When Do We Really Know What We Think We Know? Determining Causality

Janet Currie
UCLA and NBER

June 2003

I would like to thank Diane Halpern for helpful comments.

Social scientists are often asked to determine whether or not one thing causes another. The answer to this question of causality may have important implications for public policy. However, it is generally difficult to establish that "A causes B" beyond a shadow of a doubt, and researchers often arrive at conflicting conclusions depending on their data sources and methods. This conundrum is of course, not confined to social science. Researchers in the "hard" sciences and medicine often come to conflicting conclusions regarding questions such as what killed the dinosaurs, whether there is global warming, and whether hormone replacement therapy is safe and effective for older women.

This chapter considers some of the methods that social scientists working in the area of work and family use to get at the question of causality. It provides a general overview of some of the issues and problems, and then discusses these issues in the context of two specific examples: the effect of maternal employment on child well-being, and the effect of child care quality on children's outcomes. In both cases, studies have arrived at differing conclusions, but in the first case, a range of studies using different data sets and techniques is providing the basis for an emerging consensus, while in the second, the issue of selection has not yet been satisfactorily addressed.

I conclude with the reminder that replication is at the heart of science, and findings must be reproducible before they can provide a reliable basis for policy. Moreover, given the ubiquity of the sample selection problem in social science, the issue must be addressed, preferably using a number of techniques.

1. The Problem of Selection and Potential Responses

In social science, questions about causality are clouded by the problem of sample selection. For example, mothers who work tend to be healthier and better educated on average than those who do not. Hence, it would not be surprising to find that their children did better than those of non-working mothers on average, even if there were no causal relationship at all between maternal employment and child

outcomes. Similarly, children from more advantaged backgrounds are likely to be in better child care, making it difficult to distinguish between the effects of child care quality and the effects of family background.

In principal, the problem of sample selection can be solved using experimental methods: For example, if it were possible to randomly assign women to the "working" and "nonworking" groups, then we could just compare the outcomes of the children of the two groups of women in order to determine the causal effect of maternal employment. As long as the sample size is large enough, there will be no difference in the other observed or unobserved characteristics of the women on average, because the assignment to the two groups was randomly determined. This suggests that it is possible to check that the random assignment "worked" by comparing mean values of the observable characteristics of the two groups. If the means are not statistically significantly different, then we can assume that the unobservables are also similar across the two groups.

However, the sheer ridiculousness of this example highlights one of the main problems with relying on experiments in social science. Women are not plots of land who can be randomly assigned different fertilizer treatments. Experiments with human subjects often run into several difficulties. First, individuals who are not satisfied with their group assignment may take measures to effectively change groups. For example, Heckman, Hohmann and Smith (2000) discuss experimental evaluations of job training programs in which individuals assigned to the "no training" control groups often enrolled in training programs at their own expense. Controls may also be more likely than treated individuals to leave the study. Such differential attrition threatens the validity of the experiment by making it less and less likely that the mean characteristics of the controls will be equal to the mean characteristics of the treatment group.

Conversely, not everyone assigned to the "treatment" group showed up for training. It is important to keep in mind that what is randomly assigned is one's initial allocation to either the treatment

or control group (this is sometimes called the "intent to treat"). Hence, considering outcomes only among those who were assigned to the treatment group AND showed up for training would partially nullify the benefit of the random assignment, since individuals who choose to actually take the training course are a self-selected sub-sample of the treatment group which probably does not have the same average characteristics as the controls.[1]

These problems suggest that experiments in social science are not a panacea. In addition to problems with the implementation of random assignment and attrition, social experiments are sometimes objected to on ethical grounds. For example, it may be objectionable to some that a potentially beneficial "treatment" could be withheld from controls. However, it should be kept in mind that if we actually knew for certain that the treatment was beneficial, there would be no need to conduct an experiment. Secondly, "treatments" are often rationed because funding is insufficient to serve all those who might benefit. For example, only about two-thirds of children eligible for quality preschool education services through Head Start can be served. Where funding is limited, random assignment can replace administrator discretion in deciding who gets access to the treatment without violating any ethical principal.

Heckman and Smith (1995) raise two additional objections to experiments. First, they are often very costly, especially relative to analysis of already existing data sets. Second, it may be difficult to generalize the results from an experiment to the wider population. For example, experimental evaluations of Head Start tend to show "fade out" in its effects on children's cognitive test scores. However, these experiments focused on inner-city African-American children. A series of studies (Currie and Thomas, 1995; Currie and Thomas, 2000; Garces, Thomas, and Currie, 2002) of national populations of children enrolled in Head Start show that effects on outcomes including cognitive test scores tend to be

[1]One way to get around this problem is to use the random assignment as an instrumental variable for whether or not the person got "treated". See Katz, Kling, and Leibman (2001) for an example of this approach in the context of a public housing mobility experiment. Instrumental variables methods are discussed further below.

more long-lived among Hispanic and non-Hispanic white children. Thus, critics of Head Start had incorrectly extrapolated from experimental studies of Head Start to conclude that the program had no longer-term effect on children, when in fact, it appears to have longer-term effects on at least some groups of children.

One potential response to the selection problem in non-experimental data is to try to control for the confounding variables directly. In the example above, if a spurious correlation between maternal employment and child outcomes was driven by maternal health and education, then including adequate controls for these two variables would eliminate the selection problem. The problem of course is that it is always possible that there are other, unobserved confounders. Researchers sometimes try to investigate this possibility by progressively adding variables to regression models. If the parameter of interest does not vary when controls are progressively added, and if unobservable variables are likely to be correlated with the observables, then it may be the case that adding further controls would also have little effect.

