Differential Effects of High-Quality Child Care

Jennifer Hill
Jane Waldfogel
Jeanne Brooks-Gunn

Abstract

In policy research a frequent aim is to estimate treatment effects separately by subgroups. This endeavor becomes a methodological challenge when the subgroups are defined by post-treatment, rather than pre-treatment, variables because if analyses are performed in the same way as with pre-treatment variables, causal interpretations are no longer valid. The authors illustrate a new approach to this challenge within the context of the Infant Health and Development Program, a multisite randomized study that provided at-risk children with intensive, center-based child care. This strategy is used to examine the differential causal effects of access to high-quality child care for children who would otherwise have participated in one of three child care options: no non-maternal care, home-based non-maternal care, and center-based care. Results of this study indicate that children participating in the first two types of care would have gained the most from high-quality center-based care and, moreover, would have more consistently retained the bulk of these positive benefits over time. These results may have implications for policy, particularly with regard to the debate about the potential implications of providing universal child care. © 2002 by the Association for Public Policy Analysis and Management.

INTRODUCTION

As a greater percentage of women with young children join the labor force (Bachu and O'Connell, 1998), an increasing number of children are being placed in non-maternal child care arrangements at younger ages and for longer periods than ever before (Phillips and Adams, 2001). Accumulating evidence that children's early experiences can have lasting effects on subsequent development (Shonkoff and Phillips, 2000) highlights the importance of making child care available that is of sufficient quality to prepare children for future academic growth. Effects of potential policies such as universal provision of center-based child care (even high-quality child care) are as yet unclear, however. While consensus seems to be emerging regarding the benefits of center-based care over more informal arrangements (Shonkoff and Phillips, 2000), whether center-based care is better than maternal care is still debated (Belsky, in press). In addition these effects would be expected to vary based on the quality of the center-based care involved.

Manuscript received received October 2001; review completed December 2001; revision completed March 2002; accepted May 2002

Journal of Policy Analysis and Management, Vol. 21, No. 4, 601–627 (2002)

Published by Wiley Periodicals, Inc. Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/pam.10077

The primary goal of this paper is to apply a new methodological strategy to examine the differential causal impacts of access to high-quality center-based child care for children who would otherwise participate in one of three different child-care options: no non-maternal care, home-based non-maternal care, or existing center-based care. This paper takes advantage of the randomized Infant Health and Development Program (IHDP) experiment, which provided very high-quality child care to children from ages 1 to 3. The authors introduce methods that allow refinement of the analyses to focus on comparisons between IHDP center-based care and these three different types of existing child care. The complication in performing these comparisons is that these child-care choices reflect post-treatment (mediating) decisions and were thus affected by the randomization—in this case, those randomized to receive the intervention were provided with an option to which those in the control group had no access. Therefore it is unclear at the outset even how to define these subgroups. For example, if all IHDP children participated in those centers, the subgroup of "no non-maternal care" would contain only control group members. Yet, subgroup definitions must mean the same thing to both randomization groups.

Accordingly, the authors use an innovative approach (extending from conceptual work by Frangakis and Rubin [2002]) that defines subgroups of people based on their behavior (choices of child care) both in the presence of and absence of the intervention (only one of which is ever observed for any individual). This framework places this methodological challenge back into the context of a familiar estimation issue, viz., subgroup analyses. The complication now becomes how to identify these subgroups, since for some people it is based on information that cannot be observed. An adaptation of propensity score matching (Rosenbaum and Rubin, 1983) is used to address this issue.

Results indicate that for the children who would have experienced, in the absence of the intervention, some home-based (non-maternal) care but no center-based care from ages 1 to 3, the effects of high-quality center-based child care are quite large and endure fairly strongly and consistently through to age 8. For children who would have been cared for by their mother in these years (in the absence of the intervention), these effects are nearly as strong and consistent. For children who would have experienced center-based care anyway, however, the effects are strong at the end of the intervention at age 3 (indicating a quality effect) but appear to diminish in subsequent years. Some descriptive evidence suggests that this attenuation may reflect the fact that the control children who used center-based care during the intervention years and the intervention children who would have done so as well (had they never been exposed to the intervention) participate in more similar (existing, center-based) care post-intervention. However, the differences in treatment effects across these subgroups may result from the fact that different types of families (and thus children) selected into the different types of non-intervention care.

Other Studies

Many observational studies (see Blau, 2001 for a critical review) have examined the effect of heterogeneity in child care quality by collecting detailed information on the child care settings the study participants used. The observational nature of this research, however, leaves analyses vulnerable to selection bias. That is, families that chose to place their children in relatively high-quality child care tend to differ from families that chose (or were forced by financial or logistical constraints) to place their children in lower quality child care. If these differences in family characteris-

tics also affected children's development, it can be difficult or impossible to disentangle these effects from those caused by differential child care quality. Currie and Thomas (1995) make one of the few aggressive attempts to address this bias through the use of an innovative sibling-comparison analysis, however, even this strategy is not without problems.

Another potential shortcoming of observational studies is that, since they focus on existing heterogeneity in child care quality, they do not necessarily provide insight into the types of quality effects that might be evident if even better child care quality existed (for example, through government programs or due to changes in government regulation of the child care industry). Moreover, the existing heterogeneity may possibly be narrow enough to preclude discernable and lasting differences in subsequent developmental outcomes.

Examples of experimental programs (for reviews see Currie, 2001 or Karoly et al., 1998) that have provided very high-quality child care to high-risk young children include the Abecedarian Project, Project CARE, and the intervention examined here, the IHDP. Randomized evaluations of these interventions all show positive benefits (for different mixes of cognitive and behavioral measures) though for most measures these effects fade over time (some still remain positive and significant but all are reduced in magnitude, often by at least one-half). Abecedarian and Project Care, however, both targeted extremely at-risk children and started long enough ago (1972 and 1978, respectively) that the mix of child care choices made by the control populations is potentially quite different from what it would be today; both of these factors limit the generalizability of the findings. The IHDP began somewhat later (1985) and targeted premature, low-birth-weight infants, a subgroup of whom more closely resemble average infants. To date, however, analyses of all three of these experimental studies have generally failed to investigate more detailed aspects of the differences in child care experiences across the children involved, such as what the differential effect would have been on control children who, in the absence of the intervention, participated in one of several alternate child care experiences.

Therefore, although past literature has addressed the question of the extent to which differences in the quality of child care affects children's development, most of this research suffers either from self-selection into child care categories or lack of generalizability of results due to targeting of particularly at-risk children. Moreover, none of the past studies has attempted to break down the effect of very high-quality child care based on what kind of care the child would have experienced in its absence. This last question is particularly important with regard to the debate about the provision of universal center-based child care, for instance as opposed to child care provision or subsidies solely for working mothers. Some critics argue that this option might lead to worse outcomes for children, particularly those who otherwise would have been cared for primarily by their mother. Results of this study speak to this important question and offer evidence in support of high-quality universal center-based child care.

A Challenge for Policy Research

The problem of how to appropriately assess the role of mediating (post-randomization) variables is widespread in policy research. Simple examples that occur frequently in intervention research involve examination of the effect of a program on study participants who actually use the program; these can be referred to as dosage or non-compliance analyses or sometimes as analyses of participation or "take-up"

rates. Examples include analyses of the effect of a school choice program on participants who actually attended private school (Barnard et al., 1998, 2001) and the effect of a housing voucher program on the individuals who actually took advantage of the vouchers (Katz, Kling, and Liebman, 2001; Leventhal and Brooks-Gunn, 2001). These examples can become more complicated when partial dosage levels are observed (for another application of the strategy described in this paper, see, for example, Hill, Brooks-Gunn, and Waldfogel, 2001) or when multiple program components are offered (a recent and interesting example, related to the strategy presented here, is illustrated in Gibson, 2001).

As an example where the link between the treatment offered and the mediating variables was slightly more indirect, consider the RAND Health Insurance Experiment (Newhouse et al., 1993). This large-scale randomized field experiment was designed to investigate the effect of a variety of health insurance plans on health care spending and utilization, as well as measures of health status. Although the straight experimental effects were indeed of interest in this evaluation, the processes by which changes occur were also explored. In particular, when improved health status was observed, to what extent could it be attributed to an increase in health care utilization as opposed to differing physician behavior across plans? The researchers attempted some relatively simple analyses to address these questions (see for instance Newhouse et al., 1993, p. 225) while acknowledging the potential biases inherent in these efforts.

A more recent example is discussed within the National Evaluation of Welfare-to-Work Strategies study (Zaslow, McGroder, and Moore, 2000). Within the context of a randomized study, an attempt was made to determine whether the effect on family or parental outcomes (e.g., employment status, parenting skills, human capital development) represented the pathways to change for children's outcomes (health, cognitive development, behavioral and emotional adjustment). The relationships are explored but the techniques used do not allow for clear determinations of differential effects for children across different types of parental outcomes.

The methodological strategy presented in this paper provides a potentially more flexible alternative to the approaches used in the above analyses because it could be used in any of the above settings. It comes with its own set of assumptions, however, which, although arguably less stringent than standard approaches in many of the situations in which it is applicable, still must be closely considered in any potential application.

