

Selection and Attrition in the NICHD Childcare Study's Analyses of the Impacts of Childcare Quality on Child Outcomes

Greg J. Duncan

Christina Gibson

Northwestern University

July 14, 2000

Abstract

We review much of the literature coming out of the NICHD study on the impacts of childcare quality. We comment on one major issue that arises in this literature – selection into childcare - and two more that should be addressed - sample design and attrition and benefit/cost considerations. Although understanding childcare quality impacts is a crucial policy issue, we conclude that empirical estimates of these impacts are far from definitive. We applaud the careful measurement in the NICHD study, but hope that these data can be used to provide more convincing analyses of the important quality issue. We also include an appendix with reactions to earlier drafts of our paper from Belsky, Besharov, Burchinal, Friedman and Layzer, as well as our responses to them.

Contact information:

Greg J. Duncan
Institute for Policy Research
Northwestern University
2046 Sheridan Road
Evanston, IL 60208
847-467-1503
847-491-9916 (fax)
greg-duncan@northwestern.edu

We would like to thank Mark Appelbaum, Aletha Huston and Katherine Magnuson for comments on a prior draft, but in no way hold them responsible for remaining errors.

Selection and Attrition in the NICHD Childcare Study's Analyses of the Impacts of Childcare Quality on Child Outcomes

We consider the effects of selection and attrition in studies of the effects of childcare quality on child outcomes in the NICHD childcare study. We first discuss the issue of biases from difficult-to-measure factors affecting selection of children into childcare settings, and describe how these issues are treated in existing studies and might be treated in future studies. We also document the more than 50% unplanned sample attrition, and provide ideas for how one might investigate and adjust for attrition-related bias problems. Earlier drafts of this paper produced a number of written reactions. In an appendix, we include a number of these comments as well as our reactions to them.

The “selection” problem

All studies assessing the effect of a chosen child “context” (e.g., childcare, neighborhood, school) risk bias from unmeasured characteristics of the child and/or parent that affect both the selection of that context and child outcomes.

There are large disciplinary differences in how seriously this bias is viewed and treated in study designs and analyses. Thoughtful economists attach high value to thick description of the family and institutional forces surrounding childcare choices and outcomes, but most would judge as inadequate existing NICHD-study-based studies of quality effects that attempt to address the problem with the addition of a handful of family characteristics such as maternal education, family income and maternal psychological characteristics.

Economists’ worries stem in part from the fact that the potential “endogeneity problem” is probably more severe with childcare than with any child context other than the peer group. 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 out of the reach of many, but the range of choices remains broad for most families.

The potential childcare choice set for a given family expands further if one considers the options available to families that might move into the neighborhood of a relative, decide to work less to provide care themselves, be willing to travel substantial distances to get to a childcare provider, or allocate larger amounts of family income to pay for care. We may object that parents should not have to make these sacrifices for the sake of obtaining quality childcare, but the key fact is that the arrangements observed in nonexperimental studies such as NICHD’s reflect these kinds of decisions. Some parents in the study will be making such unusual sacrifices for the sake of their children; the difficult selection problem is to somehow adjust for what distinguishes them from other parents.

The most obvious version of the “selection bias” story is that parents who make moderate to extreme sacrifices to obtain quality childcare for their children are probably also promoting their child’s development in other ways. If this exceptional concern for their children’s

development is not adequately captured by the “selection” variables included in a model, then it will impart an upward bias to the apparent effect of childcare quality. A similar upward-bias story can result from a negative pattern of correlations from parents who, perhaps due to mental health problems, are both unwilling or unable to arrange for good quality care for children and less able to promote their children’s healthy development in other ways. Failure to adequately model either type of parental factors will impart an upward bias to the quality estimates.

