CONNECTING CHILD CARE QUALITY TO CHILD OUTCOMES

Drawing Policy Lessons from Nonexperimental Data

GREG J. DUNCAN
Northwestern University, Evanston, Illinois
CHRISTINA M. GIBSON-DAVIS
Duke University, Durham, North Carolina

Effective early childhood intervention and child care policies should be based on an understanding of the effects of child care quality and type on child well-being. This article describes methods for securing unbiased estimates of these effects from nonexperimental data. It focuses on longitudinal studies like the one developed by the National Institute of Child Health and Human Development's Early Child Care Research Network. This article first describes bias problems that arise in analyses of nonexperimental data and then explains strategies for controlling for biases arising from parental selection of child care. Next, it comments on attrition in longitudinal studies and outlines some strategies for addressing possible attrition bias. Finally, it discusses the need to translate "effect sizes" derived from these studies into the kinds of cost and benefit information needed by policy makers.

Keywords: *child care quality; causal links; experimental design*

Informed child care and early childhood intervention policies require knowledge of causal links between the quality and quantity of child care "inputs" and child-based "outputs." Are intensive, early education intervention programs worth their high cost? By how much will children's school readiness improve if smaller staff-to-student ratios or better trained teachers

---

AUTHORS' NOTE: *An earlier draft of this article was titled "Selection and Attrition in the NICHD Childcare Study's Analyses of the Impacts of Childcare Quality on Child Outcomes." We would like to thank Mark Appelbaum, Aletha Huston, Steven Barnett, Jay Belsky, Doug Besharov, Peg Burchinal, Martha Cox, Sara Friedman, Ann Hungerford, Jean Layzer, Katherine Magnuson, and Deborah Phillips for comments on the prior draft, but in no way do we hold them responsible for remaining errors.*

EVALUATION REVIEW, Vol. 30 No. 5, October 2006 611-630
DOI: 10.1177/0193841X06291530

are required? Do $1,000 per child subsidies that promote the use of child care centers improve child outcomes by more than $1,000? Policy analysts and policy makers must grapple with these kinds of difficult questions.

Although certainly not without problems (e.g., Manski 1993; Shadish, Cook, and Campbell 2002), random-assignment intervention studies are viewed by many analysts as the gold standard for providing the causal inferences needed to answer questions like the ones listed above. In the case of child care policies, however, precious few such studies exist, and virtually all of them assess the effects of very expensive, high-quality programs (Barnett 1995; Karoly et al. 1998). The recent evaluation of the Head Start program (Administration for Children and Families 2005) is an important addition to our knowledge of the short-run impacts of center-based programs that have been implemented nationally. Although some of the evaluations of intensive programs have established the costs and benefits of what is possible, neither they nor the Head Start evaluation help in judging the social profitability of less intensive and less expensive policies that, say, promote the use of center-based care or increase the training of child care providers. In the terminology of Shadish, Cook, and Campbell (2002), the expensive "efficacy" trials of the past need to be followed up with "effectiveness" trails of a range of realistic programs.

A common source of data for policy analysis of child care arrangements and quality is the nonexperimental sample survey. The National Longitudinal Survey of Youth (NLSY) and the Panel Study of Income Dynamics (PSID) are national studies that have been used for nonexperimental research on the effects of child care quality (Blau 1999) and Head Start (Currie and Thomas 1995; Garces et al. 2002). The most ambitious longitudinal study of child care quality is the National Institute of Child Health and Human Development's (NICHD's) Study of Early Child Care and Youth Development. The study has followed a cohort of roughly 1,200 children from birth to middle school in ten sites. The study is unique in its comprehensive measurement of the family and child care environments of children in its sample, and its results have been widely cited. Yet because the NICHD and other survey-based studies do not randomly assign children to different child care quality levels, analyses of their data must cope with sources of potential bias.

This article reviews the most important of these sources of potential bias and provides guidance for how nonexperimental studies can best be used to present policy makers with informed answers regarding child care issues. It first discusses the issue of biases from difficult-to-measure factors that affect selection of children into child care settings and then describes how these issues are treated in existing studies. Next, it analyzes the problem of attrition in longitudinal studies and provides ideas for how one might investigate and

adjust for attrition-related bias problems. (A third common source of bias—error in the measurement of child care quality—is discussed in Layzer and Goodson 2006 [this issue]). Finally, it discusses the need to translate estimates of child care "effect sizes" into the kind of cost-benefit calculations policy makers need most.

CHILD CARE SELECTION AND THE OMITTED-VARIABLES PROBLEM

Whether it be child care, neighborhood, or school, all nonexperimental studies assessing the effects of a chosen child "context" on child outcomes have the potential for bias from unmeasured characteristics of the child or the parent that affect both the selection into that context and child outcomes (Manski 1993; Vandell and Wolfe 2000).

To understand how bias arises, we follow Blau (1999) and NICHD and Duncan (2003) in viewing child *i*'s cognitive or social development at time *t*, the point of school entry (Y_{it}), as an additive function of the child's birth-to-time-*t* history of the quality and quantity of home ($HOME_t$) and child care ($CARE_t$) environments, plus time-invariant child ($CHILD$) and maternal and family (FAM) influences. The goal is to estimate β_1, the impact of the quality of the child's history of child care inputs on development:

$$Y_{it} = \alpha_1 + \beta_1 CARE_{it} + \beta_2 HOME_{it} + \beta_3 CHILD_i + \beta_4 FAM_i + e_{it}. \quad (1)$$

The omitted-variable problem arises if difficult-to-measure characteristics of the child, mother, or family environment—elements of $CHILD_i$ and FAM_i in Equation 1—are correlated with both choice of child care quality and children's development. Random assignment of $CARE_{it}$ would ensure that these correlations are effectively zero. If, however, $CARE_{it}$ is chosen by the family, then failure to control adequately for child, mother, and family characteristics will bias estimates of β_1 to the extent that these characteristics are correlated with both child care quality and child development. As shown below, the direction of the bias will depend on the direction of these correlations (NICHD and Duncan 2003).