A second approach to sample selection is to use "fixed effects" to control for fixed, unobservable, characteristics that may be associated both with selection into the sample and with outcomes. For example, suppose that women with a certain personality type (i.e. very nurturing) are both less likely to work, and more likely to have children with good outcomes. I can regard the woman's personality as a fixed characteristic that is unobserved, but correlated with both the probability of employment and child outcomes. Estimates that do not take account of differences in personality type across the population will produce biased estimates of the effects of employment–it will appear that employment causes bad child outcomes, whereas in reality it is the less nurturing personalities of the employed women that cause these outcomes.

However, if there are women who work during the childhood of one sibling but not during the childhood of the other, then I can get an alternative estimate of the effect of maternal employment by comparing the outcomes of siblings. This estimate will not be affected by personality type (as long as

both children respond to personality type in the same way, which is a strong assumption) because both children are exposed to the same type. That is, by "differencing" the observations on the siblings, one can "difference out" the effect of the omitted variable. In other contexts, we might wish to include child-specific fixed effects (e.g. to look at changes in a child's outcomes with changes in their circumstances), or child care-center specific fixed effects, depending on the type of unobserved variables we were attempting to control.

Fixed effects estimates are subject to a number of potential problems. First, and most obvious, the relevant omitted variable may not be fixed over time. In the example above, a mother might become more nurturing between children, or might be more nurturing with one child (a girl) than with the other (a boy). Second, the fixed effects strategy often ends up reducing the effective sample size a good deal. In the example above, the effect of employment is identified only from women who change employment status between children. Mothers who are either always employed, or always not employed do not contribute to the identification of the effects of employment.

Third, the mothers who change status may not be representative of the initial population of mothers. Ideally, we would like to know why some mothers changed employment status between the birth. Suppose for example, that a major economic downturn led many women to lose their jobs. One might expect this crisis to have its own negative effect on child outcomes. Alternatively, a woman who worked while one child was an infant, might not work during the infancy of a child with health problems or developmental problems (c.f. Powers, 2001). In this case, there would be a spurious positive correlation between maternal employment and child outcomes in the fixed effects models.

A fourth problem is that in the presence of random measurement error, fixed effects estimates are generally biased towards zero. Intuitively, we can divide a measured variable such as a test score into a true "signal" and a random "noise" component. The true signal may be very persistent between siblings (e.g. if both children have high IQ), while the noise component may be more random (e.g. one child has a

bad day on the day of the test). Hence, when we look at the difference between siblings, we can end up differencing out much of the true signal (since it is similar for both siblings) and being left only with the noise.

While it is important to keep these potential problems in mind, fixed effects do offer a very powerful way of controlling for constant unobserved background characteristics. The Head Start studies discussed above typically find that child outcomes are negatively correlated with having been to Head Start in Ordinary Least Squares models. However, Head Start children come from very poor backgrounds relative to all children. When this is controlled for by including household fixed effects, then positive effects of Head Start become apparent. In other words, Head Start children do systematically better than siblings who did not attend Head Start, though they do worse than other children on average.

A third way to deal with selection is to use an instrumental variable of some sort. An instrument is something that is correlated with the "endogenous variable" of interest, but is not correlated with the omitted variables, and therefore is not correlated with the outcome variable (except through its effects on the endogenous variable). In order to implement instrumental variables one can estimate a Two-Stage Least Squares model, in which the endogenous variable is first regressed on the instrument (and all the other exogenous variables in the model). In the "second stage", the model of interest is estimated including the predicted value of the endogenous variable derived from the first stage and adjusting the standard errors appropriately (standard statistical packages do this automatically). Given that the instrument is uncorrelated with the omitted variables, the predicted value of the endogenous variable will also be uncorrelated with them so that the estimation is purged of the bias that results from these omitted variables.

In a model in which child outcomes depend on maternal employment, maternal employment is the endogenous variable. It is chosen by the mother, and may be affected by other unmeasured variables, which are in turn also related to child outcomes. The instrument should be something which predicts

maternal employment, but is uncorrelated with the other omitted variables and has no independent effect on outcomes (once its effect through maternal employment is accounted for). To continue with the example from above, an economic downturn might be predictive of maternal employment, but probably would not be a valid instrument because it might be expected to have effects on child outcomes though pathways such as reductions in father's employment and reductions in school spending and other social services.

It is generally very difficult to find valid instruments, though a number of studies in recent years have used changes in laws as instruments for involvement in various social programs (these studies are often referred to as "natural experiments"). Moreover, it is not really possible to test the validity of ones maintained assumptions about the instruments, though it is possible to test that different instrumental variables yield consistent results (this is the essence of what an over-identification test does). An additional problem is that even when valid, in the sense that they are uncorrelated with other omitted variables, instruments may be "weak" in the sense that they do not explain much of the variation in the endogenous variable. Weak instruments often lead to large standard errors in the "second stage" regression, and may also in some cases lead to biased and misleadingly precise estimates (see Staiger and Stock (1997) and Bound, Jaeger and Baker (1995) for further discussion of the problem of weak instruments and some diagnostic tests).