THE INFANT HEALTH AND DEVELOPMENT PROGRAM AND STUDY DATA

The data comprise a subset of 361 subjects (the heavier low-birth-weight children) across eight sites who participated in an evaluation of the IHDP. The IHDP was an intervention targeting pre-term, low-birth-weight (LBW) infants.[1] The focus in this study is on the heavier of the LBW infants (those between 2001 and 2500 grams) because they can be considered to be quite similar to normal-weight babies in terms of their expected developmental trajectory (McCormick et al., 1992). The goal of the IHDP was to help counteract the negative effects of LBW on children's subsequent cognitive and behavioral developmental outcomes. The intervention consisted of two primary components: intensive, high-quality child

[1]For more complete descriptions of the program see Brooks-Gunn et al. (1993), Infant Health and Development Program (1990), and Ramey et al. (1992)

care provided in child development centers (CDCs) available for children between 12 months and 36 months of age; and home visits with the mother by trained staff available once a week in the first year of the child's life and every other week in the following 2 years. Parent meetings were also available somewhat regularly in years 2 and 3 of the child's life. Participants were randomized either to the intervention arm or a control arm. All study participants received pediatric surveillance that included referrals to appropriate resources. Henceforth the intervention group will be referred to as IHDP and the follow-up-only group as FU. Papers describing the effects of this intervention reveal that it had generally very strong positive effects on cognitive development by the end of the program; these effects faded over time, however, to be smaller and not statistically significant (Brooks-Gunn et al., 1994, McCarton et al., 1997; McCormick et al., 1998). Recent work (Hill, Brooks-Gunn, and Waldfogel, 2001) reveals that this attenuation is less noticeable for children attending a high percentage of the available CDC days.

Three characteristics of the IHDP study make it useful for examining differential effects of child care quality. First, researchers collected data on the child care choices of parents whose children did not receive the intervention (that is, those randomized to the FU arm). This creates an opportunity to examine how treatment effects vary across FU child-care subgroups. Second, the CDCs provided extremely high-quality center-based care. This creates variation in the level of quality in care that is greater than would exist in the typical range of child care options available to parents, which may make it easier to detect consequent differences in outcomes.[2] Third, the initial randomization of the study ensures that control groups exist for each of the subgroups examined. This last point will be discussed in greater detail in the next section.

Child Care Measures (Mediating Variables)

The variables most central to the primary analyses are the child care choice categorizations made for children during the intervention years, but in the absence of the intervention: no non-maternal care; some non-maternal but no center-based care (home-based); and some center-based care. For a child to be classified as engaging in non-maternal or center-based care, he must have participated in more than four hours a week of this type of care. Note that home-based care encompasses child care arrangements that range from paternal and other relative care, to sitter care, to family day care settings. For the control group (the analyses ignore values of this variable for the IHDP group) the rate of missing data for this measure is 13.5 percent.

In addition, measures were used to note whether a child had center-based care between ages 3 and 4 (10 percent missing), and whether the mother had additional births within 36 months of the study child's birth (9 percent missing). All of these measures are post-randomization variables that are not the primary outcomes of interest but that might represent a path through which the intervention affects the outcomes; thus they can be described as mediating variables.

[2]The authors know relatively little about the quality of center-based and family-based child care available in the eight sites. Given that the sites are all based near universities, formal care arrangements in these areas may have been of higher quality than the norm for the United States, which would reduce the amount of variability in the care used by the control groups (and likely dampen estimated treatment effects).

Table 1. Background and outcome variables across subgroups and treatment, means and standard errors.

	Follow-up Only							
	No non-maternal child care		Home-based child care		Center-based child care		IHDP Intervention	
	N=44		N=107		N=68		N=120	
Background Variables	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
Child								
birth weight	2244	22	2269	13	2240	17	2257	12
head circumference at birth	31.61	0.21	31.41	0.15	31.36	0.20	31.23	0.11
sex of child	0.51	0.09	0.45	0.05	0.46	0.06	0.51	0.04
weeks pre-term	4.92	0.22	5.04	0.14	5.10	0.20	5.21	0.14
birth order	2.08	0.15	1.96	0.10	1.81	0.11	1.92	0.09
neonatal health index	103.3	2.1	97.7	1.5	97.3	2.0	99.5	1.2
child a twin	0.08	0.04	0.06	0.02	0.04	0.03	0.06	0.02
Mother								
age	23.8	0.9	24.7	0.6	26.3	0.9	24.1	0.5
black	0.51	0.08	0.50	0.05	0.49	0.06	0.46	0.04
Hispanic	0.10	0.05	0.17	0.04	0.06	0.03	0.11	0.03
white	0.39	0.08	0.34	0.05	0.45	0.06	0.44	0.04
married at birth	0.43	0.08	0.42	0.05	0.63	0.06	0.45	0.04
less than high school	0.43	0.08	0.38	0.05	0.31	0.06	0.49	0.04
high school	0.23	0.07	0.27	0.04	0.21	0.05	0.27	0.04
some college	0.27	0.07	0.20	0.04	0.27	0.06	0.11	0.03
finished college	0.07	0.04	0.15	0.03	0.21	0.05	0.13	0.03
cigarettes during pregnancy	0.33	0.08	0.31	0.05	0.24	0.06	0.37	0.04
alcohol during pregnancy	0.06	0.04	0.13	0.03	0.18	0.05	0.13	0.03
drugs during pregnancy	0.00	0.00	0.05	0.02	0.01	0.02	0.04	0.02
worked during pregnancy	0.44	0.09	0.58	0.05	0.65	0.06	0.53	0.05
prenatal care	1.00	0.01	0.95	0.02	0.92	0.03	0.96	0.02
Father								
black	0.49	0.08	0.51	0.05	0.50	0.06	0.47	0.04
Hispanic	0.14	0.05	0.16	0.04	0.06	0.03	0.13	0.03
white	0.37	0.08	0.33	0.05	0.44	0.06	0.39	0.04
Sites								
Arkansas	0.12	0.05	0.09	0.03	0.21	0.05	0.15	0.03
New York	0.12	0.05	0.20	0.04	0.08	0.03	0.11	0.03
Harvard	0.19	0.06	0.14	0.03	0.11	0.04	0.10	0.03
Miami	0.10	0.05	0.10	0.03	0.06	0.03	0.11	0.03
Pennsylvania	0.11	0.05	0.06	0.02	0.16	0.05	0.15	0.03
Texas	0.10	0.05	0.15	0.04	0.14	0.04	0.12	0.03
Washington	0.14	0.06	0.11	0.03	0.19	0.05	0.11	0.03
Yale	0.11	0.05	0.13	0.03	0.05	0.03	0.15	0.03
Outcome Variables								
PPVT-R (age 3)	80.2	3.1	81.1	1.9	87.1	2.4	92.3	1.4
S-B (age 3)	80.5	3.0	82.5	2.0	90.5	2.3	97.5	1.6
PPVT-R (age 5)	73.9	4.3	78.0	2.6	84.6	2.8	84.8	2.1
WPPSI: full-scale (age 5)	89.0	2.6	92.0	1.8	95.6	2.2	94.7	1.5
WPPSI: verbal (age 5)	87.8	2.9	90.3	1.8	92.6	2.1	93.4	1.5
WPPSI: performance (age 5)	92.3	2.3	95.5	1.7	99.5	2.2	97.3	1.5
PPVT-R (age 8)	83.2	3.7	88.8	2.7	86.7	2.8	90.4	2.1
WISC: full-scale (age 8)	91.0	2.6	91.8	1.8	95.7	2.3	94.8	1.6
WISC: verbal (age 8)	92.7	2.8	94.0	1.8	97.3	2.3	96.9	1.6
WISC: performance (age 8)	90.6	2.4	91.1	1.7	94.6	2.3	93.6	1.5

Background Variables

Background variables measured at baseline include a wide variety of measures pertaining to both children and parents. The analyses use the pre-treatment variables listed in Table 1. The child's weight and head circumference, recorded at birth, in addition to the sex, number of weeks pre-term, birth order, a neonatal health index (Scott et al., 1989), and twin status were all considered predictive of developmental outcomes. Characteristics of the child's mother perceived as relevant include her age, ethnicity, marital status, and educational status at the time the child was born, indicators for substance abuse while pregnant, and indicators for whether she worked and whether she received prenatal care during pregnancy. Father's ethnicity is also included. Finally, the sites in which the participants resided (as listed in the table) are used as pre-treatment variables. Sites could be acting as proxies for a variety of sociodemographic factors or child care supply issues.

Fortunately, all of these baseline characteristics are fully observed, with the exception of the indicator for whether the child's mother worked during pregnancy. This variable is missing for just fewer than 6 percent of participants used in this study.

Table 1 presents means and associated standard errors for a variety of characteristics for FU children who during the intervention years (ages 1 to 3) experienced each of the three child care options, as well as for the children who received the IHDP intervention. The first two columns correspond to FU children who experienced no non-maternal care between the ages of 1 and 3. The third and fourth columns correspond to FU children who experienced some non-maternal care during this period but no center-based care (home-based). The fifth and sixth columns correspond to FU children who experienced at least some center-based care. The last two columns of the table present these means and standard errors for the IHDP children.

The means in the first six columns of Table 1 highlight the differences in the types of people who self-select into each of these child-care arrangements. Children later in the birth order or with higher values of the neonatal health index are more likely to be cared for solely by their mother. The ethnicity of both parents seems to play a role, with relatively lower percentages of Hispanics as compared with whites and blacks choosing center-based care. Older, married, more educated mothers and those who worked during pregnancy are more likely to have children participating in center-based care than in other arrangements. Drug and alcohol use during pregnancy also appear to be predictive of child care choices. Differences across sites also exist, particularly for Arkansas, New York, and Yale.