Most parents have a fairly rich set of child care choices available to them that include family-based, informal, and center-based care. To be sure, the highest-quality, most expensive care is beyond the reach of many, but the range of choices remains broad for most families. The potential child care choice set for a given family expands further if one considers the options available to families if they move into the neighborhood of a family care

provider, decide to work less to supply care themselves, are willing to travel substantial distances, or allocate larger amounts of family income to pay for care. Although those types of parental sacrifices are not restricted to child care choices, the key fact is that the arrangements observed in nonexperimental, survey-based studies such as the NLSY or NICHD Study of Early Child Care and Youth Development reflect parents' decisions.

Child care selection can bias estimates of the effects of child care "quality" in several ways. One bias story involves parents who make sacrifices to obtain high-quality child care for their children. It is likely that such parents promote their children's development in other ways, such as spending their weekends engaging in child-centered activities or reading to their children every night. If this exceptional concern for their children's development is not adequately captured by the *FAM* variable included in Equation 1, then the estimated impact of child care quality will overattribute to child care quality effects that are actually caused by parental concerns and actions.

A similar upward bias story can result from a negative pattern of correlations from parents who, perhaps due to mental health problems, are either unwilling or unable to arrange for good-quality care for their children and are less able to promote their children's healthy development in other ways. Here again, omitting key family variables will create the spurious impression that child care quality matters more than it does.

Although most researchers who speculate about omitted-variable problems believe that they impart upward bias to the coefficients of quality measures, good reasons exist to suspect the opposite. Suppose, as most modern developmental theories allege, that parents and the children themselves are active agents in the child's development. For example, a difficult-to-measure developmental delay in early childhood might motivate a parent to seek unusually high-quality care to address the problem. An adolescent-based example of this phenomenon comes from the Moving to Opportunity residential mobility experiment (Ludwig, Duncan, and Hirschfield 2001). Parents assigned to the "treatment" group were offered counseling and financial assistance to move from public housing in high-poverty neighborhoods to rental housing in low-poverty neighborhoods. Only about half of the families offered the chance to move actually moved. Ludwig, Duncan, and Hirschfield (2001) examined the prior juvenile justice records of teens in families that did and did not move in response to this offer and found twice as many arrests among teens in families that moved compared with teens in families that did not move. Thus, it appeared that the developing behavior problem of their teens motivated parents to take advantage of the program.

If parents were likewise motivated to provide care for their children with developmental or behavioral difficulties, then failure to adjust for child

characteristics prior to entry into child care would impart a *downward* bias to the quality estimates. Delayed or problematic children may be flourishing in their new, high-quality care environments, but their improvement has only just brought them up to a level equal to that of normally developing peers placed in lower quality environments. A study of *changes* in behavior surrounding these child care environments would reveal the behavior improvements associated with the higher quality care, but a typical study that fails to control for child characteristics prior to entry into the child care setting, would not reveal behavior improvements and would thus produce a downward-biased estimate of child care quality.

To round out the omitted-variables bias picture, suppose that parents choose between two strategies: (a) Both parents work and choose high-quality care to compensate for hours they are spending at work rather than with the child; or (b) one parent works part-time and the child is enrolled in less expensive, lower quality care but has more high-quality parental time. If a child develops equally well under the two scenarios, then failure to control for parental employment strategies or the preferences that produce them would make it appear that child care quality does not matter.

Thus, the nasty specter of the omitted-variables problem is revealed: Omitted variables can arise from either parent- or child-based characteristics and can impart either upward or downward bias to the estimated child care quality coefficients. Simple correlations between child care quality and child outcomes are not informative about the nature of "true" effects, given the uncertain direction of bias, and attempts to control for some but not all of them will fail to provide policy makers with valid bottom-line impact estimates.

APPROACHES TO THE OMITTED-VARIABLES PROBLEM

Random assignment of families to different child care quality settings solves the omitted-variables problem by ensuring that characteristics of families and children and, thus, both measured and unmeasured aspects of *FAM* and *CHILD* in Equation 1, are not correlated with the experimental assignment to higher or lower quality care. Lacking random-assignment data, the policy analyst has no truly convincing solution to the omitted-variables problem; nor does it help much to summarize meta-analytic results from large numbers of nonexperimental studies, all of which can suffer from possible bias. What can be done?

Measure the typically unmeasured. One obvious approach to the omitted-variables problem is to try to measure well (and include in the regression

analysis) all the relevant factors that have the potential to affect selection of both child care type and quality as well as child outcomes. It is important to measure selection factors early in the child's life, before actual child care experiences have had a chance to influence them (Hungerford and Cox 2006 [this issue]). Because dozens of possible selection factors exist and many are correlated with one another, regression-based approaches to control simultaneously for all of those factors risk multicollinearity problems. Evidence of such problems includes large standard errors and unstable coefficient estimates (Greene 1993) as well as diagnostics such as the variance inflation factor (Chatterjee, Hadi, and Price 2000). Analysts should certainly be wary of multicolinearity, but lacking evidence of it, they are better off controlling for as many selection factors as possible.