Instrumental variables estimation is closely related to Heckman's (1979) "selection correction" method. In this procedure, one first estimates a probit model predicting the probability that an individual is selected into the sample. One then uses this model to construct a control for the probability of selection (the "inverse Mill's ratio"). This term is then included in the model of interest in order to "correct" for selection bias. In principal, there ought to be variables (akin to instruments) that predict the probability of being selected into the sample, but which do not have any independent effect on outcomes. But because the selection correction term is a non-linear function of the data, it is possible in practice to

estimate Heckman selection-correction models without these exclusion restrictions. Such a model is estimable only because the selection correction term is assumed to be a non-linear function of the data, while the main equation of interest is assumed to be linear (or to have non-linearities of some other, known form). Since these functional form assumptions seldom have any basis in theory, they form a tenuous basis for identification. Many standard econometrics textbooks now recommend that the Heckman correction method be used only if there are credible exclusion restrictions (c.f. Johnston and Di Nardo, 1997 p. 450) in which case one could also use instrumental variables.

More recently, increasing numbers of social scientists are turning to "propensity scores" as a way to deal with selection bias, following Rosenbaum and Rubin (1983). The main idea of propensity scoring is that treatments and controls should be matched. That is, we want to compare the child outcomes of employed and non-employed mothers who have the same characteristics. As in the Heckman method, the first step is to estimate a probit model of the probability of selection into the treatment (in this example, employment). A predicted propensity of being in the treatment group is then assigned to each person in the data set. There are generally no variables that are excluded from the main model and included in the selection equation, unlike in the Heckman selection procedure.

There are several ways to use these scores to actually construct the match between treatments and controls. By far the most common is to divide people into a number of "bins" (often starting with five) on the basis of propensity scores. Within each bin, the average observable characteristics of those who participate and don't participate in the treatment should be equal (just as they are in an experiment with random assignment). One can then obtain the effect of the "treatment on the treated" by taking the weighted sum of the differences between the treatments and the controls in each bin.

If the average characteristics are not equal within bins, then it is necessary to develop a more refined model of the propensity to be in the treatment group. Also, it is often helpful to exclude treatments and controls who do not overlap at all in terms of their observable characteristics. For

instance, if all women with Ph.Ds worked, and all women with less than 8th grade education did not work, then one might wish to exclude these two categories from the analysis, since these women are sufficiently different that comparing them could not shed any light on the effect of maternal employment per se.

The Propensity Score approach does not rely on distributional assumptions (in contrast to the "Heckman correction") or exclusion restrictions (in contrast to Instrumental Variables analysis). As in the fixed effects method, a limitation of propensity scoring is that there have to be both treatments and controls within a bin in order for the observations in the bin to tell one anything about the effect of interest. If for example, the top and bottom quintile of the propensity score distribution held only non-employed and employed women, respectively, then we would ignore 40 percent of the observations in our propensity score analysis. However, if these women have very little in common, it might be argued that comparing the outcomes of their children was not going to be a very meaningful way to assess the effects of maternal employment in any case. The propensity score method then does make it very clear what part of the data is providing the comparison.

The main limitation of the Propensity Score method is that it does not truly address the question of selection on unobservable variables. Treatment and controls end up being "balanced" within bins on the basis of observables, but this does not rule out the possibility that they are systematically different in terms of unobserved characteristics.

This brief overview suggests that all the methods for dealing with selection bias have pros and cons. In the next two sections, we consider what can be learned from a comparison of studies using different methods in the context of two specific examples.

2. Assessing the Assessments: Maternal Employment and Child Outcomes

Concern about the effects of maternal employment on child well-being has been spurred by dramatic changes in the labor force participation rates of mother's with young children. This increase has

been greatest among married women with young children: For example, in 1950 11.9% of married women with children under six were in the labor force, compared to 62.8% in 2000. Never married, separated, and divorced mothers also increased their labor force participation dramatically with the most rapid growth occurring in the last three decades. In 2000, 65.3% of single women with children under 6 were in the work force (BLS, 2001).

At the same time, studies of brain development in young children, and the development of attachment theory, have led to concern that children will suffer harmful effects from being separated from a primary care-giver (usually the mother) at too early an age (c.f. Ainsworth et al. 1978; Bowlby, 1969). In addition to potentially reducing the amount of time the mother spends with the child, maternal employment could reduce the quality of interaction by increasing the mother's stress levels. On the other hand, mothers presumably work to earn income, and there are many studies suggesting a positive association between income and child well-being. The effects of non-maternal care are also likely to be mediated by the quality of that care. For example, Bianchi (2000) suggests that time that mothers spend in employment may be offset by fathers who spend increased time with their children. It may not be surprising then, that numerous studies of the effects of maternal employment on young children have found evidence of positive, zero, and negative effects.

This section of the chapter attempts to classify studies by methodology, in order both to highlight the ways in which the conclusions vary, and to detect whether there is a consensus emerging from the wide variety of different studies. The survey provided here is not intended to be comprehensive. Readers interested in recent, more comprehensive surveys of the literature on the effects of maternal employment should consult Hoffman and Youngblade (1999) and Zaslow and Emig (1997). In order to keep the discussion focused, this chapter emphasizes the effects of maternal employment on young children. For a discussion of effects on adolescents, see Zaslow et al. in this volume.

Somewhat surprisingly perhaps, there is a good deal of experimental evidence available about the

effects of maternal employment in the specific population of welfare mothers. The data come from a number of "welfare-to-work" studies that have randomly assigned groups of women on welfare to various treatments ranging from monetary incentives to work to mandates that women participate in work or training activities. Surveys of these experiments in Grogger et al. (2002), Morris et al. (2002), and Zaslow et al. (2002), conclude that treatments that increased household income were associated with small positive effects on cognitive and behavioral outcomes, while treatments that reduced maternal time with children without increasing household income were sometimes associated with negative effects. These welfare-to-work experiments are well executed, with careful attention to randomization and minimizing attrition. Thus, they provide compelling evidence for the subset of the population they examine.