Primary Outcome Measures

Primary outcome measures reflect cognitive development at ages 3, 5, and 8. The Peabody Picture Vocabulary Test Revised (PPVT-R) is used in all 3 years to measure receptive language. IQ is measured with the Stanford-Binet Intelligence Scale in year 3, the Weschler Preschool and Primary Scale of Intelligence (WPPSI) in year 5 (full, plus verbal and performance subscales), and the Weschler Intelligence Scale for Children (WISC) in year 8 (full, plus verbal and performance subscales). Outcome variable missingness in year 3 is just 9 percent for the Stanford-Binet but 16 percent for the PPVT-R. The higher missingness rate for the PPVT-R reflects the fact that approximately 7 percent of the children did not reach basal PPVT scores at age 3, and so were not assessed for that measure. In year 5, missingness rates are 18.5 percent across measures and in year 8 they range from 13.5 to 15 percent.

Table 1 reveals that, among FU children, the outcome variables tend to have the highest values on average for those participating in some center-based child care and the lowest values on average for those who experience only maternal child care. The IHDP children outscore all of the FU subgroups in years 3 and 5 but seem comparable to the center-based FU children in year 8.

ESTIMANDS AND METHODS

A major challenge in policy research concerns the proper way to form subgroups based on, or otherwise "control for," post-treatment variables. This section first describes the estimands in which the authors are interested and then delineates inappropriate and appropriate ways of estimating these. The authors also discuss their approach to the missing data issues in this study.

What Do We Want to Estimate?

Of interest here is the role played by a mediating (post-randomization) variable in the causal link between high-quality child care (i.e. the IHDP) and developmental outcomes, in particular, the differential effects of high-quality child care (i.e. the IHDP) on children who, in the absence of such care, would engage in any of several different types of care. Starting from the favorable vantage point of a randomized experiment, the authors would like to continue to pose and estimate the answers to causal questions.

The focus is on three estimands. Each reflects the effect of high-quality child care on a particular subgroup of children. These subgroups are defined by the care the children either did or would have received in years 2 and 3 in the absence of being assigned to the intervention group:

A. No non-maternal care,
B. Some non-maternal care (>4 hours/week), but no center-based care, and,
C. Some center-based care (>4 hours/week).

Note that these groups are both mutually exclusive and exhaustive of the possibilities.

Simplistically, these classifications might appear to represent subgroups defined by values of the outcome variable "child care choice in years 2 and 3." Closer examination reveals that these subgroups are defined based on potential outcomes (see Rubin, 1978). A potential outcome for individual i and for treatment 1 is the outcome that would have been experienced by that individual had she received treatment 1. If the individual did indeed receive treatment 1, then this potential outcome is observed. If she did not, it will not be observed. The subgroups described above are defined by the potential outcome "child care choice in years 2 and 3" that is manifested only under assignment to control. The advantage of these definitions is that potential outcomes can be considered to be pre-treatment variables because they are not influenced by the actual treatment assigned or received. Therefore, in the context of a randomized experiment, potential outcomes, like other pre-treatment variables, are independent of treatment assignment. Consequently, comparisons across treatment groups within subgroups based on potential outcomes represent causal effects. In contrast, comparisons across treatment groups within subgroups based on standard outcomes do not represent causal effects. Because these child care groups correspond to a somewhat simple example of what Frangakis and

Rubin (2002) call "principal strata" (because they stratify the population into mutually exclusive subgroups based on theoretical pre-treatment variables), these three groups will be referred to henceforth as strata: stratum A for no non-maternal care, stratum B for some home-based care, and stratum C for some center-based care.

Why Might Standard Approaches Be Problematic?

The complication in calculating estimates for these estimands is that in general a valid estimate cannot be obtained for any of these three strata simply by comparing the FU children who experience the type of care in question with all IHDP children. This means that running a regression with dummy variables for these categories and treatment assignment will not produce valid estimates in general (Rosenbaum [1984] and Frangakis and Rubin [2002] both present formal justifications for why these types of estimates are misguided). This is because, as demonstrated in Table 1, families, and even the children themselves, differ across these groups in ways that would be expected to be related to children's development. If the estimate is to address a causal question, treatment effects must be calculated only within a given stratum.

Instrumental variables (IV) techniques have been growing in popularity as a strategy for some types of mediating analyses. While this technique is quite appropriate for some scenarios (see for instance Angrist, Imbens, and Rubin, 1996), it will not be appropriate for many others. In particular, one of the most important assumptions inherent in this approach is that of the exclusion restriction. This assumption requires that the instrument (randomized treatment assignment in this example) affects the primary outcomes of interest only through the specified mediating variable (or variables if multiple instruments are available). Current policy examples that highlight these issues in the context of randomized anti-poverty experiments can be found in Bos and Granger (2002) and Magnuson, Gibson, and Huston (1999). The strategy presented here does not impose such a requirement and, in addition, is much more explicit regarding for what groups of people inferences are being made for each estimand—this nuance is sometimes downplayed or overlooked in IV analyses.

Approach

One benefit of randomization is that it ensures the existence of a group of children who received the IHDP intervention who are similar on average to the FU children in each stratum, both in terms of observed and unobserved characteristics. That is, the IHDP children in each stratum actually exist (this is not necessarily true in an observational study). The challenge, then, is to find the "matched" children in the IHDP for each of the three FU strata, that is, the children who in the absence of the intervention would have participated in the same kind of child care.

Propensity score matching (Rosenbaum and Rubin, 1983) is a technique that was developed as a means of reducing selection bias when estimating causal effects in the context of observational studies. It attempts to accomplish this goal by estimating the probability of treatment given observed covariates, that is, the propensity score, generally using either logistic or probit regression; and then finding for each treatment group member a "matched" control group member that has the closest propensity score (or an appropriate transformation of that score). The process should create two matched groups that have similar values on average for each of

the background variables included in the estimation. If these variables are the only confounding variables, these groups can be conceptualized as arising from a randomized experiment. Propensity score matching has been gaining in popularity over the years[3] and has recently been advocated as an approach to causal inference in studies of child care and early childhood intervention by an expert panel of the National Academy of Sciences (Shonkoff and Phillips, 2000).

The approach here is distinct from, but similar to, classic propensity score matching. In this study, to calculate the treatment effect for stratum A—no nonmaternal care—assume selection into this stratum can be predicted using observed characteristics X. Then calculate, for each person, her predicted probability of belonging to this stratum given observed pre-treatment variables, that is Pr(P=A | X) where P is a random variable for stratum. Playing off both the principal strata and the propensity score terminology, these predicted probabilities will be called "principal scores." Pr(P=A|X) can be modeled using any number of techniques for categorical outcomes. Recall that P is unobserved for the members of the IHDP group. The randomization ensures however that Pr(P | X, T=1) = Pr(P | X, T=0), where T denotes treatment assignment. So a model can be built and principal scores can be calculated using data from the FU group (T=0). Then using the coefficients from this model along with data from the IHDP group (T=1), principal scores for the IHDP group are calculated. These principal scores allow the authors to find, for each FU member in a given stratum, a similar IHDP member in the same stratum because they serve as one-number summaries of all the covariates used to estimate them. This approach is different from standard propensity score estimation because the model is estimated based on only a subset of the individuals—those not assigned to the intervention.

Following similar logic to that used to justify propensity score matching (Rosenbaum and Rubin, 1983), matching on these principal scores should create "true" comparison groups among the IHDP children. That is, comparisons of average outcomes across these matched groups should yield unbiased treatment effect estimates for each stratum. Intuitively, matching on, for example, stratum A principal scores creates a comparison group of IHDP children who would have engaged only in maternal care had they not had access to the IHDP intervention. The authors can assess the success of the matching by comparing the means of background variables across matched groups. The more similar these are, the more confident one can be that similar groups have been created, that is, groups that would have engaged in the same type of behavior in the absence of the intervention. Thus balance in background variables across matched groups will be used as a diagnostic for the success of the method.

In practice, the probability of belonging to a given child-care group is being modeled separately for each stratum, in each case as compared with the other two strata. Therefore a logistic regression can be used to obtain the estimated coefficients

[3]Examples of propensity score matching applications can be found in medicine, epidemiology, psychiatry, and economics (Fiebach, 1990; Heckman, Ichimura, and Todd, 1997; Imbens, Rubin, and Sacerdote; Lavori, Keller, and Endicott, 1995, Lechner, 1999). Dehejia and Wahba (2000) present an example of an evaluation context within which propensity score matching can be shown to closely replicate the true answers. However, there has been controversy in recent years about the potential fragility of propensity score estimates and lack of applicability to all evaluation contexts (Heckman, Ichimura, and Todd, 1997; Smith and Todd, 2000; Wilde and Hollister, 2002). The authors agree that inferences from propensity score analyses, as with all types of analyses, should be interpreted only with the underlying assumptions firmly in mind.

in each model.[4] When matching, children who never attended any days at a IHDP child development center and received no "treatment" as defined herein, are excluded from the IHDP group; as long as appropriate matches can still be found, this should not cause problems. Then, for each FU child, the IHDP child who has the closest principal score is chosen as a match, even if that child has already been chosen as a match for another FU child; this is called matching with replacement (see Dehejia and Wahba, 1999 for a justification and comparison to other matching methods). This matching was performed within sites because of research that points to the important role geographic location can play in representing unmeasured confounders (Hotz, Imbens, and Mortimer, 2000).

