3 or 4 fell from −2.61 (significant at $p < .04$) to −1.31 (not significant even at $\alpha = .10$). Moving to propensity score models made less of a difference. Indeed, the fully matched treatment effect estimates (in columns 7 and 8) were roughly comparable to the MI regression estimates and were more negative; although they were statistically significant for only two, the PIAT–R and PIAT–M at age 5 or 6 (effect sizes in the range of −.13 to −.15). Additionally, appropriately increased standard errors for propensity score estimates yielded estimated effects on age 7 or 8 PIAT–R that were not significant as compared with the significant regression estimate (effect size −.08), though the coefficient magnitudes were similar. It is important to note that the sample sizes for the multiply imputed estimates were substantially larger, so the smaller estimates also have smaller standard errors (with the exception of one full matching estimate). Thus, the message of this table is that the negative effects of first-year full-time employment on cognitive outcomes was slightly smaller in magnitude than those that were estimated with regression models on samples that looked more like our complete case sample than our full sample (particularly for our longer-term out-

Table 4
Comparison of Treatment Effects

Analysis	Complete case regression[a]		MI regression[b]		MI propensity score–MWR[c]		MI propensity score–FM[b]	
	TE	*SE*	TE	*SE*	TE	*SE*	TE	*SE*
PPVT–R, ages 3–4	−2.61*	1.25	−1.31	0.98	−1.20	0.87	−1.48	1.33
PIAT–M, ages 5–6	−2.09*	0.98	−1.65*	0.60	−1.67*	0.67	−1.79*	0.83
PIAT–M, ages 7–8	−2.48*	0.92	−1.07*	0.47	−1.11	0.75	−1.33	0.84
PIAT–R, ages 5–6	−1.75	0.91	−1.53*	0.53	−1.66*	0.48	−1.98*	0.75
PIAT–R, ages 7–8	−1.86*	0.96	−1.08	0.59	−1.24	0.66	−1.40	0.79
BPI								
Internalizing, age 5–6	0.06	0.15	0.13	0.09	0.09	0.16	0.08	0.16
Internalizing, age 7–8	0.25	0.16	0.03	0.10	−0.01	0.14	−0.01	0.16
Externalizing, age 5–6	0.14	0.25	0.28	0.15	0.25	0.22	0.23	0.25
Externalizing, age 7–8	0.47	0.26	0.42*	0.18	0.32*	0.16	0.40	0.24

Note. Table shows treatment effect (TE) estimates for the comparison of children of mothers who went back to work full-time in the first year of their child's life and children of mothers who did not work at all in the first 3 years. Effect is estimated for children of mothers who went back to work full-time in the first year of their child's life. MI = multiple imputation; MWR = matching with replacement; FM = full matching; PPVT–R = Peabody Picture Vocabulary Test—Revised; PIAT–M = Peabody Individual Achievement Test—Math; PIAT–R = Peabody Individual Achievement Test—Reading; BPI = Behavioral Problems Index.
[a] Complete case data for cognitive outcomes (n = 1,057); complete case data for behavioral outcomes (n = 651).
[b] n = 3,895.
[c] n = 3,473.
* $p < .05$.

comes); but, nevertheless, some significant negative effects at age 5 or 6 remained.

We found some differences in behavioral outcomes across methods, though no consistent patterns were evident. The strongest change in substantive conclusions was for externalizing behaviors at age 7 or 8, for which the multiply imputed estimates achieved statistical significance, which was due, perhaps, to the smaller standard errors yielded through the increased sample size (effect sizes for the significant estimates were .10 for the regression estimate and .08 for the matched estimate).

This comparison, between children whose mothers worked full time in the first year and children whose mothers did not work until after the first year, has been the focus of much prior research. Moreover, prior studies in this literature have sometimes found varying effects for different subgroups. Therefore, we repeated the same analyses as shown in Table 4 for a series of subgroups defined by the child's ethnicity, the mother's marital status, the family's income, the child's sex, the child's birth order, the mother's high school graduation status, and the mother's score on the AFQT. Unfortunately, with small sample sizes within subgroups, we were unable to estimate precise effects within subgroups. However (in results not reported here but available from Jennifer L. Hill on request) the overall pattern of the results indicated that the significant negative effects on cognitive outcomes were concentrated on the children with the most resources: those who were first born, those with married parents, those in households with higher income, or those with mothers who were more educated or who had higher AFQT scores. As for the detrimental externalizing impacts, the patterns were a bit different. The significant results tended to be concentrated on those children who were male, first born, and whose mothers were more educated or had higher AFQT scores.

Given that the imputation process included in the full-time sample more children with fewer resources and more children who were not first born, the differences in our results across missing data analyses move in a sensible direction.

Comparisons Between Worked Full Time in the First Year Versus Worked Part Time in the First Year

Finally we compared children whose mothers worked full time in the first year versus those whose mothers worked part time that year. The matched samples had better balanced pretreatment variables across treatment groups compared with those of the unmatched MI samples for this comparison as well. For the unmatched MI samples, 9 of the 19 covariates had significant differences in mean *t* scores ranging as high as 8.90. For the matched samples, none of the difference in mean *t* scores for either matching method was above 1.00.