The strategy of comprehensive measurement has been adopted by the NICHD Study of Early Childcare (NICHD 1996), an unusually expensive and high-quality longitudinal study of young children. The study recruited mothers from hospitals near the following locations throughout 1991: Little Rock, Arkansas; Irvine, California; Lawrence, Kansas; Boston, Massachusetts; Philadelphia, Pennsylvania; Pittsburgh, Pennsylvania; Charlottesville, Virginia; Morganton, North Carolina; Seattle, Washington; and Madison, Wisconsin. Potential participants were selected from among 8,986 mothers giving birth during selected 24-hour sampling periods. A series of intensive, in-home interviews were conducted at 1, 6, 15, 24, 36, and 54 months, and the children's school progress has been tracked every other year between kindergarten and sixth grade.

The NICHD study has devoted much of its energy to measuring child care quality on the basis of videotapes of caregiver-child interactions (but see Layzer and Goodson [2006] for a discussion of issues related to child care quality measurement). It also has measured an impressive set of child outcomes—both cognitive and behavioral—in most of its interviewing waves (Zaslow et al. 2006 [this issue]). Most relevant to our discussion is the study's relatively comprehensive measurement of selection variables for the mother and family (income-to-needs ratio, maternal education, marital status, mother's receptive vocabulary test score, personality inventory, home learning environment, and maternal "sensitivity"), child (temperament and gender), and early mother–child interaction (attachment ratings).

To assess the NICHD strategies for controlling for omitted-variables bias in its estimation of the impact of child care quality on child outcomes, we identified 20 of its published papers that contain original analyses of this topic and two additional review papers.[1] Most of these papers acknowledge the selection problem, and some discuss it explicitly (NICHD 1997, 1998, 1999, 2000, 2001b, 2002a, 2002b, 2003). For example, one study that acknowledges

"unobserved selection factors . . . cannot be accounted for" also indicates that the rich set of NICHD characteristics allows them to "get closer to accurate estimates" of the effects of child care (NICHD 2003, 452).

The correlation-based variable-selection strategies applied in NICHD-authored articles lead to relatively few covariates in the final regression models. Of the 33 family and child selection factors available in the data (and used in at least one of the articles), most studies included fewer than half of those factors, and none included more than 11. Although that approach may make sense in the context of small-sample, lab-based studies consisting of a few dozen subjects, the NICHD study includes more than 1,000 children and thus does not have the degree-of-freedom problem that may plague smaller studies. Although most of the articles noted the authors' concern that the inclusion of many selection factors would introduce unacceptable amounts of multicollinearity and imprecise coefficient estimates, none of the results provided evidence that those problems occurred. Several of the earlier articles distinguish between selection and family factors (NICHD 2001a, 2001b, 1999). As defined by the authors, selection factors are potentially biasing variables as a result of their correlations with both the child care measures of interest and child outcome. Family factors, however, are selected "on the basis of theory" along with prior research that has demonstrated the relevance of the variable.

The distinction between selection and family factors is puzzling. All variables in a regression model should be chosen on the basis of theory regarding how that variable influences child outcomes: either directly or indirectly through the choice of child care arrangements. And because small, bivariate correlations sometimes mask important multivariate relationships,[2] strategies for selecting covariates on the basis of simple correlations with child care and child outcomes is inadvisable. It is better to control for as many theoretically relevant factors as possible, provided that they are measured early in the child's life and do not produce unacceptable levels of multicollinearity.

Related to this concern is that comparisons of estimates of the impact of child care quality from regression models with no, some, or complete controls for child- and family-selection factors provide valuable information on the likely scope of still-unmeasured selection factors. NICHD (2002a) adopted this strategy and, in an effort to adjust for selection factors, added five maternal covariates to its original model. NICHD and Duncan (2003) estimated a set of models of effects of child care quality that included various numbers of control variables. They found that family-based control variables reduced the association between child care quality and 54-month cognitive development by roughly one half and that the bulk of the reduction occurred when maternal schooling was first entered into the regression.

One could view the large reductions in coefficient in two different ways. The optimist would say, "Look at the size of the adjustment when we control for the selection factors; surely we have captured most of the selection process." The pessimist would say, "The adjustments indicate that selection factors are obviously important. If measured selection factors can make that much difference, think of how much more bias reduction would come from further adjustments."

Altonji, Elder, and Taber (2005) developed a formal model of omitted-variables bias and uses the size of the adjustments when including observable selection factors to bound the likely total amount of bias. Their model sides with the pessimists. If the included selection factors are a random subset of all potential selection factors, then the substantial changes in child care quality coefficients that occur when selection factors are introduced into the equation suggest that further controls for unmeasured selection factors may matter just as much.

Change models. Another approach to the omitted-variables problem is to estimate change models. To motivate such a model, recall first that Equation 1 considered child development at the point of school entry (t) to be a function of the child's history of child care quality and family inputs plus child- and family-specific variables. Now suppose that an analogous relationship to Equation 1 relates a child's development at a younger point—say, age 2 (denoted as s)—to the child's birth-to-age-s history of the quality and quantity of home ($HOME_s$) and child care ($CARE_s$) inputs, plus invariant child ($CHILD$) and maternal and family (FAM) effects:

$$Y_{is} = \alpha_1 + \beta_1 CARE_{is} + \beta_2 HOME_{is} + \beta_3 CHILD_i + \beta_4 FAM_i + e_{is}. \quad (2)$$

A simple difference model of Equations 1 and 2, using Δ to denote the t–s difference, is

$$\Delta Y_i = \beta_1 \Delta CARE_i + \beta_2 \Delta HOME_i + e_i, \quad (3)$$

where $CARE_i$ is average child care quality *between* age 2 and the point of school entry. The obvious advantage of Equation 3 over Equations 2 or 1 is that the biases associated with the unmeasurable and persistent child and maternal and family characteristics (e.g., $CHILD_i$ and FAM_i) have been differenced out. Note that β_1 has identical interpretations in Equations 1 and 3; both reflect changes in child development associated with a change in child care quality.