Assumptions

The crucial role of the assumptions of this method must be emphasized, however. The first assumption is that all confounding covariates have been observed. This is referred to as ignorability of the assignment mechanism (Rubin, 1978) in the statistics literature and selection on unobservables (Heckman and Robb, 1985) in the economics literature. If an important pre-treatment variable was not included, and that variable affects both choice of non-intervention child care and subsequent cognitive outcomes (e.g., "inherent parenting skills"), then the observed covariates could be balanced and the method could still yield biased treatment effect estimates. Therefore results must always be interpreted with the thought of omitted variables in mind. That said, due to the randomization involved, to the extent that dependencies are present between observed and omitted variables, balancing the former should help to balance the latter. Moreover, an important paper by Heckman, Ichimura, and Todd (1997) demonstrates for an evaluation of a large-scale job training program (JTPA) that bias due to selection on unobservables in that example is small relative to bias caused by other sources (selection on observables, differences in measures across comparison groups, etc.).

The second assumption is that the covariate distributions across the treatment groups overlap sufficiently; in other words, it must be possible to find reasonable matches for each FU child for a given stratum. This second assumption should be trivially satisfied due to the randomization.

Linear regression relies on the first assumption. It does not, strictly speaking, rely on the second. However, to the extent that there is insufficient overlap, regression approaches are forced to extrapolate the model over portions of the covariate space where there is no information, which can be quite dangerous. Regression also relies heavily on parametric assumptions about the relationship between the outcomes and the pre-treatment variables that principal score matching (and propensity score matching) can avoid if desired.

Missing Data

A further complication in this study, one that is common to studies with human subjects, is missing data. While missingness rates are fairly low for studies of this

[4]Alternative choices include probit regression, discriminant analysis, or even use of polytomous regression on all three categories at once. Sensitivity of inferences to these types of modeling choice is a subject for further methodological exploration. Diagnostics (e.g., checks of pre-treatment balance across matched groups) help to ensure that model choice is not critical for obtaining valid inferences; the model is just a means to an end for creating well-balanced groups.

sort, use of common approaches to missing data, such as complete case analyses, can lead to biased estimation (Little and Rubin, 1987) even when the missingness occurs in relatively small proportions. Moreover, complete case analyses on all the variables used would limit the sample size to 243 observations, which is prohibitively small for the analyses performed. If separate complete case analyses were performed for each outcome, sample sizes would range from 274 to 307; however this strategy hinders comparisons across analyses, which is particularly problematic when investigating trends over time.

Missing data are addressed by using multiple imputation (MI) (Rubin, 1987; Schafer, 1997). In particular, the authors use software developed by Joseph Schafer (1997) that relies upon the General Location Model (Olkin and Tate 1961), a fairly flexible model that accommodates both categorical and continuous data. This model is used to predict missing values for each person based on observed values for other variables. Use of multiple predictions, or imputations, for each missing value helps to properly account for our uncertainty about these imputed values.[5] Moreover, MI allows for retention of the original sample size. Although MI does rely on assumptions about the missing data process, these assumptions are more plausible than are those required for other standard approaches (such as complete-case analyses). For example, for valid inference a complete-case analysis requires that the complete-case sample is a random sample of the original sample; this means that one must assume that any two people have the same chance of having missing values. Multiple imputation, on the other hand, requires only that two people who have the exact same values on all observed variables have the same chance of having missing values. Therefore, complete-case analyses are generally more likely to yield biased results relative to MI analyses.

A further important point is that the validity of the primary analyses relies, in part, on the randomization in the study. Complete case analyses, however, generally destroy the initial randomization by removing members from the sample differentially from each treatment group. As an example, even if adequate sample sizes were available to perform complete case analyses, there would no longer have been any guarantee that true matches existed within the IHDP group for each member of the FU group in any given strata. Multiple imputation allows for retention of all the benefits of the initial randomization.

RESULTS—THE DIFFERENTIAL EFFECTS OF HIGH-QUALITY CHILD CARE

This section presents treatment effect estimates for each principal stratum. The coefficients from the logistic regressions used to estimate principal scores, as well as goodness-of-fit measures, are presented in Appendix A. The principal scores were estimated using all of the background variables as predictors, except cigarette, alcohol, and drug use while pregnant. These last three measures were deemed too unreliable for this purpose, but have been included in the balance diagnostics. For the treatment effects for each stratum, three estimates are presented for each outcome. The first two sets of estimates are based on the comparison between the rel-

[5]Multiple imputation is becoming accepted practice for dealing with missing values in survey data. This is evidenced by its use in the following large-scale surveys: National Health and Nutrition Evaluation Survey (NHANES), sponsored by the National Center for Health Statistics; Fatal Accident Reporting System (FARS), sponsored by the Department of Transportation for the National Highway Traffic Safety Administration; and Survey of Consumer Finances (SCF), sponsored by the Federal Reserve Bank; among others.

Table 2. Balance across treatment groups for both matched and unmatched samples corresponding to each of the three principal strata.

	A: No non-maternal child care		B: Home-based child care		C: Center–based child care	
	N=164	N=76	N=227	N=168	N=188	N=109
Background Variables	Unmatched	Matched	Unmatched	Matched	Unmatched	Matched
Child						
birth weight	0.35	–0.09	–0.84	0.01	0.60	–0.18
head circumference at birth	–1.28	–0.51	–0.61	–0.73	–0.25	0.21
sex of child	–0.10	–0.50	0.77	0.20	0.52	1.25
weeks pre-term	0.96	0.29	0.67	–0.55	0.34	–0.14
birth order	–1.06	–0.30	–0.53	–1.09	0.55	0.05
neonatal health index	–1.47	–0.51	0.94	0.46	0.95	0.72
child a twin	–0.04	–0.53	0.42	–0.03	0.88	0.67
Mother						
mother's age	0.05	0.08	–1.14	–0.23	–2.41	–0.31
mother black	–0.28	–0.55	–0.18	–0.27	–0.04	0.08
mother Hispanic	–0.07	–0.15	–1.46	–0.61	0.85	–0.50
mother white	0.32	0.69	1.18	0.66	–0.41	0.14
mother married at birth	–0.13	–0.17	–0.10	–0.46	–2.77	0.13
less than high school	0.71	0.32	1.69	0.55	2.47	0.93
high school	0.75	0.20	0.30	0.45	1.22	0.36
some college	–1.95	–0.22	–1.47	–0.73	–2.35	–1.91
finished college	0.49	–0.76	–1.29	–0.48	–2.03	0.06
cigarettes during pregnancy	0.44	0.33	0.92	0.28	1.74	0.51
alcohol during pregnancy	1.46	0.76	0.27	–0.11	–0.67	–0.55
drugs during pregnancy	–2.50	–1.06	–0.07	0.32	–1.40	–0.23
worked during pregnancy	1.04	0.42	–0.63	0.17	–1.48	–0.05
prenatal care	–1.31	–0.41	0.67	1.41	1.16	1.10
Father						
father black	0.20	–0.21	0.03	–0.13	0.17	0.12
father Hispanic	–0.04	–0.31	–0.65	–0.08	1.63	–0.01
father white	–0.17	0.45	0.45	0.18	–1.10	–0.13
Site						
Arkansas	0.45	0.00	1.29	0.00	–1.04	0.00
New York	0.04	0.00	–1.45	0.00	1.00	0.00
Harvard	–1.34	0.00	–0.96	0.00	–0.13	0.00
Miami	0.01	0.00	–0.08	0.00	0.85	0.00
Pennsylvania	0.74	0.00	2.28	0.00	–0.07	0.00
Texas	0.67	0.00	–0.39	0.00	–0.10	0.00
Washington	–0.61	0.00	–0.31	0.00	–1.49	0.00
Yale	0.31	0.00	–0.03	0.00	1.93	0.00
Mean of t-scores	–0.11	–0.09	–0.01	–0.04	–0.03	0.07
Standard deviation of t-scores	0.91	0.41	0.94	0.47	1.34	0.55

Columns contain t-statistics of the difference in means for each covariate across treatment groups. These statistics are summarized in the bottom two rows by their mean and standard deviation.

evant FU group and the entire IHDP group (unmatched). For this comparison both unadjusted differences in means and regression estimates are presented. These regressions use the same variables that were used to estimate the principal strata. The unmatched regression estimate is provided to gauge the extent to which regression alleviates some of the selection bias incurred by comparing potentially non-comparable groups. The third treatment effect estimate is based on a comparison between the relevant FU group and a comparison group created by matching on principal scores (matched). Only the regression estimates are presented for this comparison. (For the matched groups, the differences between unadjusted difference in means and regression estimates are generally much smaller than these differences when using unmatched groups.)