However, there is good reason to believe that the effects of maternal employment might be substantially different in this population than in the population at large: On the one hand, higher maternal income due to employment may have larger positive effects in poor households, and on the other hand, the loss of maternal time may be more important in single-parent households than in households with a head and spouse. Thus, these studies illustrate the general point that experiments can yield credible evidence regarding very specific questions and sub-populations, but that it may be difficult to generalize from their results.

Many studies attempt to control for observable variables that may be correlated both with maternal employment and outcomes using non-experimental, observational data, with mixed results. In one of the earliest studies of this issue, Leibowitz (1977) uses data from 1969, and finds no effect of maternal employment on PPVT. More recently, many studies examine data from the children of the National Longitudinal Survey of Youth (NLSY).[2] Baydar and Brooks-Gunn (1991) find that maternal

[2]The National Longitudinal Survey of Youth (NLSY79) began with a sample of 12,652 individuals who were aged 14-21 in 1979. They have been surveyed regularly since then, and beginning in 1986 the children of female sample members have been given a battery of

employment during the first year of life has negative effects on a test of vocabulary (Peabody Picture Vocabulary, or PPVT) and on behavior problems. They find no effect of working in the second or third year of life. Desai, Chase-Lansdale, and Michael (1989) find that maternal employment in the second and third years has positive effects that largely offset negative effects of employment in the first year. Belsky and Eggebeen (1991) found that children whose mothers were employed full time in the first two years of life had poorer scores on a test of behavior problems. Greenstein (1995) again uses NLSY data and finds no significant effect on PPVT, while Harvey (1999) reports that there are initially negative effects on PPVT and PIAT (Peabody Individual Achievement Tests in Mathematics and Reading) scores which dissipate with age. Mott (1991) reports that maternal employment of more than 20 hours per week in the second quarter of a child's life had a negative effect on PPVT scores at age 3 and 4, but that employment in the first quarter of life was not related to test scores. Parcel and Menaghan find positive effects of maternal employment during the first year on PPVT scores. Two studies of more disadvantaged children (welfare mothers and low income children, respectively) also find evidence of positive effects on test scores (Moore and Driscoll, 1997; Vandell and Ramanan, 1992) which is consistent with the experimental evidence cited above for poor women. Han et al. (2001) look at 7 and 8 year old children and find that the negative effects of maternal employment in the first year appear to persist for white children, but not for black children.

In a thorough exploration of the approach of dealing with selection by controlling for a wide range of observables, Ruhm (2000) shows how estimates of the effects of maternal employment change, when increasing numbers of covariates are added to the Ordinary Least Squares Regression model. The correlation between test scores and maternal employment is initially positive. Controlling for a "standard" set of covariates such as mother's age, education and race generally reduces this correlation to

---

developmental assessments every other year. In addition, mothers are asked a series of questions about the home environment and home inputs to child development as well as about child care.

zero, though there is still a positive effect of maternal employment in the second and third years of life on PPVT scores. Adding more variables to the model, starting with variables such as AFQT scores (the Armed Forces Qualifications Test, a test of job skills), mother's background at age 14, mother's marital status, and household poverty status in the quarter before the child's birth, religion, and foreign language use causes the effects to become negative (though not generally statistically significant). Adding measures of the mother's previous and subsequent employment probabilities causes some estimated effects to become significantly negative. In further analyses of this last specification, Ruhm finds that maternal employment is associated with negative effects only in households with the father present.

Variables measuring the mother's past and future employment may be viewed as a way to capture the unobserved characteristics of the mother that are correlated with employment in the child's early years. On the other hand, past, present, and future employment are likely to be highly correlated, which complicates the interpretation of each single coefficient. Perhaps the safest conclusion to be drawn from this suite of studies is that since the results are extremely sensitive to which observable variables are included in the models, they may well be sensitive to the presence of omitted, unmeasured variables.

A few recent papers employ a fixed effects strategy to examine the effects of maternal employment, controlling for unobservable variables. James-Burdumy (2002) and Waldfogel et al. (2002) again use NLSY data. Waldfogel et al. report that in OLS models maternal employment in the first year has negative effects on some outcomes/age groups among white children, though not among Hispanic or black children. The fixed effects models yield similar point estimates, but larger standard errors, so that they cannot reject the hypothesis that the OLS and fixed effects estimates are the same. On the other hand, they also cannot reject the hypothesis that there is no effect of maternal employment in the fixed effects models. Similarly, Neidell (2002) finds little evidence of an effect of maternal employment in household fixed effects models. As discussed above, fixed effects estimates may be biased towards zero if there is measurement error in test scores, and often rely on relatively small subsets of the data (which

helps to explain the increase in standard errors when researchers move to a fixed effects design).

Several studies address the endogeneity of maternal employment using instrumental variables methods. As instruments, Blau and Grossberg (1992) use fitted values from a two-limit Tobit model of maternal employment. That is, their instruments are non-linear functions of the included exogenous variables. Hence, the identifying assumption is that the covariates enter the second stage model linearly, rather than non-linearly. Perhaps unsurprisingly, this procedure yields large standard errors in the second stage, so that the hypothesis that the instrumental variables and OLS estimates are the same cannot be rejected. The OLS estimates are similar to those of many of the studies discussed above, in that they find negative effects of maternal employment during the first year on PPVT, but positive effects of maternal employment in subsequent years.