Psychologists and educators have been reluctant to rely on change-score analyses because of their greater measurement error. Typically, change scores

are substantially less reliable than the original two scores when those scores are moderately to highly correlated (Cronbach and Furby 1970). Thus, counteracting the advantageous reductions in bias associated with a change equation is the possible disadvantage of greater measurement error in the outcome variable. But measurement error in a dependent variable produces larger standard errors for parameter estimates in a change equation such as Equation 3 than in level Equations 1 or 2 but, importantly, it does not usually bias parameter estimates (Allison 1990; Greene 1993).[3] Measurement error in independent variables, on the other hand, can bias parameter estimates. Note, however, that in Equation 3, $\Delta CARE$ is the difference in average care quality between birth and age t and between birth and age s. This difference is average care quality between s and t, which is well measured in the NICHD study.

Another worry concerning a change equation like Equation 3 is the validity of the assumption that Equations 1 and 2 have the same parameters. If, for example, the impact of child care quality on outcomes falls over time, then β_1 in Equation 1 will be smaller than β_1 in Equation 2, and the simplicity of Equation 3 is lost. Another worry about the change model is that it is much more difficult to measure children's outcomes, especially cognitive outcomes, at younger ages than at older ages. The implications of both possibilities are discussed in NICHD and Duncan (2003), which provides estimates of child care quality impacts based on Equations 1 and 3.

Sibling models. Sibling models (sometimes called "family fixed effects") are another approach to the bias problem. These approaches subtract each sibling's score on the dependent and independent variables from the average scores of all siblings in his or her family. In the case of two siblings per family, the deviation-from-family-means model becomes a sibling difference model. Replacing the subscript i in Equation 1 with 1 (for Sibling 1) and 2 (for Sibling 2) and assuming that sufficient cross-sibling variability exists in family and contextual conditions to reference *FAM*, *CARE*, and *HOME* with the sibling subscripts, the sibling difference model takes the following form:

$$Y_2 - Y_1 = \beta_1(CARE_2 - CARE_1) + \beta_2 (HOME_2 - HOME_1) + \beta_3(CHILD_2 - CHILD_1) + (e_2 - e_1). \quad (4)$$

In terms of measured variables, this approach amounts to estimating a regression in which sibling differences in the outcome of interest are regressed on sibling differences in observed child care quality and family characteristics. Here again β_1 has an identical interpretation in Equation 4 as it did in Equation 1.

Observed parental factors (such as parental educational attainment) that are the same for all siblings in a family are differenced out of Equation 4 and not included in the sibling difference regression. A key advantage of sibling models is that persistent *un*observed elements of parents are differenced out as well, thus eliminating the omitted-variable bias caused by the unmeasured persistent family factors shared by siblings.

Time-varying family factors, especially those that might be producing the sibling differences in the child care context (e.g., divorce and income changes), are a potential source of bias in Equation 4 and should be controlled explicitly in the regression, if possible. Note that those factors will bias estimates only to the extent that they are correlated with the child care differences. If uncorrelated with them, the unmeasured family differences between siblings will lower the explanatory power of a sibling difference model (often a trivial cost) but will not bias the parameter estimates (a key benefit). As with change models, a possible disadvantage of sibling models is that error-ridden measurement of sibling differences in childcare quality can bias parameter estimates toward zero.

Sibling model estimation is not possible with NICHD study, although it is with the PSID and NLSY. Currie and Thomas (1995) and Garces et al. (2002) used sibling models to assess the developmental effects of the Head Start program. Specifically, they compared outcomes of siblings in the NLSY's children cohorts and in the PSID. In each case, they related sibling differences in outcomes to sibling differences in Head Start attendance. They argued that such sibling differences provide a less biased estimate of the effect of Head Start than do nonexperimental studies that compare outcomes of Head Start attendees and a matched set of children from different families who did not attend Head Start. The case for the Currie and Thomas approach is strengthened to the extent that persistent and difficult-to-measure family factors (e.g., unusual parental concern for children's development and maternal depression) influence both enrollment decisions and outcomes common to all children in a family.

In the case of Head Start, a weakness of the sibling approach is that decisions about Head Start enrollment may reflect unmeasured differences in a mother's perceptions of the different needs of her children for Head Start, or the decisions may reflect unmeasured events (e.g., marriage, employment, or eviction) that produce the enrollment differences and have an independent effect on the child outcomes. Key to the success of sibling models is understanding and statistically adjusting for the process by which children from the same family end up in different contexts of interest. If Head Start use varies because of factors beyond the control of families (e.g., introduction or

expansion of the program into a specific geographic area), then the resulting sibling model estimates are more convincing. If Head Start use differs between siblings because of family events that may themselves have detrimental effects on the children (e.g., an eviction forced the family to move away from a service area), then it is more likely that sibling models estimates are biased. Much like the "measure the unmeasured" strategy, a sibling approach requires researchers to devote considerable effort to understanding the determinants of key contexts and why those determinants might differ between siblings (Griliches 1979).