Table 5 displays the results of this comparison. The complete case regression estimates on cognitive outcomes were strongly negative and statistically significant for four of five outcomes. However, our previously reported results suggest that we should have expected these effects to diminish as we moved to analyses that used more robust methods. And indeed this is what we found. Regression estimates and MWR propensity score estimates with multiply imputed data (in columns 3–6) yielded smaller estimated negative effects, although these were still statistically significant for three of the five outcomes (with effect sizes ranging from −.09 to −.19). Full matching propensity score models with multiply imputed data reduced the estimates still more, with only one that was statistically significant (effect size = −.17). Again, the sample sizes for the multiply imputed estimates were substantially bigger so these smaller estimates also had smaller standard errors. Thus, the policy conclusion of the models in columns 7 and 8 is different from the conclusion that would be drawn from columns 1 and 2. Although all the point estimates were negative, there is no evidence in columns 7 and 8 that children were significantly disadvantaged in their cognitive development if their mothers worked full time rather than part time in the first year.

As for the behavioral outcomes, there is evidence from the multiply imputed regression results that children of mothers who worked full

Table 5
Comparison of Treatment Effects

	Complete case regression[a]		MI regression[b]		MI propensity score–MWR[c]		MI propensity score–FM[b]	
Analysis	TE	*SE*	TE	*SE*	TE	*SE*	TE	*SE*
PPVT–R, ages 3–4	−4.94*	1.20	−3.07*	0.81	−3.35*	0.88	−3.14*	1.01
PIAT–M, ages 5–6	−2.43*	0.94	−1.61*	0.62	−1.28*	0.57	−1.27	0.80
PIAT–M, ages 7–8	−1.44	0.89	−0.65	0.50	−0.95	0.70	−0.68	0.81
PIAT–R, ages 5–6	−2.43*	0.92	−1.49*	0.55	−1.56*	0.59	−1.43	0.76
PIAT–R, ages 7–8	−1.88*	0.94	−0.46	0.65	−0.60	0.64	−0.44	0.88
BPI								
Internalizing, age 5–6	−0.01	0.15	0.16	0.10	0.17	0.18	0.16	0.19
Internalizing, age 7–8	0.21	0.16	0.11	0.11	0.00	0.20	0.01	0.26
Externalizing, age 5–6	0.21	0.25	0.41*	0.16	0.31	0.25	0.32	0.19
Externalizing, age 7–8	0.32	0.26	0.38*	0.16	0.12	0.24	0.12	0.26

Note. Table shows treatment effect (TE) estimates for the comparison of children of mothers who worked full-time in the first year and children of mothers who worked part-time in the first year. Effect is estimated for children of mothers who worked full-time in the first year. MI = multiple imputation; MWR = matching with replacement; FM = full matching; PPVT–R = Peabody Picture Vocabulary Test—Revised; PIAT–M = Peabody Individual Achievement Test—Math; PIAT–R = Peabody Individual Achievement Test—Reading; BPI = Behavioral Problems Index.

[a] Complete case data for cognitive outcomes (n = 986); complete case data for behavioral outcomes (n = 611).

[b] n = 3,528.

[c] n = 3,416.

* $p < .05$.

Table 6
Comparison of Treatment Effects

Analysis	Complete case regression[a]		MI regression[b]		MI propensity score–MWR[c]		MI propensity score–FM[b]	
	TE	*SE*	TE	*SE*	TE	*SE*	TE	*SE*
PPVT–R, ages 3–4	1.35	1.38	−0.30	1.15	−0.69	1.70	−0.82	1.33
PIAT–M, ages 5–6	0.73	1.06	−0.09	0.61	0.38	0.79	−0.33	0.84
PIAT–M, ages 7–8	1.99	1.01	−0.57	0.51	−0.42	0.71	−0.87	0.85
PIAT–R, ages 5–6	2.95	1.03	−1.01	0.58	−0.79	0.52	−1.21	0.82
PIAT–R, ages 7–8	1.48	1.07	−0.05	0.60	0.08	0.81	−0.56	0.82
BPI								
Internalizing, age 5–6	0.06	0.17	0.06	0.10	0.13	0.10	0.06	0.15
Internalizing, age 7–8	0.04	0.27	0.20	0.13	0.31	0.16	0.22	0.19
Externalizing, age 5–6	0.37	0.17	0.40*	0.16	0.51*	0.13	0.31	0.25
Externalizing, age 7–8	0.20	0.28	0.49*	0.22	0.73*	0.29	0.48	0.33

Note. Table shows treatment effect (TE) estimates for the comparison of children of mothers who did not work in the first 3 years of their child's life and children of mothers who worked full-time in the first year. Results from propensity score matching with replacement (MWR) and full matching (FM) models are for the effect of the treatment on the controls. That is, the treatment is considered to be working full-time in the first year, but we are making inferences about the children whose mothers did not work at all in the first 3 years of the child's life. MI = multiple imputation; PPVT–R = Peabody Picture Vocabulary Test—Revised; PIAT–M = Peabody Individual Achievement Test—Math; PIAT–R = Peabody Individual Achievement Test—Reading; BPI = Behavioral Problems Index.
[a] Complete case data for cognitive outcomes (n = 1,060); complete case data for behavioral outcomes (n = 1,485).
[b] n = 4,297.
[c] n = 2,121.
* $p < .05$.

time might exhibit more externalizing problems than they would have had their mothers worked part time (effect size of .10 for each). However, the propensity score results do not uphold this finding.

Additional Comparisons

Here we report the results of additional analyses that might be particularly relevant to examining the effects of policies such as TANF and proposed extensions of the FMLA. We begin with analyses that may be relevant for understanding TANF impacts.

We sought to determine what would happen to the children of mothers in our dataset who did not work in the first three years or who put off work until after the first year if, instead, their mothers had worked full time in the first year. Table 6 presents the first of these comparisons (never vs. full time), and the results provided no evidence for significant negative effects on children's cognitive outcomes. They did, however, show significant detrimental effects on children's externalizing behavior at both of the last assessment time points (effect sizes ranging from .11 to .19).