Another form of potential (upward) bias arises from the child’s own characteristics. Suppose a child has a difficult temperament and bites or fights in a childcare situation. Higher-quality childcare providers may expel such children from their centers and force parents to choose among lower-quality providers. If, as seems likely, a problematic temperament also affects adversely the school readiness or other child outcome of interest, then failure to control for child temperament will also induce an upward bias in the quality estimates: the ill-tempered child has not done well AND has not enjoyed high-quality care, but the association between quality of care and outcomes is spurious, not causal.

Although most people who speculate about omitted-variable problems believe that they impart upward bias to the coefficients of quality measures, there are good reasons, as well as some evidence, to suspect the opposite. Suppose, as most modern developmental theories allege, that parents take agentic actions on behalf of their children. For example, a difficult-to-measure developmental delay in early childhood might motivate a parent to seek out unusually high-quality care to address the problem. Failure to adjust for child characteristics prior to entry in this case will impart a downward bias to the quality estimates; the delayed child may be flourishing in his or her high-quality care environment, but still not doing better than observationally (as measured in the data) identical peers in lower-quality environments. As with the temperament example, the analytic need here is to include controls for child characteristics and/or outcomes prior to entry into the childcare setting.

To round out the picture, Susan Mayer has presented a downward bias scenario involving parental rather than child characteristics. Suppose that parents choose between: i) a two-earner strategy involving high-quality care to compensate for hours parents are spending at work rather than with the child; or ii) a part-time-work-for-one-of-two-earners strategy involving less expensive, lower-quality formal care combined with more high-quality parental time. If children develop equally well in these two regimes, then failure to control for parental employment strategies will make it seem that childcare quality doesn’t matter.

Thus, the nasty specter of selection problems: they can arise from either parent or child-based characteristics and can impart either upward or downward bias to the coefficients of quality measures. Simple correlations between quality and outcomes tell us practically nothing about the nature of “true” effects given the uncertain direction of bias, and attempts to control for some but not all of them will prevent researchers from obtaining the bottom-line impact estimates that policy-makers seek.

Incidentally, yet another source of bias - downward in this case - comes from the nature of most of the quality measures. In pursuing construct validity, study PIs have opted for intensive measurement of a very small sample of time slices of childcare setting. The ORCE samples four

44-minute cycles during two half-day sessions scheduled two weeks apart. The quality of the observation during those time slices is very high. But there is always the danger that the center was having “one of those days” on either or both of the sampled occasions. Assuming the selected slices are reasonably random (it is not clear in study documentation whether they are or whether providers were given plenty of advance warning and/or choice in when the observations were taken), one can gauge the likely sampling error associated with the fact that ORCE samples from time and adjust (upward in this case) the coefficients on the quality measures.

The NICHD Study’s approaches to the selection problem

Apart from randomly assigning families to different childcare quality settings (or, second best, perhaps random assignment to radically different childcare subsidy schemes), there are no truly convincing solutions to the endogeneity problem. Nor does it help much to summarize results from large numbers of studies, all of which suffer from possible endogeneity bias.

At its heart, the endogeneity (selection) problem stems from omitted-variable bias. Accordingly, and playing to developmentalists’ natural inclinations, one could try to measure well and include in the analysis all of the relevant factors that have the potential for affecting selection of childcare mode and quality as well as child outcomes.

As revealed piecemeal across the papers, the NICHD study contains an impressive set of selection variables for the mother/family (income, education, mother’s PPVT receptive vocabulary score, CESD depression, the NEO personality inventory, the HOME learning subscale, maternal “sensitivity”); child (temperament, birth weight); early mother/child interaction (attachment security); and site dummies. NICHD papers have controlled for different combinations of the variables. We offer the following observations on how these papers treat the selection bias problem:

- Studies that control for just two or three variables (e.g., mother’s education and/or income/needs, maternal “sensitivity”) and then claim to have controlled for selection are not credible. A few variables cannot capture the bulk of what theories (both developmental and economic-choice) suggest might affect selection and child well being. Surely potent biases remain after these kinds of controls are introduced. More recent NICHD-data-based analyses control for a more complete set of selection factors, but none control for as many as half of the total available set.
- Child-based selection factors are rarely considered in these models. An exception is that “Early Child Care and Self-Control, Compliance and Problem Behavior at 24 and 36 Months” (1998) included controls for early temperament. If measured prior to enrollment in childcare, these and all other theoretically driven child-based selection factors should be routine controls in all of these studies.
- We found the often-substantial differences between the uncontrolled and selection-controlled quality effect size estimates to be interesting, and would urge that this comparison be included in all of the papers. One could view these differences in two radically different ways. The optimist says: “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 says: “The adjustments indicate that selection factors are obviously important. If measured selection factors can make this much difference, think of how much more of a change would come from further adjustments!” Altonji et al. (2000) use the size of the adjustments when including observable selection factors to bound the likely total amount of bias. Regrettably for the NICHD study, the paper sides with the pessimists: the more the adjustment caused by the introduction of observable selection factors, the more you should worry about unobservables.

- The 1999 SRCD paper “Effect Sizes from the NICHD Study of Early Child Care” claims that true parameter estimates probably lie “between zero-order correlations, which are undercontrolled, and the partial correlations, which may be overcontrolled.” We do not understand why partial correlations, based in this case only on adjustments for site, income-to-needs, mother’s adjustment, home quality, gender, percentage of time in center care and hours in childcare, probably overcontrol for selection factors.
- For the endogeneity-obsessed skeptic, a particularly worrisome dimension of childcare quality is stability. It would seem crucial to distinguish two aspects of stability: turnover in center staff, a center going out of business or other events that are largely outside the control of the family, on the one hand, and more voluntary changes in chosen arrangements that result from conscious choice or other actions (e.g., didn’t pay for care on time, child misbehaved) that are a function of family or child characteristics, on the other. The former type of stability produces fewer endogeneity biases, the latter is probably loaded with them.

Alternative approaches to the selection problem

What might be done?

Our ideas about how the problem might be approached with nonexperimental, longitudinal data are spelled out in greater detail in Duncan et al. (2000), which is available upon request. (Blau [1999] is also a good reference here; his NLSY-based quality measures may not be the best, but the nature of and motivation for his methods are worth careful study.)

As mentioned above, one helpful descriptive step is to be more systematic about the nature of selection on the “observables.” Here we have in mind using developmental and choice-based theories to group parental and child selection factors and run models that introduce various sets of these selection factors to gauge the direction and size of change in the quality coefficients. This would not solve the problem, but would help inform future work.

Another note here on disciplinary culture differences. We don’t understand why selection variables are kept to a minimum; entered in only linear and additive ways; and, sometimes, dropped when they prove insignificant. Economists typically include a very long list of intercorrelated selection variables. Why is this a bad thing? If there were evidence that standard errors were blowing up under the weight of the multicollinearity, then we could see a reason for parsimony. Since that does not appear to be a problem in these data, additional controls should

only serve to reduce bias, which is highly desirable. One possible objection to this advice is that missing data on certain selection measures (e.g., mother's PPVT score) reduces sample sizes needlessly. But this problem can be overcome by including all cases as well as a dummy variable indicator of missing data. Thus, we would recommend retaining all selection controls in all analyses.

And, in the interest of flexible controls, why not allow for noteworthy nonlinear effects of selection variables by using sets of dummies for selection variables? The data provide enough observations to support these extra variables. Although it would be nice to have single coefficients that summarize the effects of individual selection variables, the more important goal is to reduce bias, and so readers should care less about coefficients on selection measures than about making sure that selection bias has been controlled for in flexible ways.

Change models?

Economists' models are often derided for their simplistic assumptions. But their great virtue is that they make the theory very explicit, and provide a framework for understanding the parameter estimates coming out of the regression analysis. Developmentalists certainly do not shy away from theory. Why don't they attempt to formalize their theories in (even crude) mathematical models? Such models would certainly help us understand how to interpret the "lagged and concurrent" models presented in some of the NICHD papers.