How the Matching Affects Balance across Comparison Groups

The gains in balance achieved are investigated by calculating t-statistics (Table 2) for the differences in means of background characteristics across both unmatched and matched groups corresponding to the three primary estimands. The smaller the t-statistic for a given variable, the better the balance across groups. Assuming that the characteristics examined here (or functions thereof) represent all confounding variables, balance in these variables across comparison groups should give confidence that the corresponding treatment effect estimates are unbiased. In a randomized experiment with sufficient sample size such t-statistics would be expected to roughly follow a normal distribution with a mean of 0 and a standard deviation of 1. Thus, means and standard deviations are also calculated for the t-statistics for each analysis using the randomized experiment expectations as a benchmark. Means close to 0 are a good sign for balance; standard deviations of less than 1 indicate the potential for greater efficiency than a randomized experiment if the assumptions (no unmeasured confounders) are met.

Comparing all children who received the IHDP intervention (thus had access to the CDCs) with the stratum A children in the control group who received no non-maternal care in years 2 and 3, indicates few worrisome differences in background characteristics. In particular, only the t-statistic for whether the mother used drugs during pregnancy achieves statistical significance (0.05 significance level), although the t-statistic for whether the mother attended college courses is also quite high. When the comparison group is limited to IHDP intervention recipients matched to follow-up children who received no non-maternal care (implicitly representing those who would have received no non-maternal care in the absence of the intervention), however, the balance in background characteristics across groups is much closer. Standard deviations across these t-statistics fall from 0.90 for the unmatched groups to 0.41 for the matched groups. The means for the t-statistics are close to 0 for both sets of groups.

If all children who received the IHDP intervention are compared with the stratum B children in the control group (home-based non-maternal care, but no center-based care from ages 1 to 3) (Table 2, columns 3 and 4), the imbalance is not particularly striking. The one significant difference is for the percentage of children from the Pennsylvania site. However, when the comparison group is limited to IHDP intervention recipients matched to the targeted follow-up children (some non-maternal care, no center-based care), the balance in background characteristics across groups is noticeably closer. Standard deviations across these t-statistics fall from 0.94 for the unmatched groups to 0.47 for the matched groups. Again the means of the t-statistics are both very close to 0.

Table 3. Treatment effect estimates for stratum A: no non-maternal care.

Comparison group:	Unmatched				Matched	
Sample size:	N=164				N=76	
Analysis:	Diff. Means		Regression		Regression	
Outcome measures	t.e.	s.e	t.e.	s.e	t.e.	s.e.
Age 3 years						
PPVT-R	**11.8**	3.4	**11.2**	2.5	**13.3**	3.8
S-B	**16.4**	3.4	**16.8**	3.1	**20.2**	4.4
Age 5 years						
PPVT-R	**10.0**	4.9	**10.1**	3.7	**12.3**	6.1
WPPSI						
Full-scale	4.6	3.1	**6.2**	2.6	**9.1**	4.0
Verbal	4.6	3.4	**5.9**	2.7	7.4	4.7
Performance	4.1	2.8	**5.7**	2.7	**9.6**	3.9
Age 8 years						
PPVT-R	6.4	4.1	**7.8**	3.3	9.2	5.3
WISC						
Full-scale	3.2	3.0	4.7	2.7	**8.2**	3.8
Verbal	3.4	3.2	4.5	2.9	8.1	5.0
Performance	2.8	2.8	4.4	2.7	**7.4**	3.7

Treatment effect estimates significant at a 0.05 level (two-tailed) are presented in boldface.

Comparisons between background characteristics for all the IHDP intervention participants and the stratum C children from the control group (at least some center-based care in years 2 and 3) are displayed in columns 5 and 6. They reveal statistically significant differences across groups with respect to mother's age, mother's marital status at the time the child was born, and three of the educational attainment variables (less than high school, some college, and finished college). No t-statistic for the differences in means across matched groups are statistically significant. Standard deviations across these t-statistics are 1.34 for the unmatched groups and 0.55 for the matched groups and means are both close to 0. Thus matching yields the largest absolute gains in overall balance (as measured by the standard deviation of the t-statistics) for stratum C comparisons. Note that for the stratum C comparisons, however, even after matching the t-statistic for the mother having attended some college is -1.91. This value (along with those for the other education variables) indicates that the mothers for the matched group of intervention children had slightly lower educational attainment than the mothers of children who actually attended standard center-based care in years 1 through 3. This implies that the treatment effect estimates for this stratum may be somewhat understated.

No Non-Maternal Child Care

The results corresponding to the unmatched and matched analyses for stratum A (Table 3) are somewhat different across all years. In particular there is a consistent pattern, for all but one measure, in which the regression estimates for the unmatched groups fall somewhere in between the unadjusted difference in means for the unmatched groups and the regression estimates for the matched groups. If the regression estimates from the matched groups are treated as the most reliable estimates on average of the three (due to superior balance across these groups), in

Table 4. Treatment effect estimates for stratum B: some non-maternal, no center-based care.

Comparison group:	Unmatched				Matched	
Sample size:	N=228				N=168	
Analysis:	Diff. Means		Regression		Regression	
Outcome measures	t.e.	s.e	t.e.	s.e	t.e.	s.e.
Age 3 years						
PPVT-R	**11.0**	2.4	**11.4**	1.9	**12.3**	2.7
S-B	**14.4**	2.6	**15.0**	2.0	**16.4**	3.0
Age 5 years						
PPVT-R	5.9	3.4	**7.1**	2.9	**8.7**	4.1
WPPSI						
Full-scale	1.7	2.3	**4.0**	1.9	**5.9**	2.5
Verbal	2.1	2.3	**4.0**	2.0	**5.9**	2.6
Performance	0.9	2.2	3.2	1.9	**4.9**	2.5
Age 8 years						
PPVT-R	0.8	3.5	2.5	2.9	5.6	4.3
WISC						
Full-scale	2.4	2.3	**4.8**	2.0	**7.2**	2.7
Verbal	2.1	2.4	**4.5**	2.0	**7.4**	2.6
Performance	2.2	2.3	**4.2**	2.0	5.7	3.2

Treatment effect estimates significant at a 0.05 level (two-tailed) are presented in boldface.

many cases the regression estimates from the unmatched groups seem to eliminate some of the selection bias caused by comparing unbalanced groups but do not eliminate quite as much as the matched principal score analyses.

The results for the more closely balanced matched analyses demonstrate very large and significant positive effects (0.05 significance level) of attending high-quality child care for outcomes in year 3, as well as large effects in years 5 and 8, five of which are statistically significant. These results are consistent with past literature that identifies a positive developmental effect of non-maternal child care versus maternal care experienced between ages one and three years that may persist until ages seven or eight (Blau and Grossberg, 1992; Han, Waldfogel, and Brooks-Gunn, 2001; Waldfogel, Han, and Brooks-Gunn, 2001). The striking difference, compared with previous findings, is that these treatment effects are so much larger. It seems likely that this difference in magnitude may be attributable to the more substantial difference in quality of child care experienced across comparisons groups in this study as compared to the heterogeneity observed in typical, existing arrangements.

Some Non-Maternal, No Center-Based Child Care (Home-Based)

Table 4 displays results for the comparisons in stratum B. Here the estimates for the effect of high-quality care are fairly similar across analyses for the year 3 results, but diverge more as time goes by. For all these measures the estimates are ordered such that the unmatched regression estimates lie between the unmatched difference in means and the matched regression estimates. The results for the most closely matched groups reveal very large and statistically significant effects for year 3. Year 5 and 8 effects are still large and statistically significant for five of the eight measures. This implies that the quality of IHDP care was superior to the home-based child care available to people in this stratum.

Table 5. Treatment effect estimates for stratum C: some center-based care.

Comparison group:			Unmatched		Matched	
Sample size:			N=188		N=109	
Analysis:	Diff. Means		Regression		Regression	
Outcome measures	t.e.	s.e	t.e.	s.e	t.e.	s.e.
Age 3 years						
PPVT-R	5.0	2.8	**8.5**	2.3	**10.3**	2.9
S-B	**6.3**	2.8	**10.4**	2.5	**11.6**	3.3
Age 5 years						
PPVT-R	–0.7	3.5	3.8	3.1	6.4	4.1
WPPSI						
Full-scale	–1.9	2.7	2.2	2.5	2.0	3.5
Verbal	–0.2	2.6	3.5	2.5	4.1	3.2
Performance	–3.1	2.6	0.6	2.6	–0.3	4.1
Age 8 years						
PPVT-R	2.9	3.4	**8.2**	2.9	8.7	4.8
WISC						
Full-scale	–1.5	2.8	2.1	2.4	1.7	3.7
Verbal	–1.2	2.7	2.5	2.5	2.9	3.8
Performance	–1.3	2.8	1.5	2.6	0.4	4.0

Treatment effect estimates significant at a 0.05 level (two-tailed) are presented in boldface.

At Least Some Center-Based Child Care

The greater imbalance for the unmatched analyses, as compared with the matched analyses in stratum C, does not lead to greater differences in treatment effect estimates between the unmatched difference in means estimates and the matched regression estimates even at age 3. Indeed the difference between these estimates across all years is generally smaller than observed for the previous analyses. This might imply that the linear model used is more appropriate for the members of this stratum or it could simply be a reflection of the lack of treatment effect in years 5 and 8. The unmatched regression estimates fall between the other two for all measures except one (WISC performance, age 8).

The estimates of the effect of high-quality child care relative to the care received by those who participated in at least some center-based care are quite large and statistically significant in year 3. These effects appear to have attenuated by year 5 at which point no statistically significant results are evident. By age 8 the results are still positive but are moderate and not statistically significant. Interestingly, the PPVT-R measure in year 8 displays a very large and nearly statistically significant effect, but this is difficult to interpret (it would be expected roughly to mirror results for the WISC verbal measure for that year).