Instrumental variables. Instrumental variables are another technique used to estimate unbiased estimates in nonexperimental settings. The technique rests on the premise that a "instrument" can be identified that is (a) highly correlated with both the key independent variable of interest (*CARE* in our case) and the outcome of interest and (b) only affects the outcome through its influence on *CARE*. State child care regulations is a possible example, because they are beyond the control of families making care choices, affect the quality of care, and may well only affect children's outcome via their affect on quality of care. If such an instrument can be identified, then a first-stage equation regresses quality of care on the instrument and other independent variables:

$$CARE_{it} = \gamma_1 + \gamma_2 INSTRUMENT + \gamma_3 HOME_{it} + \gamma_4 CHILD_i + \gamma_5 FAM_i + e_{it}. \quad (5)$$

Next, the predicted value of child care participation ($PredCARE_{it}$) obtained from the first stage is regressed on the outcome:

$$Y_{it} = \alpha_1 + \beta_1 PredCARE_{it} + \beta_2 HOME_{it} + \beta_3 CHILD_i + \beta_4 FAM_i + e_{it}. \quad (6)$$

The measure of child care quality in Equation 6 thus is purged of any correlation with unobserved characteristics.

Although instrumental variables can be an effective tool for providing unbiased estimates, finding an appropriate instrument is very difficult (Angrist, Imbens, and Rubin 1996; Heckman, Ichimura, and Todd 1997). In the case of child care, it would require identifying a variable that affects child development, but that effect must be fully mediated through child care use. This criterion is difficult to meet, as most variables related to child care selection—maternal education, parenting style and preferences, even residential location—are also likely to affect child outcomes and are therefore not suitable instruments. To date, we know of no studies that have used this technique to estimate the effect of child care on child outcomes.

Propensity scores. Another methodology that has also been used in nonexperimental contexts is propensity scores. Propensity scores involve estimating the probability of participation in a given level of child care quality based on observed characteristics and then generating a predicted probability. This probability can then be used to match similar families who did and did not participate in care, thus creating two "matched" groups that are equivalent save for chosen child care quality. Similar to a randomized experiment, any difference between the two groups must be due to child care use and not because of preexisting characteristics. The advantage of propensity scores is that comparisons of the two groups need only involve the propensity score rather than all of the variables used to create that score. This simplification greatly facilitates the use of various matching schemes (Heckman, Ichimura, and Todd, 1997, 1998). Hill, Waldfogel, and Brooks-Gunn (2002) used a variant of propensity scores to estimate the effect of center-based care for children in the Infant Health and Development Program. The authors found that center-based care offered strong advantages over other care strategies and that these advantages persisted well into the elementary school years.

Constructing propensity scores can be difficult. The technique's "selection on observables" (Heckman and Robb 1985) or the "unconfoundeness assumption" (Imbens 2004), requires that, conditional on the variables used to calculate the propensity score, participation does not depend on unobserved determinants of treatment (Imbens 2004). In the case of child care, the unconfoundedness assumption requires that child outcomes are not influenced by the same factors that influenced a family to use care, unless those factors are also used to construct the propensity score. However, the better the propensity score model is at predicting child care use, the more unlikely it is that there exists a comparable family that did not use care. There must be similarly situated families in both groups, or comparisons cannot be made. This overlap in the distribution of cases is known as the common support region, and propensity scores cannot be used if there is an insufficient common support region (Imbens 2004; Smith and Todd 2005). As illustrated in Gibson-Davis and Foster (2006), meeting both of these assumptions in a nonexperimental contexts can be quite challenging.

Regression discontinuity designs. One of the most promising approaches for analysis of nonexperimental data is the regression discontinuity design (Shadish, Cook, and Campbell 2002). Gormley et al. (2005) used this approach in their evaluation of the Tulsa pre-K program, a high-quality program model that uses college-trained teachers in its classrooms. A strict September 1 deadline determined age-eligibility for the program and produced instances where children born just before or after the program's

deadline either had or did not have pre-K program experience a year later, when the children's achievement and behavior was tested. The analysis focused on whether the otherwise monotonic relationship between a child's birth date and the outcomes was broken around the point of September 1.

Longitudinal studies such as the NLSY or NICHD Study are often ill suited for regression discontinuity approaches because the method often requires gathering longitudinal data in a specific geographic location or surrounding a particular point in time in which a particular policy was in effect. Nor are they well suited for studies of child care quality in most child care settings, because there is typically no policy for which eligibility requirements are sharply discontinuous. This is not to argue against such methods. Indeed, it makes the most sense to begin with the research questions and then seek the best data—with random assignment data at the top of list—to answer the question rather than to begin with a data set and ask what policy questions might be answered with it.

All told, analysts who seek policy-relevant impact estimates of child care quality face a menu of analytic choices. Because all estimating methods have strengths and weaknesses, the preferred strategy is to seek convergence among estimates obtained from different methods.

SAMPLE ATTRITION

All longitudinal studies suffer from sample attrition. In the PSID, 59% of 1968 respondents also responded 26 years later to its 1994 follow-up. In the NLSY, the cumulative response in 1996, 17 years after its initial interviewing wave, was 79% (Hernandez 2005). As documented in Duncan and Gibson (2000), unplanned attrition as a result of refusals and inability to locate families at various stages of the NICHD study selection process produced a 52.5% cumulative response rate when the children were 6 months old. Substantial nonresponse provides ample opportunity for attrition bias. Children not observed (both by in-home assessments and by child care experience) were more likely to be from less advantaged households: They were less likely to be married, had lower incomes and levels of maternal education, and were more likely to be a member of a racial or ethnic minority (NICHD 1999, 2000, 2001a).