Table 7 might, at first glance, appear to consider the same comparison as was shown in Table 4; that is, the children of mothers who put off work until after the first year compared with those who worked full time in the first year, and indeed the regression results remained the same (because the regression model did not change). However, the propensity score effects reflect the effect of the treatment on controls this time. The treatment effects still measure the effect of working more; however, now the inference is made with respect to the children of mothers who worked less. In this case, the estimates were actually quite similar to those for the effect of the treatment on the treated with some evidence for significant negative effects on children's cognitive outcomes (effect sizes ranging from −.09 to −.15), particularly for ages 5 or 6. The significant detrimental externalizing effects were evident in this comparison as well for ages 7 or 8 (effect sizes of .11 and .14).

Although these results are suggestive, two limitations should be noted. The first is that these estimates are for all mothers who did not work in the first year or in the first three years. The effects for low-income mothers who were most likely to be affected by TANF might differ. We did attempt to estimate similar models for the low-income group most likely to be affected by TANF, but the small sample size made it difficult to draw out any strong conclusions.[14] The second limitation was that we were unable to distinguish women who worked as a result of welfare work requirements from those who worked for other reasons. If work requirements pushed a particular group of women into work in the first year who would otherwise not have worked that year, our ability to make projections about this group would be limited.

We also carried out some analyses that might be relevant to FMLA extensions. The full time versus no first and part time versus no first comparisons (Tables 3 and 4, respectively) are relevant here because they examined what would have happened to children of mothers who were currently returning to work in the first year, if, instead, they had put off returning to work until after the first year. The results indicated that there would be cognitive benefits to children at ages 5 or 6 if their mothers put off working instead of working full time in the first year (the point estimates for part time are in the same direction but smaller and not statistically significant). The results also suggest that mothers putting off work until after the first year, rather than working full time in the first year, would be associated with lower levels of externalizing problems for their children at ages 5 or 6 and at ages 7 or 8.

[14] Results are not shown but they are available from Jennifer L. Hill upon request. No estimates were statistically significant, although it is not clear whether this represents a true decrease in magnitude of effects or simply an increase in imprecision due to reduced sample sizes.

Table 7
Comparison of Treatment Effects

Analysis	Complete case regression[a]		MI regression[b]		MI propensity score–MWR[c]		MI propensity score–FM[b]	
	TE	*SE*	TE	*SE*	TE	*SE*	TE	*SE*
PPVT–R, ages 3–4	−2.61*	1.25	−1.31	0.97	−1.45	1.42	−1.44	1.58
PIAT–M, ages 5–6	−2.09*	0.98	−1.65*	0.60	−2.00*	0.78	−1.79	1.19
PIAT–M, ages 7–8	−2.48*	0.92	−1.07*	0.47	−1.37	0.73	−1.50	1.09
PIAT–R, ages 5–6	−1.75	0.91	−1.53*	0.53	−1.83*	0.45	−2.16*	0.80
PIAT–R, ages 7–8	−1.86*	0.96	−1.08	0.59	−1.01	0.71	−1.55	0.82
BPI								
Internalizing, age 5–6	0.34	0.24	0.13	0.09	0.07	0.09	0.09	0.17
Internalizing, age 7–8	0.11	0.24	0.03	0.10	0.08	0.14	−0.01	0.20
Externalizing, age 5–6	−0.16	0.39	0.28	0.15	0.15	0.20	0.19	0.25
Externalizing, age 7–8	−0.14	0.40	0.42*	0.18	0.55*	0.19	0.40	0.29

Note. Table shows treatment effect (TE) estimates for the comparison of children of mothers who did not work until the second year and children of mothers who worked full-time in the first year of their child's life. Results from propensity score matching with replacement (MWR) and full matching (FM) models are for the effect of the treatment on the controls. That is, the treatment is considered to be working full-time in the first year, but we are making inferences about the children whose mothers did not work until after the first year. MI = multiple imputation; PPVT–R = Peabody Picture Vocabulary Test—Revised; PIAT–M = Peabody Individual Achievement Test—Math; PIAT–R = Peabody Individual Achievement Test—Reading; BPI = Behavioral Problems Index.
[a] Complete case data for cognitive outcomes (n = 1,057); complete case data for behavioral outcomes (n = 651).
[b] n = 3,895.
[c] n = 1,746.
* $p < .05$.

Here too we might be concerned that our results for children of mothers who worked full time in the first year, or part time in the first year, may not be generalizable to the more specific subset of children whose mothers would be most affected by FMLA extensions. We estimated the same models for the group of children of mothers most likely to be affected by the FMLA and related policies, and the estimates (results available from Jennifer L. Hill on request) were very similar though just slightly more imprecisely measured as a result of the decrease in sample size, leading to some loss of significance.

Discussion

This article provides the first direct evidence on how sensitive estimates of the effects of early maternal employment on child outcomes are to the use of newer, and possibly more robust, methods designed to handle missing data and selection bias. Prior studies, including many with the NLSY, have identified missing data and selection bias as potential challenges to causal estimates but have not taken as rigorous an approach as was taken here.

Four sets of models were estimated, moving from the type of regression estimates with complete case data that have been standard in this literature to regression estimates with multiply imputed data to two forms of propensity score estimates with multiply imputed data. These are the first estimates to deal with missing data and selection bias in this manner, and the results indicate that these newer methods do change the findings and policy implications.