James-Burdumy (2002) and Baum (2003) use local labor market conditions as an instrument for whether or not a woman worked after her child was born. As discussed above, the identifying assumption underlying this specification is that local labor market conditions affect child outcomes only through the mother's probability of employment, and not, for example, through effects on the husband's income. These estimates again have fairly large standard errors, with the result that none of the estimated effects are statistically significant.

Baum also employs a crude form of "matching" by conducting an analysis that is restricted only to women who worked in the quarter before the birth. His rationale is that women who were working in the quarter before they gave birth are all relatively strongly attached to the labor market, and so maybe more homogeneous than other groups of women. In this sub-sample, he finds some significant negative effects of maternal employment in the first year on child well-being, which, however, are offset by increases in family income. For example, the effect of maternal employment on child PIAT-mathematics scores would be neutral in a household in which the mother's employment added $40,000 to household income.

Hill, Waldfogel, Brooks-Gunn and Han (2003) use propensity scores and find small negative effects of maternal employment in the first year on PPVT and PIAT scores. Interestingly, these effects are present for whites, blacks, and Hispanics and are stronger for married women in higher-income households. This suggests that it may be more difficult to find an adequate substitute for high-skilled women's time spent in child care.

What can we learn from all of this? First, crude estimates of the relationship between maternal employment and child outcomes are contaminated by selection bias. Mothers who are employed have characteristics that would cause their children to do well in any case. Controlling for observed and unobserved characteristics in a variety of ways leads to estimates of maternal employment effects that may be negative for employment in the first year of the child's life, but are generally negligible thereafter. Effects may be more negative for children of high socio-economic status mothers. For disadvantaged children, there is relatively strong evidence that maternal employment may even be beneficial, as long as it raises family income.

These conclusions can be drawn because there are a large number of studies that report generally similar findings using different methods and identifying assumptions. A key point is that given the limitations inherent in each particular method, we should avoid drawing strong conclusions from estimates based on a single method, and place more weight on conclusions that can be arrived at from a number of different directions.

3. Assessing the Assessments: Child Care Quality and Child Outcomes

The increase in labor force participation among mothers has also increased the attention paid to the quality of child care. The National Institutes for Child Health and Development (NICHD) Early Child Care Study found that most infants were placed in some sort of non-maternal care by four months of age (NICHD Early Childcare Research Network, 1997). Studies of the effects of the inputs on child care

quality, the effects of the inputs on child development, and the effects of quality on child outcomes are reviewed in National Research Council (forthcoming) and National Research Council and Institute of Medicine (2000b), Love et al. (1996), and Lamb (1998). Many studies use small non-randomly selected convenience samples, few or no measures of family and child characteristics, and few studies attempt to deal with the issue of selection into child care arrangements. Some do not use any type of control group. If children who have other advantages also get better child care (as Meyers et al. (2002) show) then it will come as no surprise if children in better quality care have better outcomes. However, is difficult to attribute the better outcomes to the causal effects of child care quality. This section makes no attempt to offer a comprehensive survey of the voluminous literature on child care quality, instead focusing on studies that are of particular interest.

There has been at least one experimental evaluation of child care quality. The National Day Care Study (NDCS; Ruopp et al., 1979) closely monitored a sample of 1,600 children in 64 day care centers serving low-income children for about nine months. The children were given baseline developmental assessments and were assessed again at the end of the nine month period during which classroom activities and inputs were monitored. The study design included two experiments in which children were randomly assigned to classrooms with different staff-child ratios and teachers with different levels of training. They found that preschool children whose teachers had training in early childhood education made greater gains on tests of language receptivity and general knowledge and showed more cooperative behavior than other children. Staff-child ratio was not associated with child development for preschoolers, but was for toddlers (ages 1-2). As has been discussed previously, one of the main limitations of experiments is generalizability. While this study suggests that care of higher quality benefits low-income children, it is not clear that any inference can be drawn about effects of quality on other children.

Evidence from many experimental studies of early intervention programs for disadvantaged

children is also relevant. Currie (2001) and Currie and Blau (forthcoming) provide surveys of this literature. Briefly, evidence from experimental interventions such as the Perry Preschool Project, the Carolina Abcedarian project, and the Infant Health and Development Program, as well as current evaluations of Early Head Start (Head Start from children zero to three years old) all show that quality preschool programs can have positive effects on disadvantaged children, though they do not identify which "inputs" are critical to producing quality.

Several large-scale, observational studies have attempted to examine child care quality in more representative groups of children. The Cost, Quality, and Outcomes Study (CQOS Study Team, 1999; Helburn, 1995) collected data from a sample of 400 day care centers in four states in 1993. Observational measures of quality were recorded, along with rich data on inputs and costs. Children who were expected to spend another full year at one of the sampled centers and then enroll in Kindergarten in Fall 1994 were selected to be given developmental assessments. They were reassessed in Kindergarten and second grade. The sample included 828 children, of whom 757 provided useable data. Controlling for maternal education, child gender and ethnicity, and the teacher's rating of her relationship with the child, Peisner-Feinberg et al. (2001) report a positive association between child care quality at age 4 and subsequent mental development, math achievement, and behavior. These are suggestive findings, but the absence of information on the home environment and a baseline development assessment leaves considerable uncertainty about whether these findings represent causal effects.