One common approach to adjusting for attrition bias is to carefully distinguish families dropping out of the study by design (i.e., as a result of planned exclusions from conditional sampling and other exclusion criteria) from the worrisome dropouts resulting from refusals, failure to locate, and so on and develop a set of weights formed by taking the inverse of the predicted response

rate of the child used in a particular analysis. The predictions would come from analysis of the complete sample of families not deliberately excluded from the study. Both the PSID and NLSY adjust their probability-of-selection weights for attrition. Other methods of modeling nonresponse (e.g., Rubin's [1974] propensity scores) could be used as well.

CAUSATION IN NONEXPERIMENTAL STUDIES

Large disciplinary differences exist in the language and practice of causal modeling. For example, economists almost universally endorse the practice of basing causal inferences from nonexperimental data on a well-specified theoretical and empirical model of the causal process along with explicit attention to cataloguing and, to the extent possible, addressing the various statistical threats to causal inference (Holland 1986; Shadish, Cook, and Campbell 2002). In this view, the validity of causal inference depends on the adequacy of the causal model and the ability of the empirical analysis to address the possible sources of bias. Theoretical and statistical assumptions about modeling are explicit and open to challenge; at the same time, nonexperimental studies that address model and data-based concerns can inform policy by providing estimates that approximate causal impacts.

Developmental studies are usually careful to point out when their data do not come from a randomized experiment. As with much of the nonexperimental literature in developmental psychology, most of the articles then go on to assert that, as a consequence, it is impossible to draw causal inferences from the analysis. Indeed, much of their language describing results is couched in terms of "associations" between child care quality and child outcomes. It is not uncommon, however, to see these papers make explicit statements about effects, and others draw explicit policy conclusions. For instance, NICHD (1997, 876) stated, "The interaction analyses provided evidence that high-quality child care served a compensatory function for children whose maternal care was lacking." On the policy side, NICHD (2002c, 199) asserted, "These findings provide empirical support for policies that improve state regulations for caregiver training and child-staff ratios."

One cannot have it both ways. Studies that do not aspire to causal analysis should make no claim whatsoever about effects and draw no policy conclusions. At the same time, it would be a terrible waste of resources to conduct expensive longitudinal studies without attempting to use them for causal modeling.

EFFECT SIZES AND COSTS

A key policy question regarding early childhood interventions and quality differences across typical child care settings involves effect sizes and whether the dollar value of the effects of increased quality are worth the costs (Blau 2002). This question is hard to answer because it is difficult to quantify costs and, especially, benefits. Nevertheless, some steps could be taken in translating effect sizes into more policy-relevant numbers.

Cohen's (1988) rules about what constitute small and large effect sizes are often cited in the developmental literature. He views a 0.20 standard deviation (*SD*) as small, a 0.50 *SD* as moderate, and a 0.80 *SD* as large. In the context of child care quality, those rules focus on the fraction of *SD* change in the child outcome (dependent) variable associated with a one-unit change in the child care quality (independent) variable. Surely, rules such as these are incomplete, if not misleading, if they are not tied to program costs and benefits.

Economists would argue that policy analysis is much better served with knowledge of the value of the benefits associated with the effects of child care quality relative to the costs of achieving those benefits (Gramlich 1990; Levin 1983). Small effect sizes that are inexpensive to generate may well be worth it, whereas big effects from expensive interventions may not be.

It is quite possible to have a cost-effective intervention that accounts for only a small percentage change in the outcome of interest. The welfare-to-work literature is filled with examples of cost-effective job search programs that have relatively small effects but cost so little that their benefit-cost ratios still exceed one (Gueron and Pauly 1991). Thus, for policy purposes, reliance on Cohen-type rules governing effect size, in the absence of cost considerations, appears misguided.

Suppose that one has somehow succeeded in obtaining unbiased causal estimates of β_1, the impact of child care quality on child outcomes. "Raw-score" coefficient estimates reflect the change in the child outcome caused by a "one-unit" change in quality. Estimates from NICHD and Duncan (2003) suggest that moving from a 2 to a 3 on the study's 4-point child care quality scale sustained for 2.5 years between ages 24 and 54 months is associated with a 0.08 to 0.16 *SD* increase in a children's cognitive test scores. On a standard IQ scale (on which the *SD* is 15), this increase amounts to 1.2 to 2.4. By Cohen's (1988) standards, those effects are small indeed. But the more relevant question is whether the dollar value of increasing children's achievement by these amounts more than outweighs the cost. Krueger and Whitmore (2001) showed that the 0.20 *SD* increase in achievement in the Tennessee Star classroom size experiment, if permanent, can translate into sizable dollar benefits.

NICHD (2000) used a top- versus bottom-quartile method for computing effect sizes. In effect, the one-unit change in the quality measure involves a change from the average child in the lowest child care quality quartile to the average child in the highest quartile. This approach does not provide the reader with a good sense of how large the average quality difference is between the two groups. The paper states that the breakpoints for the bottom and top quartiles are roughly 1 *SD* below and above the mean—a 2 *SD* difference. But the average child in the bottom-quality quartile is well below the 25th percentile, whereas the average child in the top-quality quartile is well above the 75th percentile. Thus, the implicit quality comparisons in these papers amount to perhaps a 3 *SD* "treatment." It is vital in all studies to clarify exactly what a one-unit change on key independent variables amounts to.

In the absence of a natural metric, a useful method for expressing the impact of child care quality on child outcome is to scale both child care quality and child outcomes in *SD* units. The quality score used in the NICHD study has an *SD* of about 0.5, so that a one-unit change in the quality measure amounts to 2 *SD* (NICHD and Duncan 2003). When both the quality measure and outcome are expressed in *SD* units, the "effect" of a 1 *SD* increase in child care quality in the NICHD and Duncan (2003) study reduces to 0.04 to 0.08 *SD* in school readiness.