Combining Strata A and B–No Center-Based Care

Given that the patterns in point estimates for treatment effects between stratum A and stratum B are so similar, it seems reasonable to collapse these strata into one category—no center-based care—as a means of increasing the precision of these estimates. The combined analysis yields treatment effect estimates (Table 6) of similar magnitude as the separate analyses, however standard errors are reduced sufficiently so that all the estimates for the matched comparison are now statistically

Table 6. Treatment effect estimates for strata A and B combined: no center-based care.

Comparison group:	Unmatched				Matched	
Sample size:	N=272				N=224	
Analysis:	Diff. Means		Regression		Regression	
Outcome measures	t.e.	s.e	t.e.	s.e	t.e.	s.e.
Age 3 years						
PPVT-R	**11.2**	2.1	**11.0**	1.8	**11.7**	2.1
S-B	**14.9**	2.3	**15.4**	2.0	**17.4**	2.2
Age 5 years						
PPVT-R	**7.1**	3.0	**7.9**	2.6	**11.7**	3.2
WPPSI						
Full-scale	2.6	2.1	**4.5**	1.7	**6.6**	1.9
Verbal	2.9	2.1	**4.4**	1.8	**6.4**	2.1
Performance	1.8	2.0	**3.9**	1.8	**5.6**	1.9
Age 8 years						
PPVT-R	2.4	3.0	4.2	2.5	**7.6**	3.5
WISC						
Full-scale	2.6	2.1	**4.6**	1.8	**6.6**	2.5
Verbal	2.5	2.1	**4.4**	1.8	**5.8**	2.6
Performance	2.4	2.1	**4.2**	1.8	**6.3**	2.2

Treatment effect estimates significant at a 0.05 level (two-tailed) are presented in boldface.

significant. These results strengthen the premise that children who do not receive any center-based care in years 2 and 3 could demonstrate persistent, significant, positive effects on cognitive development through to age 8 if given the chance to participate during these years in high-quality child care such as that provided by the IHDP. These results are consistent with evidence from observational studies (NICHD Early Child Care Research Network, 2000) that suggests that children are better prepared for school if they have participated in center-based care prior to starting school.

MORE MEDIATING VARIABLES

To investigate more fully the causal pathways by which some of these processes might have occurred, the authors examine the differential effect of the intervention on two additional mediating variables: an indicator for participation in center-based child care (whether as primary or secondary type of care) in months 36 to 48, and an indicator for the mother having had at least one additional birth through month 36. Since these variables are both post-treatment, in the absence of performing specific analyses such as those above, and for the reasons described above, causal statements cannot be made about the relationships between these variables and the primary outcomes.[6] It is possible, however, to make causal statements about the relationships between treatment assignment and these outcomes within

[6] Sample size restrictions prevented us from performing "mediating analyses" for these variables within the mediating analyses already performed for our primary outcomes with regard to child care choices in years two and three. If these restrictions did not exist, one approach would be to compare, for example, IHDP children in each stratum who moved into center-based care in year 4 to matching FU children (the FU children who would have experienced center-based care in year 4 had they first been assigned to receive the IHDP intervention). The authors could then examine if the long-term treatment effects for this subgroup are truly small to nonexistent.

Table 7. Mediating variables, means and standard errors.

	A: No non-maternal child care		B: Home-based child care		C: Center-based child care	
IHDP	Mean	S.E.	Mean	S.E.	Mean	S.E.
center-based care 36–48 mos.	0.70	0.06	0.58	0.07	0.60	0.09
additional birth pre-36 mos.	0.51	0.14	0.43	0.06	0.45	0.07
Follow-up						
center-based care 36–48 mos.	0.20	0.02	0.29	0.03	0.71	0.03
additional birth pre-36 mos.	0.61	0.03	0.32	0.02	0.29	0.04

strata, and thus conjecture about the subsequent relationships between these mediating outcomes and later outcomes that could inform future research.

The (unadjusted) means and associated standard errors displayed in the top panel of Table 7 correspond to the IHDP children who were matched to the FU children (bottom panel) in each of child care strata defined earlier. The mean values of the mediating variables are much more similar among the IHDP children across strata than they are for the FU children. Of particular interest, for strata A and B, a much higher percentage of IHDP children experience center-based care (70 and 58, respectively) than did the FU children (20 and 29, respectively). In stratum C, however, fewer IHDP children engaged in center-based care during this post-intervention period as compared with the FU group. Recall that FU stratum C contains children who had already been participating in the center-based care existing in their community. Thus participating in center-based care in the fourth year (post-intervention) would have represented a potentially seamless transition (i.e. they might not have had to switch providers). On the contrary, IHDP families would have had to find and move their children to new, appropriate center-based care. Difficulties in effecting this transition could be responsible for the lower rates of center-based care exhibited by the IHDP stratum C children.

If the authors can assume that center-based care provides better cognitive stimulation, on average, as compared to maternal care alone or home-based care, these findings suggest a reason for the differential attenuation patterns demonstrated in our cognitive outcomes across groups. All three groups exhibited large treatment effects at the end of the intervention (age 3 years). Post-intervention, however, the gap between the types of care received by IHDP and FU children differed across principal strata. The intervention seems to have induced a greater proclivity towards center-based care among the IHDP families in strata A and B (those who would not have chosen center-based care in the absence of the intervention). Therefore the gap between the percentage of IHDP children engaging in center-based care in strata A and B and the number of corresponding FU children engaging in center-based care, may have helped to maintain the differences in cognitive outcomes between treatment groups. The stratum C children, however, ended up in quite similar types of care across treatment groups post-intervention, with actually fewer intervention children than FU children able to engage in center-based care in their fourth year. This could have led to the attenuation of the initial gains for these children over time.

Additional births might also mediate the causal link between high-quality child care and developmental outcomes. In stratum A, the IHDP families have a lower rate, as compared to FU families, of having an additional birth in the 3 years following the birth of the study child. Additional siblings have been hypothesized to

Table 8. Effect of switching to available forms of center-based care: within control-group comparisons.

Comparison	Maternal →Center		Home-based →Center	
Estimate	matched: difference in means	matched: regression-adjusted	matched: difference in means	matched: regression-adjusted
Sample size	n=66		n=157	
Outcomes				
Age 3 years				
PPVT-R	8.7	4.1	4.7	2.2
	(6.21)	(5.71)	(5.61)	(2.65)
S-B	9.3	5.6	5.9	2.1
	(6.23)	(4.85)	(5.79)	(2.44)
Age 5 years				
PPVT-R	**15.8**	8.3	5.6	2.2
	(7.41)	(6.76)	(5.12)	(3.18)
WPPSI				
Full Scale	6.8	2.6	3.3	0.5
	(5.17)	(4.06)	(4.92)	(2.55)
Verbal	6.6	2.3	2.3	–0.3
	(5.95)	(5.05)	(5.09)	(2.62)
Performance	5.6	2.2	3.4	1.1
	(4.79)	(4.39)	(4.45)	(3.02)
Age 8 years				
PPVT-R	2.5	–1.85	–4.6	**–7.5**
	(7.12)	(5.16)	(6.94)	**(3.80)**
WPPSI				
Full Scale	5.3	1.2	3.9	1.4
	(5.60)	(4.18)	(5.87)	(2.73)
Verbal	4.3	0.3	2.8	0.0
	(6.00)	(4.91)	(5.73)	(2.86)
Performance	5.6	2.0	4.1	4.1
	(6.01)	(4.36)	(5.32)	(5.32)
Balance t-statistics				
mean	0.02		0.07	
standard deviation	0.57		0.46	

Results are treatment effect (t.e.) estimates and associated standard errors (s.e.) obtained within matched samples both from difference in mean estimates as well as from linear regression on all pre-treatment variables (except cigarette, alcohol and drug use). The mean and standard deviation of the difference-in-mean t-statistics for the pre-treatment variables are also provided as an indication of balance achieved by the matching. Treatment effect estimates significant at a 0.o5 level (two-tailed) are presented in bold.

have a negative effect on children's development (Baydar et al., 1997). Therefore it seems plausible that for children in this stratum, part of the positive treatment effect may be caused by a reduction in number of siblings added to the family shortly after the birth of the participating child. Note that for strata B and C this phenomenon works in the opposite direction.

COMPARISONS ACROSS STRATA

We should be explicit about the fact that the results presented in sections 4 and 5 do not allow us to make causal inferences across strata. For instance, it would not

be appropriate to infer from these results that existing center-based care is necessarily a better option (on average) than home-based or maternal care, just because the treatment effects for stratum C are much smaller than those for strata A and B. That is because different types of families select into different types of child care arrangements. The point of this work is that the authors can make causal inferences within each stratum. From a policy viewpoint the authors have addressed the question of what the differential effect would be on different types of people if high-quality child care were made available to them. These results then could motivate provision of high-quality center-based options to people currently planning on using some form of home-based or maternal child care.