Mocan et al. (1995) used the CQOS data to estimate a model of classroom quality (a composite of ECERS-ITERS and other measures, where ITERS is the Infant-Toddler Environment Rating Scale) as a function of child care inputs. This study includes a larger number of control variables than many others. They found positive effects of staff-child ratios, teacher wages, and the fraction of the staff with a college degree on quality, and a negative effect of teacher turnover. However, Blau (2000) re-analyzed these data using a center fixed effects approach in order to control for other unobserved fixed characteristics of the

centers that might be related to child outcomes. Within centers, he found that only the effect of workshop training for teachers was significantly different from zero. Blau (1997) used data from the National Child Care Staffing Study (NCCSS) to do a similar analysis of the effects of inputs on classroom quality, with similar results. These studies illustrate that the fixed effects methodology can be used to control for a variety of potentially omitted variables, although for reasons discussed above, these estimates may be lower bounds on the true effects.

The NICHD Study of Early Child Care (SECC; U.S. Department of Health and Human Services, 1998) has followed a sample of over 1,300 children from their birth in 1991 through the present, closely monitoring their home and child care environments and their development. The study used hospital birth records in ten sites in the U.S. during 1991 to select a sample of healthy births to English-speaking mothers over age 18 who planned to remain in the site during the next year. Families were visited periodically for assessments of the home environment, and children who were in non-maternal child care arrangements were visited in their child care arrangement. The quality of the arrangement was measured using a variety of assessment instruments, and data on child care inputs were recorded by direct observation. A novel feature of the study was the inclusion and assessment of all types of non-maternal child care arrangements, not just centers and family day care homes. Child development was assessed at regular intervals and extensive psycho-social data on the mother and data on the home environment were collected as well. As children changed child care arrangements, the new arrangements were visited and observed.

The effects of child care quality on the cognitive and social development of children and on child behavior have been analyzed in a large number of studies using these data by the NICHD Early Child Care Research Network (NICHD ECCRN 1998, 2000, 2000b). These studies find generally positive effects on cognitive measures, behavior problems, peer interactions, and child care quality. Most of these

studies attempt to control for selection by controlling for some subset of observable variables in Ordinary Least Squares models.

The results from these studies are more credible than most in the literature because of the longitudinal design of the NICHD study, the inclusion of children in all types of child care (in some but not all of the studies), and the availability of extensive information on non-child-care factors. However, the richness of the data has not been fully exploited in most of the studies. For example, baseline measures of outcomes are seldom included and most studies exclude children who were not in child care at the time of observation. This could lead to biased estimates if such children are different from the included children in unobserved dimensions.

A recent analysis of the data by ECCRN and Duncan (2002) makes an effort to overcome some of these problems. This study controls for more home and child characteristics than the other studies using these data, and also estimates models of changes in test scores. Focusing on changes in test scores is akin to estimating models with child-specific fixed effects, because factors that affected the base score are implicitly controlled. The results indicate that a two-standard-deviation (SD) improvement in child care quality in early childhood is associated with a one-fifth of a SD increase in cognitive functioning at age 54 months in a standard regression model with extensive controls; and one-sixth to one-seventh of a SD increase in cognitive functioning in a change score model that controls for the level of 24-month cognitive functioning.

There have been very few studies of the effects of child care quality which have attempted to further control for the possibility of non-random selection on unobservable characteristics. Blau (1999) uses data from the NLSY to analyze the effects of child care inputs on child development in models that controlled for a large number of family and child characteristics as well as characteristics of the. His results show mostly small and insignificant effects of “structural” measures of child care quality such as group size, staff-child ratios, and teacher training, both in OLS and family fixed effects analyses. In

contrast, measures of the home environment were all significantly different from zero, and measured in terms of elasticities, their effects were three to five times larger than those of any of the child care effects.

I am not aware of studies that use either instrumental variables to deal with the selection problem in studies of the effects of child care quality. One possibility for instrumental variables analysis, would be to use changes in child care regulations as instruments for observed changes in child care inputs. The Florida Child Care Quality Improvement Study (FCCQIS; Howes et al., 1998) does something akin to this, by exploiting changes in day care center regulations that occurred in Florida in 1992. Teachers and children in a stratified random sample of 150 day care centers in four Florida counties were interviewed and assessed before and after the new regulations went into effect. The study found that the regulations appeared to "bind" (i.e. be enforced) but that there was no significant change in classroom quality as measured by the ECERS and ITERS. In terms of child outcomes, the only significant finding was that attachment security increased. Although it is striking that this change occurred coincident with the change in regulation, the absence of any control group makes it difficult to assess. Ideally, one would like to compare the changes in affected child care centers with changes in similar centers that were not affected by the regulations. It is also possible that changes in regulation affect the pool of children using centers, so that the gains in attachment security could reflect compositional, rather than causal, effects.

Hill, Waldfogel and Brooks-Gunn (2002) use propensity scores to get at the issue that children who participated in the Infant Health and Development Program might otherwise have used care of different types. Specifically, in the absence of the IHDP, they could have used maternal care, non-maternal home-based care, or center-based care. They find that the effects of the intervention were largest for children who would not have experienced center-based care in the absence of the intervention, and smallest for those who would have been in center-based care in any case.

This brief overview suggests that the literature on the effects of child care quality is somewhat under-developed relative to the literature on effects of maternal employment, because, although there are

many studies, comparatively few pay any attention to the selection issue. Since the quality of child care is a choice that is likely to be correlated with many other characteristics of children and families, it will be necessary to conduct studies that take selection into account before strong conclusions can be drawn about the effects of child care quality on child development more generally, though the experimental evidence certainly suggests that high quality care may be beneficial for low income children.