Even if the 0.04 to 0.08 *SD* range of effect sizes brackets the impact of increments to child care quality, it remains unclear whether they are large enough to be of policy importance. They are certainly small by Cohen's (1988) standard and much smaller than those reported in experimental studies of early preschool intervention programs offering levels of quality that routinely exceed those of typical community-based child care programs focusing on children at risk due to both economic and developmental factors. For example, treatment effect sizes on IQ were 0.75 *SD* at age 5 for the 3- and 5-year treatment of the Abecedarian Project and 0.60 *SD* for the 1- to 2-year Perry Preschool Project (Ramey, Bryant, and Suarez 1985).

But intensive programs were quite expensive. The policy question for the smaller effect sizes emerging from the NICHD data is whether the cost of raising child care quality by 1 *SD* is less than the dollar value of the benefit of a 0.04 to 0.08 *SD* boost in cognitive scores. If it cost $5,000 per year for several years to produce the gains observed in the NICHD study, then it may make sense to consider cheaper alternatives (e.g., supplementing family income directly or expanding highly targeted pre-K programs) for increasing children's test scores. Unfortunately, we know precious little about the economic costs of providing different levels of childcare quality. Future work with these data should address some of these effect-size, benefit, and cost issues.

SUMMARY

Effective intervention and child care policies should be based on an understanding of the impact on child well-being of intensive early childhood interventions as well as typical improvements in child care quality. While noting attempts to evaluate Head Start and other center-based early education interventions, this article has focused on drawing policy-relevant causal inferences regarding the impact of child care quality using data from the longitudinal study developed by the NICHD Early Child Care and Youth Development Research Network. From the analysis presented in this article, one can draw the following conclusions:

1. Analytic strategies that rely on measured variables to control for bias ought to include all theoretically relevant control variables.
2. When study nonresponse is substantial, efforts should be made to investigate whether nonresponse might be biasing parameter estimates and, if so, to take measures to adjust for the bias.
3. A variety of methods, including change, sibling and regression discontinuity models, could be used to address possible bias.
4. Whether effect sizes are "large" or "small" is less important for policy than a comparison of benefits and costs.

No single approach will provide truly convincing estimates of causal impacts from nonexperimental data such as these, but policy makers cannot wait for the needed experimental studies. The best strategy for the data from the NICHD study and other child studies is to push the data as far as possible toward the goal of securing convincing conclusions about causation, search for robust findings across the set of studies, and then consider the costs and benefits associated with consensus estimates of the impact of child care quality (Light and Pillemer 1984).

NOTES

1. We are grateful to the network researchers and staff who helped us identify these studies.

2. Suppose, for example, that low- and high-income children are equally likely to be in high-quality care because low-income families are eligible for subsidies and high-income families can afford to pay for it. The simple correlation between income and quality is zero, yet income is theoretically important as a determinant of care and child outcomes. The relationship between income and care in areas with no local child care subsidies may be quite strong.

3. National Institute of Child Health and Human Development Early Child Care Research Network and Duncan (2003) found little loss in statistical precision in going from a level to a

change model. For example, the standard errors in their estimated effects of child care quality between ages 24 and 54 months were only about 5% to 10% larger in the change model than in the level model.

REFERENCES

Administration for Children and Families. 2005. *Head Start impact study: First year findings.* Washington, DC: U.S. Department of Health and Human Services. http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/first_yr_finds/firstyr_finds_title.html.

Allison, P. 1990. Change scores as dependent variables in regression analysis. In *Sociological methodology*, ed. C. Clogg, 93-114. Oxford, UK: Basil Blackwell.

Altonji, J., T. Elder, and C. Taber. 2005. Selection on observable and unobservable variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy* 113:151-84.

Angrist, J., G. Imbens, and D. R. Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 90:444-55.

Barnett, W. S. 1995. Long-term effects of early childhood programs on cognitive and school outcomes. *The Future of Children* 5:25-50.

Blau, D. 1999. The effects of child care characteristics on child development. *Journal of Human Resources* 34:786-822.

———. 2002. *The child care problem.* New York: Russell Sage Foundation.

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences.* 2nd ed. Hillsdale, NJ: Lawrence Erlbaum.

Chatterjee, S., A. Hadi, and B. Price. 2000. *Regression analysis by example.* 3rd ed. New York: John Wiley.

Cronbach, L., and L. Furby. 1970. How should we measure "change"—Or should we? *Psychological Bulletin* 74:16-21.

Currie, J., and D. Thomas. 1995. Does Head Start make a difference? *American Economic Review* 85:341-64.

Duncan, G. J., and C. M. Gibson. 2000. Selection and attrition in the NICHD Child Care Study's analyses of the impacts of childcare quality on child outcomes. Manuscript, Northwestern University, Evanston, IL.

Garces, E., D. Thomas, G. J. Duncan, and J. Currie. 2002. Longer-term effects of Head Start. *American Economic Review* 92:999-1012.

Gibson-Davis, C. M., and E. M. Foster. 2006. A cautionary tale: Using propensity scores to estimate the impact of food stamps on food insecurity. *Social Service Review* 80:93-126.

Gormley, W., T. Gayer, D. Phillips, and B. Dawson. 2005. The effects of universal Pre-K on cognitive development. *Developmental Psychology* 41:862-84.

Gramlich, E. 1990. *A guide to cost-benefit analysis.* Englewood Cliffs, NJ: Prentice Hall.