It is also possible to perform analyses that address the question of whether the FU children who received maternal care or home-based care would have had better outcomes had they instead participated in the types of center-based care in which the stratum C FU children did. To do this standard propensity score analyses (discussed below) were performed to find matches to the stratum A and stratum B FU children from among the stratum C FU children. These analyses again require inclusion of all confounding covariates. They also require that true matches for each child exist. This second assumption is not as trivially satisfied as it was in the principal score analyses above because no randomization was performed across the groups now being compared. In addition, the randomization ensured for the principal score analyses that the joint distributions of the pre-treatment variables (hence dependence structures, correlation patterns) were the same across treatment groups. Because of this symmetry, balancing observed covariates should have helped to balance unobserved covariates because the same types of dependence between the variables will exist across comparison groups. This property does not hold for the standard propensity score analyses used here. These analyses are still likely to be more robust than standard regression analyses, however.

The results displayed in Table 8 reflect estimates of the following quantities: the effect of participating in standard (existing) center-based care in years 2 and 3 for FU children who instead experienced only maternal care; and the effect of participating in standard (existing) center-based care in years 2 and 3 for FU children who instead participated in home-based care. For this table the matched difference-in-means results and the matched regression results are presented rather than the unmatched and matched regression results because there were greater discrepancies between the former pairs of estimates overall; therefore, the estimates displayed provide a fuller sense of the sensitivity of our conclusions to analytic strategy.

These analyses yield very few significant estimates. While all but one of the estimates attributable to the move from maternal care to center-based care are positive (implying a beneficial effect of center-based care), only one is statistically significant (the PPVT-R in year 5 unadjusted matched estimate). The estimates of the effect of moving to center-based care from home-based care are generally smaller than for the maternal-care versus center-based care comparison (often less than half the magnitude), but also almost all positive (only two negative, both for PPVT-R in year 8, and one zero). The only statistically significant result is the negative effect estimated for the PPVT-R in year 8 by the regression analysis on the matched samples.

As discussed above these results need to be interpreted with greater caution than the principal score analyses used for the primary estimands. In addition, they are very noisy, making it difficult to draw solid inferences anyway. Overall there is some descriptive evidence in the form of the overwhelming number of positive effect estimates indicating that there may be some advantages of center-based care over home-

based and, more probably, maternal care, however the evidence is in no way strong or conclusive. Based on these analyses, however, the possibility cannot be ruled out that the stronger effects of high-quality child care estimated for those who would otherwise experience maternal or home-based care as opposed to those who would otherwise experience standard center-based care are due to differences in the types of children and families self-selecting into each stratum (rather than to the differences in the quality of care attributable to each of these existing child care options).

DISCUSSION

This paper makes use of a new methodological strategy that allows for investigation of the role of post-randomization variables. This approach can be used to differentiate between the effect of high-quality child care on children who otherwise would have been cared for by their mothers, the effect on children who otherwise would have participated in home-based child care, and the effect on children who otherwise would have engaged in some center-based child care. It is clear that the children in the first two strata would have had much stronger cognitive outcomes if they had had access to high-quality child care (at the level provided by the IHDP intervention), with effects persisting through to age eight years. The mediating variable comparisons suggest that these effects may have been strengthened by the fact that IHDP children were more likely to participate in center-based care in the year after the intervention ended. Although it is possible that the analysis failed to control for some variable and have thus biased the estimates upward, this variable would have to be incredibly powerful (and unrelated to the ones already controlled for) to eliminate the large positive effects observed.

The differences in outcomes for the children who would have participated in some center-based care in the absence of the intervention are far less striking; with the exception of the age eight PPVT-R, the treatment effect estimates past age three are generally much smaller and none are statistically significant. Part of the reason for this attenuation may be either the failure of all the IHDP children to remain in center-based care in the year after the intervention ended or the greater likelihood of IHDP mothers to have an additional child within 3 years after the birth of the study child. It also may be that the center-based care available to control group families in the study sites (study sites were all university communities) was of higher quality than run-of-the-mill center-based care that existed in other parts of the country at the time of the study; thus the gains for IHDP care relative to the center-based care used by families in this study may understate the true potential gains nationally. A fourth possibility is that these children are different enough from the children in strata A and B that they would have performed well in any of the child care choices available. This last possibility cannot be ruled out, but is thrown into question, by the within-control-group propensity score analyses presented in Table 8.

A major policy implication of these findings is that the provision of universal, high-quality, center-based child care seems likely to be beneficial to all types of participating children, even those who otherwise would be cared for solely by their mothers. A related implication is that if choices need to be made regarding who would most benefit from a potentially limited supply of high-quality child care, targeting those who would otherwise engage in maternal or home-based care might be the most efficient solution. Whether or not this is the most equitable solution is a separate issue.

These results illustrate a new approach toward examination of the causal role played by a mediating variable (in this case, child care in years two and three post-randomization). The authors have outlined proper estimands to retain a causal inter-

pretation that rely on defining subgroups based on what an individual would do both in the presence of and the absence of a given treatment intervention. The authors have also presented a straightforward strategy, principal score matching, for estimating these causal effects. These types of analyses help the researcher to move beyond the "black-box" of a randomized experiment to explore pathways by which a treatment has its effect. Therefore this strategy could be extremely useful in answering important policy questions in the context of social experiments of anti-poverty or welfare programs, health insurance plans, housing or school vouchers, early intervention programs or any number of additional programs or interventions.

APPENDIX A

Table A displays the estimated coefficients from the logit models used to estimate the principal scores for our four primary analyses. Notice that the estimates from the analyses for strata A and B combined only apply to a subset of the coefficients as step-wise regression was used for these propensity scores in an attempt to achieve better balance. The table also displays measures of goodness-of-fit of each model in the form of classification rates.

The goodness-of-fit measure should be viewed with some caution as its implications for our final treatment effect analyses are not entirely clear. For instance a low number (close to 0.5) could either be an indication that there is little to no selection bias (randomization, roughly speaking, would produce values closer to 0.5, as sample size increases) or an indication that, in spite of selection bias, the variables included cannot account for the differences across groups. Understanding how the groups differ in terms of the characteristics measured can help to clarify the issue somewhat, but only to the extent that knowledge exists about the relationships between these variables and the outcomes of interest. Goodness-of-fit measures, therefore, cannot act as a test of the assumptions.

Table A. Estimated coefficients and goodness of fit from logistic regression models used to predict principal scores in primary analyses.

	Stratum A		Stratum B		Stratum C		Strata A/B	
Variables	coef	s.e.	coef	s.e.	coef	s.e.	coef	s.e.
intercept	–1.00	(3.81)	–1.39	(2.92)	0.14	(3.11)	0.10	(0.92)
New York	0.03	(0.51)	0.69	(0.39)	–0.87	(0.42)	0.81	(0.39)
Harvard	0.52	(0.42)	0.25	(0.34)	–0.64	(0.35)	0.61	(0.34)
Miami	0.16	(0.49)	0.10	(0.41)	–0.24	(0.42)	0.23	(0.40)
Pennsylvania	0.41	(0.49)	–0.41	(0.43)	0.05	(0.41)	–0.24	(0.39)
Texas	0.33	(0.43)	0.38	(0.36)	–0.71	(0.40)	0.67	(0.38)
Washington	–0.18	(0.46)	0.10	(0.36)	0.03	(0.37)	0.07	(0.35)
Yale	0.93	(0.44)	0.42	(0.37)	–1.40	(0.44)	1.39	(0.44)
birth-weight group	2.61	(5.42)	–3.10	(4.19)	1.07	(4.56)	–0.43	(2.88)
birth weight	–1.26	(3.65)	–0.30	(2.94)	1.60	(2.94)	0.06	(0.20)
mother's age	–1.34	(1.67)	0.83	(1.35)	0.19	(1.36)	–0.58	(0.23)
less than high school	0.90	(0.67)	–0.03	(0.49)	–0.69	(0.57)	0.11	(0.31)
high school	0.77	(0.62)	–0.25	(0.45)	–0.33	(0.50)	NA	NA
some college	0.34	(0.62)	0.00	(0.45)	–0.24	(0.50)	NA	NA
mother Hispanic	0.21	(0.46)	–0.11	(0.40)	–0.05	(0.44)	NA	NA
mother white	0.40	(0.41)	–0.02	(0.35)	–0.32	(0.40)	NA	NA

Table A. Estimated coefficients and goodness of fit from logistic regression models used to predict principal scores in primary analyses. (*Continued*).