Negative effects of maternal employment on children's cognitive outcomes were found in our analyses primarily for children whose mothers were employed full time in the first year postbirth as compared with children whose mothers postponed work until after their child's first year of life and also as compared with mothers who worked part time in the first year. Negative effects in terms of increased externalizing behavioral problems were evident in each of these comparisons involving mothers who worked full time in the first year.

Our results shed light on the potential causal effects of maternal employment on child outcomes and illustrate an approach to causal inference and missing data in the context of observational studies that could be applied in many other topics in child policy. Comparing the results across methods, controlling for selection bias and imputing missing data, rather than relying on standard regression models with complete cases, does change the results. It appears that standard missing data methods might overstate the negative effects of full-time maternal employment in the first year of life on children's cognitive development, particularly at ages 3 or 4 and at ages 7 or 8, and some might miss the detrimental impacts on externalizing behavior as well. Moreover, standard regression methods that use only complete case data might overstate the advantages associated with part-time work in the first year in terms of cognitive measures.

Overall, the change from complete case analyses to MI tended to make the biggest difference in terms of our inferences, particularly for the cognitive results. The imputed sample was relatively less advantaged and consisted of a smaller proportion of first born children. Given our suggestive subgroup results, it seems possible that the differences in the full time versus no first comparison might be directly related to this change in sample composition. To be explicit, when we added more disadvantaged mothers the overall treatment effect was reduced by their smaller negative treatment effects on cognitive outcomes. Smaller negative effects of first-year maternal employment on cognitive outcomes for less-advantaged children have been found in prior studies, and this pattern has been interpreted as these children having less to lose by being in care with someone other than their mother than children who come from more advantaged families (see, for instance, Desai, Chase-Lansdale, & Michael, 1989). Another interpretation is that disadvantaged families may come from

a longer tradition of families with mothers who worked outside the home and so have developed better coping mechanisms and support systems. Thus, transitions into work might have been less traumatic for the children.

Smaller differences between regression and propensity score estimates were found. This could have occurred for several reasons. First, although the groups compared for each estimate were often noticeably different in terms of observed covariates, it appears that these differences largely reflected different weightings of the same types of people in each group (as was evidenced by the good overlap in the propensity score distributions across all pairs of maternal employment groups). This overlap across maternal employment groups in terms of observed covariates implies that regression models were likely not forced to extrapolate dangerously. It is also possible that the linear and additive specification was appropriate for these models.

Finally, it is likely that MI played a role in the observed similarities between regression and propensity score methods, for two reasons. First, the augmented samples provided by the multiply imputed data allowed for a greater possibility of comparable samples across any pair of treatment groups being compared (this is not necessarily the case, but it appears to be in this context). Second, the models used to create the imputed values force the relationship between the imputed outcomes and the covariates to be linear, so for these data linear regression models used to estimate treatment effects would hold by default. As evidence of the role of MI in this phenomenon (in results not shown), the differences between propensity score estimates and the regression estimates from the complete case samples were much more striking than these differences for the multiply imputed data.

We also conducted some analyses of comparisons that might be particularly relevant to policy choices related to welfare and family and medical leave policies. We see these results as mainly illustrative, because our efforts to estimate treatment effects for the specific groups most likely to be affected were hampered by relatively small sample sizes. Indeed, one of the most important implications of this study is the caution that must be exercised in generalizing results across populations. Our results strongly suggest that the effects of early maternal employment are not uniform but rather vary across different types of children and families. We saw evidence of this when our estimates changed as we added back in cases that were missing data, and we also saw hints of this when we compared results across subgroups. This variation by subgroup is an important topic for further research.

That having been said, it is useful to remember that all of the estimates are based on observational and not experimental data. The causal methods we used rely strongly on the selection on observables assumption. Therefore, we need to believe that we have included in our analyses all of the variables that affected both the maternal employment decision and the children's subsequent cognitive and behavioral outcomes. Maternal depression is one variable that we were not able to directly control for that might have played a role in the observed associations, particularly because there is evidence that it is correlated with maternal reports of children's behavior (Early Child Care Research Network, 2003). To the extent that this variable, or others that we were not able to include, are important variables (and have not been proxied by the variables we did include), the estimates reported here may be biased.

References

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91,* 444–472.

Angrist, J. D., & Krueger, A. (1999). *Empirical Strategies in Labor Economics,* Vol. 3A of O. Ashenfelter & D. Card (Eds.), *Handbook of Labor Economics* (1278–1329). Amsterdam: Elsevier Science.

Baer, J. S., Kivlahan, D. R., Blume, A. W., McKnight, P., & Marlatt, G. A. (2001). Brief intervention for heavy-drinking college students: 4-year follow-up and natural history. *American Journal of Public Health, 91,* 1310–1316.

Brooks-Gunn, J. (2004). Don't throw out the baby with the bathwater: Incorporating behavioral research into evaluations. *Social Policy Report, 18(2),* 14.

Brooks-Gunn, J., Han, W., & Waldfogel, J. (2002). Maternal employment and child cognitive outcomes in the first three years of life: The NICHD study of early child care. *Child Development, 73,* 1052–1072.

Chase-Lansdale, P., Mott, F. L., Brooks-Gunn, J., & Philips, D. A. (1991). Children of the National Longitudinal Survey of Youth: A unique research opportunity. *Developmental Psychology, 27,* 918–931.

Cook, T. D. (2004). Beyond advocacy: Putting history and research on research into debates about the merits of social experiments. *Social Policy Report, 18(2),* 5–6.

Cottingham, P. (2004). Why we need more, not fewer, gold standard evaluations. *Social Policy Report, 18,* 13.