A simple model we have in mind views a child outcome at, say, age 24 months as an additive function of the quality and quantity of home environmental inputs between birth and 24 months; the quality and quantity of childcare inputs between birth and 24 months; child-specific factors such as genetic endowment; and family-specific factors such as persistent parenting styles. Attempts to estimate the relationship between childcare quality and child outcomes are biased if important but unmeasured child- or family-specific factors affect both childcare quality and child outcomes. (This is just restating the omitted-variables problem in the previous section.)

To motivate the use of longitudinal data going up to, say, 36 months, suppose that the model just suggested holds for 36 months as well, with the inputs now measured over the interval from birth to 36 months. Thus, the 36-month model is: child outcome at 36 months is an additive function of the quality and quantity of home environmental inputs between birth and 36 months; the quality and quantity of childcare inputs between birth and 36 months; child-specific factors such as genetic endowment; and family-specific factors such as persistent parenting styles.

Now consider the results when we estimate a change model by subtracting the 24-month model from the 36-month model. It has as its dependent variable the change in outcome from 24 to 36 months. Independent variables are the quality and quantity of home environmental inputs between 24 and 36 months and the quality and quantity of childcare inputs between 24 and 36 months. Persistent (both measurable and not) child-specific and family-specific factors are differenced out of this change model and thus cause no bias – a huge analytic benefit.

We do not understand why the longitudinal data are not used to estimate change models, which would constitute a more powerful approach to controlling for bias stemming from persistent selection factors. Change models have their problems (many of which are overstated in the older developmental literature – we like Allison [1990] on this) - but they will almost certainly produce less biased coefficients on the key quality variables. Change-model-based coefficients are typically less precise (i.e., bigger standard errors) and fail to control for time-varying selection factors. (Inconsequentially, they also have lower explained variance.) But the virtues of their likely reductions in bias should reign as supreme in the minds of developmentalists as they do in the minds of economists: bias is (nearly) everything.

In some cases, the time series of measures of child outcomes in a given domain are not precisely comparable. While this is unfortunate, we would argue that it is far better to estimate a variant of change model with controls for the somewhat noncomparable, lagged dependent variables on the right-hand side of the equation.

Blau (1999) uses a variant of the model we have in mind to estimate the effects of the crude set of quality measures that are available in the NLSY. He uses both change and sibling fixed effects. His quality measures are certainly not the best, but measurement error is a much more tractable analytic problem than endogeneity bias, since one can use information on the nature of the measurement error to adjust estimates (usually upward) for it. Thus, rather than dismiss data with greater measurement error out of hand, one should attempt to learn from them by adjusting for the likely impact of that error while at the same time exploiting features of the data (e.g., representative, contains siblings) that are superior to data from the NICHD study.

Sampling and sample attrition

Many aspects of sampling and attrition in the NICHD study worry us greatly, and make us wonder why adjustments for nonresponse are not incorporated routinely into research based on these data. Don't get us wrong: we value highly the NICHD study's attempt to draw a population-based sample and the clarity and completeness of the description of the sampling and nonresponse in some of the NICHD-based study papers. It is far superior to most studies in this regard.

However, although the multi-site design of the study is focused on modeling rather than population description needs, there are still a number of ways in which sampling design and attrition problems can hurt even modeling efforts.

The most obvious example of the dangers of attrition-related bias is in using the data for descriptive purposes. If the assortment of planned (e.g., elimination of cases in dangerous neighborhoods and of women who indicate that they are likely to move in the next three years) and unplanned (e.g., refusal, inability to contact; haphazard selection of hospitals within sites) reasons for sample composition leads to an underrepresentation of low-income families, then description of, say, the distribution of the sample across differential levels of childcare quality will be biased in favor of higher-SES families.