Greene, W. H. 1993. *Econometric analysis.* 2nd ed. New York: Macmillan.

Griliches, Z. 1979. Sibling models and data in economics: Beginnings of a survey. *Journal of Political Economy* 87:S37-S64.

Gueron, J., and E. Pauly. 1991. *From welfare to work.* New York: Russell Sage Foundation.

Heckman J. J., H. Ichimura, and P. E. Todd. 1997. Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. *Review of Economic Studies* 64:605-54.

———. 1998. Matching as an econometric evaluation estimator. *Review of Economic Studies* 65:261-94.

Heckman, J. J., and R. Robb. 1985. Alternative methods for evaluating the impact of interventions. In *Longitudinal analysis of labor market data*, ed. J. Heckman and B. Singer, 156-246. Cambridge, MA: Harvard University Press.

Hernandez, D. 2005. Comparing response rates for SPD, PSID, and NLSY. http://www.sipp.census.gov/spd/workpaper/spd-comp.htm.

Hill, J. L., J. I. Waldfogel, and J. Brooks-Gunn. 2002. Differential effects of high-quality child care. *Journal of Policy Analysis and Management* 21:601-27.

Holland, P. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81:945-60.

Hungerford, A., and M. J. Cox. 2006. Family factors in child care research. *Evaluation Review* 30:**PP.**

Imbens, G. W. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* 86:4-29.

Karoly, L., P. Greenwood, J. Everingham, R. Kilburn, P. Rydell, M. Sanders, and J. Chiesa. 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.

Krueger, A., and D. Whitmore. 2001. The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR. *Economic Journal* 11:1-28.

Layzer, J. I., and B. D. Goodson. 2006. The "quality" of early care and education settings: Definitional and measurement issues. *Evaluation Review* 30:**PP.**

Levin, H. M. 1983. *Cost effectiveness: A primer*. Beverly Hills, CA: Sage.

Light, R., and D. Pillemer. 1984. *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Ludwig, J., G. J. Duncan, and P. Hirschfield. 2001. Urban poverty and juvenile crime: Evidence from a randomized housing-mobility experiment. *Quarterly Journal of Economics* 116 (2): 665-79.

Manski, C. 1993. Identification of endogenous social effects: The reflection problem. *Review of Economic Studies* 60:531-42.

National Institute of Child Health and Human Development Early Child Care Research Network. 1996. Characteristics of infant child care: Factors contributing to positive caregiving. *Early Childhood Research Quarterly* 1:269-306.

———. 1997. The effects of infant child care on infant-mother attachment security: Results of the NICHD Study of Early Child Care. *Child Development* 68:860-879.

———. 1998. Early child care and self-control, compliance and problem behavior at 24 and 36 months. *Child Development* 69:1145-70.

———. 1999. Child care and mother-child interaction in the first three years of life. *Developmental Psychology* 35:1399-1413.

———. 2000. The relation of child care to cognitive and language development. *Child Development* 71:960-80.

———. 2001a. Child care and family predictors of preschool attachment and stability from infancy. *Developmental Psychology* 37:847-62.

———. 2001b. Early child care and children's peer interaction at 24 and 36 months. *Child Development* 72:1478-1500.

———. 2002a. Early child care and children's development prior to school entry: Results from the NICHD Study of Early Child Care. *American Educational Research Journal* 39:133-64.

———. 2002b. Maternal employment and child cognitive outcomes in the first three years of life: The NICHD Study of Early Child Care. *Child Development* 73:1052-63.

———. 2002c. Structure>process>outcome: Direct and indirect effects of caregiving quality on young children's development. *Psychological Science* 13:199-206.

———. 2003. Does quality of child care affect child outcomes at age 4.5? *Developmental Psychology* 39:451-69.

National Institute of Child Health and Human Development Early Child Care Research Network, and G. J. Duncan. 2003. Modeling the impacts of child care quality on children's preschool cognitive development. *Child Development* 74:1454-75.

Ramey, C. T., D. M. Bryant, and T. M. Suarez. 1985. Preschool compensatory education and the modifiability of intelligence: A critical review. In *Current topics in human intelligence*. Vol. 1, *Research methodology*, ed. D. Detterman, 247-96. Norwood, NJ: Ablex.

Rubin, D. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66:688-701.

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

Smith, J., and P. Todd. 2005. Does matching overcome Lalonde's critique of nonexperimental estimators? *Journal of Econometrics* 125:305-53.

Vandell, D., and B. Wolfe. 2000. *Child care quality: Does it matter and does it need to be improved?* Washington, DC: U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation.

Zaslow, M., T. Halle, L. Martin, N. Cabrera, J. Calkins, L. Pitzer, and N. G. Margie. 2006. Child outcome measures in the study of child care quality. *Evaluation Review* 30:**PP.**

Greg J. Duncan is the Edwina S. Tarry Professor of Education and Social Policy and a faculty fellow at the Institute for Policy Research at Northwestern University. Prior to joining the Northwestern faculty in 1995, he served as principal investigator of the Panel Study of Income Dynamics project at Michigan for the previous 13 years. He has published extensively on issues of income distribution, child poverty and welfare dependence. He was elected to the American Academy of Arts and Sciences in 2001.

Christina M. Gibson-Davis is an assistant professor at the Terry Sanford Institute of Public Policy at Duke University, with a secondary appointment in social and health sciences, and an affiliate of the Center for Child and Family Policy. Her work analyzes the well-being of low-income families, with particular emphasis on the causes and consequences of multiple family structures. She is also a research affiliate of the National Poverty Center at the University of Michigan.