4. Conclusions

At this point, it is appropriate to return the question posed in the title, "How do we know what we think we know?" The answer is familiar to all scientists: We can be reasonably confident of our results if they can be replicated in a wide range of well designed studies. In the social sciences, well designed studies should attempt to deal with the ubiquitous problem of sample selection. Studies that compare results using a number of different methodologies and/or data sets, are also more informative.

References

Ainsworth, M.D.S., Blehar, M.D., Waters, E., and Wall, S. (1978) *Patterns of Attachment: A Psychological Study of the Strange Situation* Hillsdale NJ: Lawrence Erlbaum Associates.

Bayder, N. and Brooks-Gunn, J. (1991) Effects of Maternal Employment and Child Care Arrangements in Infancy on Preschoolers' Cognitive and Behavioral Outcomes: Evidence From the Children of the NLSY. *Developmental Psychology 27*, 918-931.

Baum, C.L. (2003) Does Early Maternal Employment HarmChild Development? An Analysis of the Potential Benefits of Leave Taking. *Journal of Labor Economics 21*, 409-448.

Belsky, J. and Eggebeen, D. (1991) Early and Extensive Maternal Employment/Child Care and 4-6 Year Olds Socioemotional Development: Children of the National Longitudinal Survey of Youth. *Journal of Marriage and the Family 53*, 1083-1099.

Bianchi, S. M. (2000) Maternal Employment and Time with Children: Dramatic Change or Surprising Continuity? *Demography 37*, 401-414.

Blau, D.M. (1997) The Production of Quality in Child Care Centers. *Journal of Human Resources 32*, 354-387.

Blau, D.M. (1999) The Effect of Child Care Characteristics on Child Development. *Journal of Human Resources 34*, 786-822.

Blau, D.M. (2000) *The Production of Quality in Child Care Centers: Another Look*. Applied Developmental Science 4, 136-148.

Blau, F. D. and Grossbert, A.J. (1992) Maternal Labor Supply and Children's Cognitive Development. *Review of Economics and Statistics 74*, 474-481.

Bowlby, J. (1969) *Attachment and Loss* New York: Basic Books.

Bound, J., Jaeger, D.A., Baker, R.M. (1995) Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Exogenous Explanatory Variable is Weak. *Journal of the American Statistical Association 90*, 443-450.

Cost, Quality, and Child Outcomes Study Team (1999) *The Children on the Cost, Quality, and Outcomes Study Go to School, Executive Summary* [web page] URL www.fpg.unc.edu/~NCEDL/PAGES/eqes.htm.

Currie, J. (2001) Early Childhood Intervention Programs: What Do We Know? *Journal of Economic Perspectives 15*, 213-238.

Currie, J. and Blau, D. (forthcoming) Who's Minding the Kids?: Preschool, Day Care, and After School Care in *Handbook of Education Economics* Finis Welch and Eric Hanushek (eds). New York: North Holland.

Currie, J. and Thomas, D. (1995) Does Head Start Make a Difference? *American Economic Review 85,* 341-364.

Currie, J. and Thomas, D. (1999) Does Head Start Help Hispanic Children? *Journal of Public Economics 74*, 235-262.

Desai, S., Chase-Lansdale, P.L., and Michael, R. (1989) Mother or Market? Effects of Maternal Employment on Cognitive Development of Four Year Old Children. *Demography 26,* 545-561.

Garces, E., Thomas, D. and Currie, J. (2002) Longer-Term Effects of Head Start. *American Economic Review 92,* 999-1012.

Greenstein, T.N. (1995) Are the 'Most Advantaged' Children Truly Disadvantaged by Early Maternal Empoyment? Effects on Child Cognitive Outcomes. *Journal of Family Issues 16*, 149-69.

Grogger, J., Karoly, L.A., and Klerman, J.A. (2002) *Consequences of Welfare Reform: A Research Synthesis* Santa Monica, CA: RAND.

Han, W., Waldfogel, J. and Brooks-Gunn, J. (2001) The Effects of Maternal Employment on Children of the National Longitudinal Survey of Youth. *Developmental Psychology 35*, 445-459.

Harvey, E. (1999) Short-Term and Long-Term Effects of Early Parental Employment on Children of the National Longituidnal Survey of Youth. *Developmental Psychology 35*, 445-459.

Heckman, J.J. (1979) *Sample Selection Bias as a Specification Error*. Econometrica 47, 153-161.

Heckman, J.J. and Smith, J.A. (1995) Assessing the Case for Social Experiments. *Journal of Economic Perspectives IX,* 85-110.

Heckman, J.J., Hohmann, N. and Smith, J.A. (2000) Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment. *The Quarterly Journal of Economics 115,* 651-694.

Helburn, S.W. (Ed) (1995) *Cost, Quality, and Child Outcomes in Child Care Centers Technical Report* Denver CO: Center for Research in Economic and Social Policy, University of Colorado at Denver.

Hill, Jennifer, Jane Waldfogel, and Jeanne Brooks-Gunn (2002) Differential Effects of High Quality Child Care. *Journal of Policy Analysis and Management* 21 (4), 601-627.

Hill, Jennifer, Jane Waldfogel, Jeanne Brooks-Gunn, and Wenjui Han (2003) Towards a Better Estimate of Causal Links in Child Policy: The Case of Maternal Employment an Child Outcomes. Unpublished paper, Columbia University School of International and Public Affairs.

Hoffman, L., and Youngblade, L. (1999) *Mothers at Work: Effects on Children's Well Being* New York: Cambridge University Press.