More subtle biases may occur in modeling efforts. Suppose that child outcomes are quite sensitive to childcare quality for low-SES but not high-SES children. A model relating outcomes to quality that does not explicitly introduce such an interaction will produce a kind of sample average estimate of the impact of quality on outcomes. If the sample underrepresents low-SES families, then the sample average relationship will understate the relationship both for the population as a whole and, of course, for the low-SES portion of the population.

Naturally, one can always argue that a good analysis searches for all such important interactions. But since nonresponse bias may arise from many dimensions, and it is impractical to test for interactions across all of them, it is useful to have weights available to check for differences between weighted and unweighted model estimates. A properly specified model should produce very similar weighted and unweighted estimates of its parameters. (For details, see DuMouchel and Duncan, 1983).

Sampling and attrition in the NICHD study

Drawing from a conversation with Mark Appelbaum, the memo entitled “The Study of Early Child Care Enrollment Process,” as well as descriptions of the sample in found in three NICHD papers,¹ the situation seems to be as follows:

- Hospital selection within site was not random, with hospitals chosen by the individual study PIs based on criteria such as geographic location, patient load, and the relationship between the hospital administration and NICHD researchers. It is difficult to establish how representative births to these hospitals are, both of births in the cities in which they are located and of births more generally. This is not a problem of attrition but rather of a meaningful conception of the population of births that study births in these hospitals represent.
- 8,986 mothers were screened in the hospitals. This sample was reduced to 5,416 mothers eligible for a two week phone call because of unplanned (438 cases; mostly refusals) and planned (3,142 cases; mother <18, multiple births, mother not fluent in English, family expects to move, medical complications, baby being put up for adoption, family lives too far away, family in another study, family lives in an unsafe neighborhood) attrition.

¹ “Early Child Care and Self-Control, Compliance and Problem Behavior at 24 and 36 Months” (1997); “Familial Factors Associated with the Characteristics of Nonmaternal Care for Infants”(1998); and “Poor Children in Working Families: Parent Employment and Poverty as Predictors of Developmental Outcomes” (2000). It is worth noting that the sample sizes in these three papers are not consistent with each other, and in one case, seem suspect. In the "Poor Children in Working Families" paper, the follow-up sample size at 36 months is stated to be 1,216 children. The "Early Child Care and Self-Control" paper, however, reports a sample size of 1,041 at 36 months (the number used for response rates in this memo). Although the sample size discrepancy could be due to missing data on the various child outcomes, the papers do not report sample sizes per outcome measure. Also, the descriptive table presented in “Familial Factors Associated with the Characteristics of Nonmaternal Care for Infants” has a sample size of 1,281 at 15 months. It seems highly unlikely that over 21, months, only 65 (1281 - 1216) families dropped out.

We are troubled by some of these exclusion criteria. The study excluded families who were expected to move within three years; however, planning to move does not correlate well with actually moving, and efficacy of plans varies with socio-economic status. It also excluded families who lived in unsafe neighborhoods, but fails to indicate how unsafe neighborhoods were defined and whether definitions were uniform across sites. Furthermore, the nature of the sample necessarily excluded families who did not have a telephone. It is important to be able to assess the implications of these restrictions, which would require key demographic and other data on both response and nonresponse cases from hospital records that could be used to adjust for the forced attrition.

- A conditional sub-sampling plan was imposed to assist in eventually reaching a planned enrollment target of around 1,200 families. The subsampling attempted to ensure that single parent, low-maternal education and minority distributional targets were met. The NICHD study appears to have maintained random case selection at this stage of the process. However, the use of targets led to somewhat differential selection probabilities for enrolled cases, which could, in principle, be compensated by weighting by the inverse of these selection probabilities. Moreover, one site (Lawrence, KS) recruited an additional hospital to meet these quotas. All told, the subsampling process reduced the number of families for which two-week phone calls were attempted from 5,416 to 3,015. Apart from generating a need to weight for differential selection probabilities, this reduction should not be considered to be attrition-related nonresponse.
- Unplanned (1,153 cases; refusals and lack of success with contacts at three different times of the day), planned (151 cases; baby in hospital more than 7 days, planning to move within 3 years) and mysterious (185 cases, labeled “other” in documentation) reasons further reduced the sample from 3,015 screened mothers to the 1,364 who were scheduled to provide information for the one-month interview. *The 1,153 refusals and no-contact cases constitute the bulk of the study’s worrisome nonresponse cases.*
- Further attrition at the one-month interview itself reduced the 1,364 enrollment by 162 cases.
- Further attrition reduced the sample of families providing information for the 15-, 24-, 36-month and subsequent interviews.