D'Agostino, R. B., Jr. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17,* 2265–2281.

Dehejia, R., & Wahba, S. (1999). Causal effects in non-experimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association, 94,* 1053–1062.

Dehejia, R., & Wahba, S. (2000). *Propensity score matching methods for non-experimental causal studies* (Tech. Rep. No. 6829). Washington, DC: National Bureau of Economic Research.

Desai, S., Chase-Lansdale, P. L., & Michael, R. T. (1989). Mother or market? Effects of maternal employment on the intellectual ability of four-year-old children. *Demography, 26,* 545–561.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics, 49,* 1231–1236.

Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test—Revised.* Circle Pines, MN: American Guidance Service.

Early Child Care Research Network. (2003). Social functioning in first grade: Associations with earlier home and child care predictors with current classroom experiences. *Child Development, 74,* 1639–1662.

Efron, B., & Tibshirani, R. J. (1998). *An introduction to the bootstrap.* London: Chapman & Hall.

Fiebach, N. H. (1990). Outcomes in patients with myocardial-infarction who are initially admitted to stepdown units: Data from the Multicenter Chest Pain Study. *Biometrika, 89,* 15–20.

Foster, E. M. (2003). Propensity score matching: An illustrative analysis of dose response. *Medical Care, 41,* 1183–1192.

Foster, E. M., & McLanahan, S. (1996). An illustration of the use of instrumental variables: Do neighborhood conditions affect a young person's chance of finishing high school? *Psychological Methods, 3,* 249–260.

Fulgini, A. A., & Brooks-Gunn, J. (2000). *The healthy development of young children: SES disparities, prevention strategies, and policy opportunities*. Washington, DC: National Academy of Sciences.

Gamoran, A., Mare, R. D., & Bethke, L. (1999). Effects of nonmaternal child care on inequality in cognitive skills. *Institute for Research on Poverty Discussion Paper,* 1186–1199.

Gastwirth, J. L., Krieger, A. M., & Rosenbaum, P. R. (2000). Asymptotic separability in sensitivity analysis. *Journal of the Royal Statistical Society, Series B. Methodological, 62,* 545–555.

Han, W-J., Waldfogel, J., & Brooks-Gunn, J. (2001). The effects of early

maternal employment on later cognitive and behavioral outcomes. *Journal of Marriage and the Family, 63,* 336–354.

Harvey, E. (1999). Short-term and long-term effects of early parental employment on children of the National Longitudinal Survey of Youth. *Developmental Psychology, 35,* 445–459.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47,* 153–161.

Hill, J. L., Reiter, J. P., & Zanutto, E. L. (2004). A comparison of experimental and observational data analyses. In A. Gelman & X.-L. Meng (Eds.), *Applied Bayesian modeling and causal inference from an incomplete-data perspective.* New York: Wiley.

Hill, J. L., Waldfogel, J., & Brooks-Gunn, J. (2002). Differential effects of high-quality child care. *Journal of Policy Analysis and Management, 21,* 601–627.

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

Huber, P. (1967). The behavior of maximum likelihood estimates under non-standard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 221–223). Berkeley: University of California Press.

Imbens, G. W., Rubin, D. B., & Sacerdote, B. (2001). Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a sample of lottery players. *American Economic Review, 91,* 778–794.

James-Burdumy, S. (1999). *The effect of maternal labor force participation on child development* (Document No. PP05–02). Princeton, NJ: Mathematica Policy Research.

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association, 91,* 222–230.

Karoly, L. A., Greenwood, P. W., Everingham, S. S., Hoube, J., Kilburn, M. R., Rydell, C. P., et al. (1998). *Investing in our children: What we know and don't know about the costs and benefits of early childhood interventions.* Santa Monica, CA: RAND.

Kohen, D., Brooks-Gunn, J., McCormick, M., & Graber, J. (1997). Concordance of maternal and teacher ratings of school and behavior problems in children of varying birth weights. *Journal of Developmental and Behavioral Pediatrics, 18,* 295–306.

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics, 114,* 497–532.

Lavori, P. W., Keller, M. B., & Endicott, J. (1995). Improving the aggregate performance of psychiatric diagnostic methods when not all subjects receive the standard test. *Statistics in Medicine, 14,* 1913–1925.

Lechner, M. (1999). Earnings and employment effects of continuous off-the-job training in East Germany after unification. *Journal of Business & Economic Statistics, 17,* 74–90.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

McCall, R. B., & Green, B. L. (2004). Beyond the methodological gold standards of behavioral research: Considerations for practice and policy. *Social Policy Report, 18(2),* 3–19.

Ming, K., & Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics, 56,* 118–124.

Morris, P. (2002). The effects of welfare reform policies on children. *Social Policy Report, 16,* 4–18.

Morris, P., Duncan, G., & Rodriguez, C. (2004). *Using welfare reform experiments to estimate the impact of income on child achievement.* Unpublished manuscript, Northwestern University.

Morris, P., Gennetian, L. A., & Duncan, G. J. (2005). Effects of welfare and employment policies on young children: New findings on policy experiments conducted in the early 1990's. *Social Policy Report, 19,* 3–17.

Mosteller, F. (1995). The Tennessee Study of class size in the early school grades. *The Future of Children, 5,* 113–127.

Phillips, M., Brooks-Gunn, J., Duncan, G., Klebanov, P., & Crane, J. (1998). Family background, parenting practices, and the Black–White test score gap. In C. Jencks & M. Phillips (Eds.), *The Black–White test score gap* (pp. 103–145).Washington, DC: Brookings Institution Press.