Howes, C., Galinsky, E., Shinn, M., Gulcur, L., Clements, M., Sibley, A., Abbott-Shim, M., McCarthy, J. (1998) *The Florida Child Care Quality Improvement Study* New York: Families and Work Institute.

James-Burdumy, S. (1999) *The Effect of Maternal Labor Force Participation on Child Development* Washington D.C.: Mathematica Policy Research.

Johnston, J. and DiNardo, J. (1997) *Econometric Methods 4th Edition* New York: McGraw Hill.

Katz, L.F., Kling, J., and Leibman, J.B. (2001) Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment. *The Quarterly Journal of Economics 116*, 607-654.

Leibowitz, A. (1977) Parental Inputs and Children's Achievement. *Journal of Human Resources 12*, 242-251.

Meyers, M., D. Rosenbaum, D., Ruhm, C., Waldfogel, J (2002) *Inequality in Early Childhood Education and Care: What Do We Know?* Columbia University School of Social Work, xerox.

Mocan, N., Burchinal, M., Morris, J.R., Helburn, S.(1995) Models of Quality in Center Child Care, in *Cost, Quality, and Child Outcomes*, S. Helburn (ed) Denver: Center for Research on Economic and Social Policy, University of Colorado at Denver.

Moore, K.A. and Driscoll, A.K. (1997) Low-Wage Maternal Employment and Outcomes for Children: A Study. *The Future of Children 7*, 122-127.

Morris, P. Know, V., and Gennetian, L.A. (2002) *Welfare Policies Matter for Children and Youth: Lessons for TANF Reauthorization* Manpower Demonstration Research Corporation [web page] URL http://www.mdrc.org/Reports2002/NG_PolicyBrief/NG_PolicyBrief.htm[2002].

Mott, F.L. (1991) Developmental Effects of Infant Care: The Mediating Role of Gender and Health. *Journal of Social Issues 47*, 139-158.

Parcel, T.J. and Menaghan, E.G. (1994) *Parents' Jobs and Children's Lives* New York: Wlater de Gruyter.

National Research Council and Institutes of Medicine (2000) *From Neurons to Neighborhoods: The Science of Early Childhood Development* J. Shonkoff and D. Phillips (eds) Washington D.C.: National Academy Press.

National Research Council and Institute of Medicine (forthcoming) *Working Families in the United States: Challenges, Opportunities and Options* E. Smolensky and J.A. Gootman (eds.) Washington D.C.: National Academy Press.

Neidell, M. (2002) *Early Time Investments in Children's Human Capital Development: Effects of Time in the First Year on Cognitive and Non-Cognitive Outcomes* Chicago: Dept. of Economics, University of Chicago.

NICHD Early Child Care Research Network (1997) Child Care During the First Year of Life *Merrill-Palmer Quarterly 43*, 340-360.

NICHD Early Chid Care Research Network (1998) Early Child Care and Self-Control, Compliance, and Problem Behavior at Twenty-Four and Thiry-Six Months *Child Development 69*, 1145-1170.

NICHD Early Child Care Research Network (2000) The Relation of Child Care to Cognitive and Language Development *Child Development 71*, 960-980.

NICHD Early Child Care Research Network (2000) Characteristics and Quality of Child Care for Toddlers and Preschoolers *Applied Developmental Science 4*, 116-135.

NICHD Early Child Care Research Network and Duncan, G.J. (2002) *Modeling the Impacts of Child Care Quality on Children's Preschool Cognitive Development* presented at Society for Research on Child Development, Minneapolis, April 2001.

Peisner-Feinberg, E.S., Burchinal, M.R., Clifford, R.M., Culkin, M.L., Howes, C., Kagan, S.L. and Yazejian, N. (2001) The Relation of Preschool Child-Care Quality to Children's Cognitive and Social Development Trajectories Through Second Grade *Child Development 72*, 1534-1553.

Powers, E. (2001) New Estimates of the Impact of Child Disability on Maternal Employment *American Economic Review 91*, 135-140.

Rosenbaum, P.R. and Rubin, D.B. (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects *Biometrika 70*, 41-55.

Ruopp, R., Travers, J., Glantz, F., and Coelen, C. (1979) *Children at the Center* Cambridge: ABT Books.

Ruhm, C. (2000) *Parental Employment and Child Cognitive Development* Working Paper #7666, Cambridge MA: NBER.

Staiger, D. and Stock, J. (1997) Instrumental Variables Regression with Weak Instruments. *Econometrica 65*, 557-587.

U.S. Bureau of Labor Statistics (2001) *Report on the American Workforce* Washington: Bureau of Labor Statistics.

U.S. Dept. of Health and Human Services (1998) *The NICHD Study of Early Child Care* Washington: National Institute of Child Health and Human Development.

Vandell, D.L. and Ramanan, J. (1992) Effects of Early and Recent Maternal Employment on Children from Low-Income Families. *Child Development 63*, 938-949.

Waldfogel, J, Han, W.J. and Brooks-Gunn, J. Early Maternal Employment and Child Cognitive Development. *Demography 39*, 369-392.

Zaslow, M.J., Moore, K.A., Brooks, J.L., Toot, K., Redd, Z.A. and Emig, C.A. (2002) Experimental Studies of Weflare Reform and Children. *Future of Children 23*, 79-98.

Zaslow, M.J. and Emig, C.A. (1997) When Low-Income Mothers Go to Work: Implications for Children. *Future of Children 7*, 1001-115.