	Stratum A		Stratum B		Stratum C		Strata A/B	
Variables	coef	s.e.	coef	s.e.	coef	s.e.	coef	s.e.
mother married at birth	0.66	(0.33)	0.06	(0.30)	–0.66	(0.31)	0.61	(0.28)
(birth weight)^2	0.38	(1.24)	0.14	(1.00)	–0.58	(1.01)	0.08	(0.14)
(weeks pre-term)^2	0.00	(0.02)	–0.01	(0.01)	0.01	(0.01)	0.00	(0.00)
(mother's age)^2	0.21	(0.30)	–0.23	(0.24)	0.08	(0.24)	NA	NA
weeks pre-term	–0.03	(0.24)	0.11	(0.18)	–0.06	(0.20)	NA	NA
child's sex	–0.42	(0.23)	0.21	(0.19)	0.12	(0.19)	NA	NA
child's birth order	0.32	(0.14)	–0.01	(0.12)	–0.25	(0.14)	0.26	(0.12)
received prenatal care	0.79	(0.61)	0.18	(0.50)	–0.88	(0.46)	0.81	(0.46)
neo-natal health index	0.01	(0.01)	0.00	(0.01)	0.00	(0.01)	NA	NA
worked during preg.	–0.60	(0.29)	0.15	(0.25)	0.33	(0.22)	–0.37	(0.20)
child a twin	0.48	(0.37)	–0.21	(0.31)	–0.18	(0.37)	NA	NA
bw group*bw	–1.32	(2.55)	1.42	(1.99)	–0.53	(2.12)	0.36	(1.03)
bw group*mom's age	–0.16	(0.48)	0.35	(0.39)	–0.26	(0.41)	0.10	(0.29)
bw group*<hs	0.44	(1.06)	–0.57	(0.76)	0.61	(0.83)	–0.16	(0.44)
bw group*hs	0.10	(0.99)	0.03	(0.69)	0.16	(0.75)	NA	NA
bw group*some coll.	1.21	(0.96)	–0.83	(0.68)	0.21	(0.73)	NA	NA
bw group*mom Hispanic	–0.55	(0.76)	0.79	(0.64)	–0.57	(0.75)	NA	NA
bw group*mom white	0.48	(0.60)	–0.10	(0.50)	–0.22	(0.55)	NA	NA
bw group*mom married	–0.85	(0.58)	–0.70	(0.49)	1.60	(0.51)	–1.28	(0.44)
goodness of fit	0.77		0.62		0.69		0.68	

The IHDP was supported by grant 91-01142-00 from the Pew Charitable Trusts, Philadelphia, PA; by grant 5R01 HD 27344 from the National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD; by grants from the Maternal and Child Health Bureau (Title V, Social Security Act), Washington, DC; the Department of Health and Human Services, Washington, DC; and the Robert Wood Johnson Foundation, Princeton, NJ. The authors would also like to thank the NICHD Research Network on Child and Family Well-being and the March of Dimes Foundation. Support for this paper was provided by grant 5R29 HD 35150-04 from the National Institute of Child Health and Human Development and by the William T. Grant Foundation. The authors are also indebted to the editor and two anonymous reviewers for helpful comments and suggestions.

JENNIFER HILL is an Assistant Professor at Columbia University School of International and Public Affairs.
JANE WALDFOGEL is an Associate Professor at Columbia University School of Social Work.
JEANNE BROOKS-GUNN is a Professor at Teacher's College, Columbia University.

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.

Bachu, A., & O'Connell, M. (1998). Fertility of American women. Washington, DC: U.S. Census Bureau.

Barnard, J., Du, J., Hill, J.L., & Rubin, D.B. (1998). A broader template for analyzing broken randomized experiments. Sociological Methods and Research, 27, 285-317.

Barnard, J., Frangakis, C., Hill, J.L., & Rubin, D.B. (2001). School choice in NY city: a Bayesian analysis of an imperfect randomized experiment, in Kass et al. (Eds.), Case studies in Bayesian Statistics,. vol. 5. New York: Springer-Verlag.

Baydar, N., Hyle, P., & Brooks-Gunn, J. (1997). A longitudinal study of the effects of the birth of a sibling during preschool and early grade school years. Journal of Marriage and the Family, 59 (4), 957-965.

Belsky, J. (in press). Development risks (still) associated with early child care. Journal of Child Psychology and Psychiatry.

Blau, D.M. (2001). The child care problem: an economic analysis. New York: Russell Sage.

Blau, F.D., & Grossberg, A.J. (1992). Maternal labor supply and children's cognitive development. Review of Economics and Statistics, 74, 474-481.

Bos, J.M., & Granger, R.C. (2002). Estimating effects of day care use on children's school-readiness: evidence from the New Chance demonstration. New York: MDRC.

Brooks-Gunn, J., Klebanov, P.K., Liaw, F., & Spiker, D. (1993). Enhancing the development of low-birthweight, premature infants: changes in cognition and behavior over the first three years. Child Development, 64(3), 736-753.

Brooks-Gunn, J., McCarton, C.M., Casey, P.H., McCormick, M.C., Bauer, C.R.. Bernbaum, J.C., Tyson, J., Swanson, M., Bennett, F.C., Scott, D.T., et al. (1994). Early intervention in low-birth-weight premature infants. Results through age 5 years from the Infant Health and Development Program. Journal of the American Medical Association, 272(16), 1257-1262.

Currie, J. (2001). Early childhood education programs. Journal of Economic Perspectives, 15 (2), 213-238.

Currie, J., & Thomas, D. (1995). Does Head Start make a difference? American Economic Review, 85(3), 341-364.

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, Technical Working Paper 6829. Cambridge, MA: National Bureau Economic Research.

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.

Frangakis, C.E., & Rubin, D.B. (2002). Principal stratification in causal inference. Biometrics, 58(1), 20-29.

Gibson, C. (2001). Privileging the participant: the importance of take-up rates in social welfare evaluations. Chicago: Joint Center for Poverty Research.

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(2), 336-354.

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

Heckman, J., & Robb, R. (1985). Alternative methods for evaluating the effect of interventions, in Heckman and Singer (Eds.), Longitudinal analysis of labor market data (pp. 156-246). New York: Wiley.

Hill, J., Brooks-Gunn, J., & Waldfogel, J. (2001). Sustained effects of high participation in an early intervention for low-birth-weight premature infants. New York: Columbia University.

Hotz, J.V., Imbens, G., & Mortimer, J.H. (2000), Predicting the efficacy of future training programs using past experiences. Los Angeles: University of California.

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 (4), 778-794.

Infant Health and Development Program. (1990). Enhancing the outcomes of low-birth-weight, premature infants. A multisite, randomized trial. The Infant Health and Development Program. Journal of the American Medical Association, 263(22), 3035-3042.

Karoly, L.A., Greenwood, P.W., Everingham, S.S., Hoube, J., Kilburn, M.R., Rydell, C.P., Sanders, M., & Chiesa, J. (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.

Katz, L.F., Kling, J.R, & Liebman, J.B. (2001). Moving to opportunity in Boston: early results of a randomized mobility experiment. Quarterly Journal of Economics, 116(2), 607-654.

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 and Economic Statistics, 17, 74-90.

Leventhal, T., & Brooks-Gunn, J. (2001). Moving to opportunity: an experimental study of neighborhood effects on mental health. New York: Center for Children and Families, Teachers College, Columbia University.

Little, R.J.A., & Rubin, D.B. (1987). Statistical analysis with missing data. New York: Wiley.

Magnuson, K.A., Gibson, C., & Huston, A. (1999). Structured after-school care programs and children's behavior: findings from an experimental anti-poverty program. Chicago: Northwestern University.

McCarton, C.M., Brooks-Gunn, J., Wallace, I.F., Bauer, C.R., Bennett, F.C., Bernbaum, J.C., Broyles, R.S., Casey, P.H., McCormick, M.C., Scott, D.T., Tyson, J., Tonascia, J., & Meinert, C.L. (1997). Results at age 8 years of early intervention for low-birth-weight premature infants. The Infant Health and Development Program. Journal of the American Medical Association, 277(2), 126-132.

McCormick, M.C., Brooks-Gunn, J., Workman-Daniels, K., Turner, J., & Peckham, G.J. (1992). The health and developmental status of very low-birth-weight children at school age. Journal of the American Medical Association, 267(16), 2204-2208.

McCormick, M.C., McCarton, C., Brooks-Gunn, J., Belt, P., & Gross, R.T. (1998). The infant health and development program: interim summary. Journal of Developmental and Behavioral Pediatrics, 19(5), 359-370.

Newhouse, J.P., & The Insurance Experiment Group. (1993). Free for all? Lessons from the RAND health insurance experiment. Cambridge, MA: Harvard University Press.

NICHD Early Child Care Research Network (2000). The relation of child care to cognitive and language development. Child Development, 71(4), 960-980.

Olkin, I., & Tate, R.F. (1961). Multivariate correlation models with mixed discrete and continuous variables. Annals of Mathematical Statistics, 32, 448-465.

Phillips, D., & Adams, G. (2001). Child care and our youngest children. Future of Children, 11(1), 35-52.

Ramey, C.T., Bryant, D.M., Wasik, B.H., Sparling, J.J., Fendt, K.H., & LaVange, L.M. (1992). Infant Health and Development Program for low birth weight, premature infants: program elements, family participation, and child intelligence. Pediatrics, 89(3), 454-465.

Rosenbaum, P.R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. Journal of the Royal Statistical Society, A, 147(5), 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.

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

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

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

Scott, D.T., Bauer, C.R., Kraemer, H.C., & Tyson, J. (1989). A neonatal health index for preterm infants. Pediatric Research, 25, 263A.

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

Smith, J., & Todd, P. (2000). Does matching overcome Lalonde's critique of nonexperimental estimators? London, ON: University of Western Ontario.

Waldfogel, J., Han, W., & Brooks-Gunn, J. (in press). The effects of early maternal employment on child cognitive development. Demography.

Wilde, E.T., & Hollister, R. (2002). How close is close enough? Testing nonexperimental estimates of effects against experimental estimates of effect with education test scores as outcomes. Madison, WI: Institute for Research on Poverty.

Zaslow, M.J., McGroder, S.M., & Moore, K.A. (2000). The national evaluation of welfare-to-work strategies: effects on young children and their families two years after enrollment: findings from the child outcomes study. Child Trends and DHHS.