Robins, J. M., & Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine, 16,* 285–319.

Rosenbaum, P. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society, Series B. Methodological, 53,* 597–610.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *The Journal of the Royal Statistical Society, Series A. 147,* 656–666.

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,* 516–524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39,* 33–38.

Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics, 29,* 185–203.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics, 6,* 34–58.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association, 74,* 318–328.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association, 95,* 573–585.

Ruhm, C. (2003). Parental employment and child cognitive development. *Journal of Human Resources, 39*(1), 155–192.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* London: Chapman & Hall.

Shonkoff, J. P., & Phillips, D. (2000). *From neurons to neighborhoods: The science of early child development.* Washington, DC: National Academies Press.

Sianesi, B. (2004). An evaluation of the Swedish system of active labor market programs in the 1990s. *The Review of Economics and Statistics, 86,* 133–135.

Smolensky, E., & Gootman, J. A. (Eds.). (2003). *Working families and growing kids: Caring for children and adolescents.* Washington, DC: National Academies Press.

Waldfogel, J., Han, W., & Brooks-Gunn, J. (2002). The effects of early maternal employment on child cognitive development. *Demography, 39,* 369–392.

Appendix A

Rates (in Percentages) of Missing Data for All Variables by Treatment Group

Variable	Missing observations				Missing rate overall (n = 6,114)
	Never (n = 1,494)	No first (n = 1,092)	Part time (n = 725)	Full time (n = 2,803)	
Worked before first born	0	0	0	0	0
Mom's mom worked	46	37	33	32	36
Ethnicity	0	0	0	0	0
Female	0	0	0	0	0
Child is first born	0	0	0	0	0
Married	4	2	1	2	2
Household in poverty	5	3	6	6	5
Mom's education at birth	0	0	0	0	0
Family size	8	6	7	8	8
Log income	27	22	20	21	23
Child's age in 2000 (months)	0	0	0	0	0
Child's first assessment age	5	3	6	6	5
Mom's age at birth	0	0	0	0	0
Mom's IQ test	6	5	5	3	4
Birth weight (low)	9	5	6	6	7
Preterm birth (weeks)	0	0	0	0	0
PPVT–R, ages 3–4	31	29	29	31	31
PIAT–M, ages 5–6	17	14	17	16	16
PIAT–M, ages 7–8	22	23	19	19	20
PIAT–R, ages 5–6	20	15	19	18	18
PIAT–R, ages 7–8	22	23	19	19	20
BPI					
Internalizing, age 4	56	57	57	57	57
Externalizing, age 4	56	58	57	58	57
Internalizing, ages 5–6	17	14	15	15	15
Externalizing, ages 5–6	17	13	15	16	16
Internalizing, ages 7–8	26	23	20	23	23
Externalizing, ages 7–8	26	23	21	23	24

Note. This table lists the percentage of missing observations for each variable listed. The overall rate is in the last column. The first four columns show these rates broken down by maternal employment group. PPVT–R = Peabody Picture Vocabulary Test—Revised; PIAT–M = Peabody Individual Achievement Test—Math; PIAT–R = Peabody Individual Achievement Test—Reading; BPI = Behavioral Problems Index.

Appendix B

Multiple Imputation

Viewed in a causal framework (Holland, 1986; Rubin, 1978), we considered outcomes for individuals in different treatment groups to represent different variables (potential outcomes) and imputed accordingly (each outcome variable was translated into a separate outcome variable for each treatment, with missing values for anyone not in that treatment group). This strategy allowed for the maximum amount of freedom in how relationships between each outcome in each treatment group and the predictor variables were defined. In the absence of such an approach, it is possible for treatment effects to be artificially biased toward zero (for a discussion of MI combined with propensity score matching in a related context see Hill et al., 2004, and for a more general example see Hill et al., 2004).

(*Appendixes continue*)

Table B1

Means and Standard Deviations of Background Variables Across Treatment Groups for the Different Data Sets

	Overall[a]		Never[b]		No first[c]		Part time[d]		Full time[e]	
Variable and data set	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Work before first born										
CC-C	0.76	0.43	0.39	0.49	0.55	0.50	0.83	0.38	0.94	0.23
CC-B	0.76	0.43	0.39	0.49	0.57	0.50	0.83	0.38	0.94	0.23
MI	0.69	0.46	0.32	0.47	0.53	0.50	0.85	0.35	0.91	0.28
Mom's mom worked										
CC-C	0.38	0.49	0.42	0.49	0.38	0.49	0.39	0.49	0.36	0.48
CC-B	0.37	0.48	0.39	0.49	0.39	0.49	0.39	0.49	0.35	0.48
MI	0.36	0.48	0.37	0.48	0.38	0.48	0.38	0.49	0.36	0.48
Hispanic										
CC-C	0.16	0.37	0.19	0.39	0.18	0.39	0.14	0.35	0.15	0.36
CC-B	0.16	0.37	0.19	0.89	0.17	0.38	0.11	0.32	0.17	0.37
MI	0.20	0.40	0.23	0.42	0.22	0.41	0.14	0.35	0.20	0.40
Black										
CC-C	0.29	0.45	0.27	0.45	0.35	0.48	0.18	0.39	0.30	0.46
CC-B	0.28	0.45	0.26	0.44	0.34	0.48	0.17	0.38	0.27	0.45
MI	0.28	0.45	0.31	0.46	0.30	0.46	0.17	0.37	0.29	0.45
White										
CC-C	0.55	0.50	0.54	0.50	0.47	0.50	0.68	0.47	0.55	0.50
CC-B	0.56	0.50	0.55	0.50	0.49	0.50	0.71	0.45	0.56	0.50
MI	0.52	0.50	0.46	0.50	0.48	0.50	0.69	0.46	0.51	0.50
Female										
CC-C	0.49	0.50	0.49	0.50	0.44	0.50	0.48	0.50	0.51	0.50
CC-B	0.48	0.50	0.47	0.50	0.44	0.50	0.49	0.50	0.51	0.50
MI	0.49	0.50	0.49	0.50	0.47	0.50	0.50	0.50	0.50	0.50
Child is first born										
CC-C	0.41	0.49	0.26	0.44	0.38	0.49	0.39	0.49	0.48	0.50
CC-B	0.43	0.50	0.27	0.44	0.40	0.49	0.38	0.49	0.48	0.50
MI	0.38	0.49	0.26	0.44	0.37	0.48	0.40	0.49	0.45	0.50
Married										
CC-C	0.73	0.45	0.63	0.49	0.70	0.46	0.84	0.37	0.75	0.44
CC-B	0.73	0.44	0.65	0.48	0.69	0.46	0.84	0.37	0.76	0.43
MI	0.66	0.48	0.55	0.50	0.64	0.48	0.80	0.40	0.68	0.47
Household in poverty										
CC-C	0.73	0.41	0.39	0.49	0.32	0.47	0.17	0.37	0.12	0.33
CC-B	0.21	0.41	0.40	0.49	0.33	0.47	0.17	0.37	0.13	0.33
MI	0.26	0.44	0.42	0.49	0.34	0.47	0.15	0.36	0.17	0.38
<High school										
CC-C	0.18	0.38	0.33	0.47	0.24	0.43	0.13	0.33	0.11	0.31
CC-B	0.18	0.38	0.32	0.47	0.27	0.44	0.12	0.33	0.12	0.32
MI	0.31	0.46	0.47	0.50	0.37	0.48	0.21	0.40	0.22	0.42
High school										
CC-C	0.45	0.50	0.41	0.49	0.52	0.50	0.41	0.49	0.45	0.50
CC-B	0.46	0.50	0.41	0.49	0.50	0.50	0.40	0.49	0.45	0.50
MI	0.38	0.49	0.33	0.47	0.42	0.49	0.37	0.48	0.40	0.49
<College										
CC-C	0.24	0.43	0.18	0.38	0.19	0.39	0.29	0.46	0.26	0.44
CC-B	0.22	0.42	0.18	0.38	0.18	0.38	0.29	0.46	0.26	0.44
MI	0.20	0.40	0.13	0.34	0.16	0.37	0.25	0.43	0.23	0.42
College										
CC-C	0.13	0.34	0.09	0.28	0.05	0.23	0.17	0.38	0.17	0.37
CC-B	0.14	0.35	0.09	0.29	0.06	0.23	0.19	0.39	0.17	0.38
MI	0.11	0.32	0.06	0.24	0.06	0.23	0.18	0.38	0.14	0.35
Family size										
CC-C	3.21	1.67	3.43	1.82	3.52	1.82	3.24	1.57	3.02	1.56
CC-B	3.24	1.75	3.50	1.82	3.62	2.03	3.13	1.43	3.08	1.70
MI	3.44	1.86	3.69	1.94	3.69	1.97	3.21	1.61	3.27	1.82
Log scaled income										
CC-C	10.01	1.30	9.63	1.80	9.58	1.57	10.16	1.24	10.26	0.84
CC-B	10.02	1.32	9.63	1.74	9.60	1.54	10.18	1.16	10.24	0.96
MI	9.89	1.41	9.52	1.75	9.60	1.63	10.22	1.05	10.12	1.09
Child's age in 2000										
CC-C	159.04	38.97	160.59	40.99	171.42	37.56	160.42	38.58	153.73	37.81
CC-B	161.03	39.72	159.55	40.88	171.42	36.92	159.27	39.71	151.60	39.51
MI	155.18	44.53	157.69	44.47	165.96	42.62	153.26	44.89	150.15	44.35

(*table continues*)

Table B1 (*continued*)

	Overall[a]		Never[b]		No first[c]		Part time[d]		Full time[e]	
Variable and data set	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Mom's age at birth										
CC-C	25.04	3.62	24.86	3.81	23.82	3.49	25.08	3.52	25.52	3.52
CC-B	24.91	3.64	25.20	3.90	23.72	3.53	25.25	3.67	25.63	3.60
MI	25.37	4.06	25.23	4.12	24.32	4.07	25.88	4.01	25.73	3.96
Mom's IQ test										
CC-C	52.49	25.41	46.47	26.65	47.59	23.10	61.32	25.14	54.05	24.97
CC-B	54.01	25.50	46.12	26.04	48.35	23.65	62.58	24.94	54.25	25.06
MI	47.06	26.17	38.86	25.91	43.09	23.98	58.19	26.60	50.09	25.42
Birth weight										
CC-C	118.40	19.92	117.04	22.84	118.61	18.67	119.88	18.56	118.43	19.56
CC-B	119.08	19.90	117.52	22.10	117.30	22.08	119.95	19.13	118.40	20.22
MI	117.70	21.06	116.10	22.45	117.46	21.42	120.07	20.87	118.03	20.11
Preterm										
CC-C	0.70	1.62	0.83	1.79	0.59	1.51	0.68	1.62	0.71	1.60
CC-B	0.69	1.60	0.79	1.80	0.70	1.69	0.62	1.51	0.74	1.72
MI	0.69	1.68	0.72	1.73	0.65	1.59	0.67	1.80	0.70	1.65

Note. Table B1 displays the means and standard deviations for all background variables in each of three different samples. CC-C denotes the sample that retains observations that have no missing data on either background variables or the cognitive outcomes. CC-B denotes the sample that retains observations that have no missing data on either background variables or the behavioral outcomes. MI denotes the multiply imputed sample.

[a] CC-C (n = 1,543); CC-B (n = 2,129); MI (n = 6,114).

[b] CC-C (n = 280); CC-B (n = 389); MI (n = 1,494).

[c] CC-C (n = 277); CC-B (n = 369); MI (n = 1,092).

[d] CC-C (n = 206); CC-B (n = 275); MI (n = 725).

[e] CC-C (n = 780); CC-B (n = 1,096); MI (n = 2,803).

(*Appendixes continue*)

Table B2

Means and Standard Deviations of Outcome Variables Across Treatment Groups for the Different Data Sets

	Overall[a]		Never[b]		No first[c]		Part time[d]		Full time[e]	
Variable and data set	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
PPVT–R, ages 3–4										
CC-C	89.44	20.56	86.40	20.80	87.50	19.51	96.17	17.47	89.45	18.31
MI	86.72	20.25	81.78	21.56	85.31	20.06	94.42	18.01	87.91	19.37
PIAT–M, ages 5–6										
CC-C	101.12	12.32	99.58	12.99	100.28	13.13	104.43	12.52	101.09	12.46
MI	99.03	13.85	95.99	14.19	98.63	13.75	102.84	13.49	99.81	13.46
PIAT–M, ages 7–8										
CC-C	102.14	13.32	101.30	12.74	101.78	12.47	104.48	11.31	101.95	12.22
MI	100.43	12.83	97.97	13.21	99.79	13.09	103.32	12.13	101.25	12.46
PIAT–R, ages 5–6										
CC-C	105.62	10.67	104.61	12.89	104.23	10.79	107.68	12.20	105.93	12.31
MI	104.35	13.65	101.67	14.02	103.51	12.91	107.20	13.43	105.37	13.52
PIAT–R, ages 7–8										
CC-C	106.11	13.56	104.60	13.72	105.31	12.67	108.62	12.05	106.28	12.45
MI	104.07	13.59	101.06	14.70	103.17	13.58	106.67	12.83	105.36	12.83
BPI Internalizing, ages 5–6										
CC-B	2.49	2.25	2.62	2.47	2.66	2.30	2.34	2.15	2.43	2.16
MI	2.61	2.24	2.75	2.37	2.69	2.31	2.34	2.09	2.56	2.18
BPI Internalizing, ages 7–8										
CC-B	2.45	2.29	2.45	2.37	2.58	2.34	2.27	2.05	2.45	2.31
MI	2.64	2.33	2.68	2.37	2.82	2.42	2.46	2.26	2.59	2.28
BPI Externalizing, ages 5–6										
CC-B	5.59	3.75	5.91	3.98	5.89	3.80	5.18	3.66	5.48	3.66
MI	5.65	3.75	5.76	3.85	5.87	3.92	5.18	3.52	5.62	3.69
BPI Externalizing, ages 7–8										
CC-B	5.62	3.83	6.02	3.98	5.87	3.87	5.27	3.61	5.48	3.81
MI	5.80	3.88	5.98	4.03	5.96	3.97	5.42	3.65	5.75	3.81

Note. Table B2 displays the means and standard deviations for all outcome variables in each of three different samples. CC-C denotes the sample that retains observations that have no missing data on either background variables or the cognitive outcomes. CC-B denotes the sample that retains observations that have no missing data on either background variables or the behavioral outcomes. MI denotes the multiply imputed sample. PPVT–R = Peabody Picture Vocabulary Test—Revised; PIAT–M = Peabody Individual Achievement Test—Math; PIAT–R = Peabody Individual Achievement Test—Reading; BPI = Behavioral Problems Index.

[a] CC-C (n = 1,543); CC-B (n = 2,129); MI (n = 6,114).

[b] CC-C (n = 280); CC-B (n = 389); MI (n = 1,494).

[c] CC-C (n = 277); CC-B (n = 369); MI (n = 1,092).

[d] CC-C (n = 206); CC-B (n = 275); MI (n = 725).

[e] CC-C (n = 780); CC-B (n = 1,096); MI (n = 2,803).

Appendix C

Full Matching Algorithm for Effect of Treatment on Treated

The following algorithm can be used to create a full matching partition appropriate for estimating the effect of the treatment on the treated.

1. Order treatment units by their propensity scores.

2. Create an initial partition by dividing all observations into adjacent and nonoverlapping subclasses based on their propensity scores where the dividing lines between subclasses are the midpoints between the propensity scores of two adjacent treatment units. In this initial partition, each subclass will have at least one treatment unit but possibly no control units.

3. Starting with the treatment unit with the highest propensity score, examine each treatment unit in turn. For each treatment unit that has no control units in its subclass do the following: (a) Identify the control unit with the closest propensity score; if multiple control units fit this criterion, randomly sample to choose one; (b) if this control unit is the only control in its subclass, merge that subclass with the subclass with the original treatment unit (so it now contains two treated and one control unit); (c) if this control unit is one of two or more controls in its subclass, simply move it to the subclass of the original treatment unit.

The resulting partition divides the data into non-overlapping subclasses in which each treatment unit is matched to its closest control as well as possibly other controls with close propensity scores.

Received August 2, 2004
Revision received March 21, 2005
Accepted March 29, 2005 ■