Attrition-related nonresponse rates

It is useful to try to construct some response rates for the sample. For these response rates we want to count as attrition the “refusals” and “unable to locate” but not count as attrition the planned exclusions from the sample.

One way to do this is with the product of attrition-related response rates (the “Rs” below) at each stage of the sampling process. Here we ignore deliberate exclusions from the sample, as well as the 185 mysterious “other” cases lost when the one-month interviews were scheduled.

R1: Attempt to established eligibility for two-week phone call: $[8986 - (308 \text{ phone refusals} + 130 \text{ in-hospital refusals})] / 8986 = 95.1\%$

R2: Attempt to schedule one-month interview: $[3015 - (512 \text{ three unsuccessful attempts} + 641 \text{ refusals})] / 3015 = 61.8\%$

R3: One-month interview itself: $[1526 - 162 (\text{unclear origin})] / 1526 = 89.4\%$

R5: Fifteen-month follow up: $[1364 - 83 (\text{unclear origin})] / 1364 = 93.9\%$

R6: Two-year follow up: $[1281 - 196 (\text{unclear origin})] / 1281 = 84.7\%$

R7: Three-year follow up: $[1085 - 44 (\text{unclear origin})] / 1085 = 95.9\%$

Thus the attrition-related cumulative response rate as of the one-month interview is a product of R1, R2 and R3, which equals **52.5%** $[=(.951)(.618)(.894)]$.

The attrition-related cumulative response rate as of the three-year interview is the product of R1 through R7, or **40.0%** $[=(.951)(.618)(.894)(.939)(.847)(.959)]$

There is further nonresponse owing to the failure to observe childcare quality or obtain child outcomes measures. Again, there are two types of nonresponse: “purposeful,” stemming from children not in “eligible” childcare settings, and “genuine,” where measurement was desired but not obtained. The latter type constitutes another multiplicative factor needed to obtain a “true” response rate for a given analysis. The former may impart bias, but can be dealt with using Heckman-esque sample-selection bias adjustments. It is difficult to know the magnitude of this problem, as the NICHD materials do not present sample sizes per measured outcome.

Nonresponse rates, conservatively estimated, of 50% and higher are troubling and provide ample opportunity for attrition bias. A February 1993 Appelbaum memo provides reassuring comparisons of characteristics of catchment and enrollment samples. However, a comparison of children with follow-up data at 36 months versus nonparticipating children (in “The Relation of Child Care to Cognitive and Language Development”, in press) reveals very worrisome differences – a nearly one-point difference in mean income to needs (2.88 for respondents vs. 2.01 for nonrespondents); a more than one year difference in maternal schooling (14.4 vs. 13.2 years), a 1sd difference in maternal psychological adjustment (.59 vs. -.52 on a standardized scale); and, in what surely must be a typo, a nearly 3-to-1 higher rate of two-parent family structure (78% vs. 27%!).

Approaches to detecting and adjusting for nonresponse bias

The study is to be applauded for designing its procedures the study so that patterns of nonresponse can be quantified this precisely, which is much better than the typical developmental study where possible nonresponse problems are completely ignored. But because it provides this extra information, extra steps can be taken to ensure that the results are robust to at least some forms of nonresponse bias.

One way of doing this would be as follows: as with the response rate calculations listed above, distinguish families dropping out of the study by design (owing to planned exclusions from conditional sampling and other exclusion criteria) from the more worrisome dropouts owing to refusals, failure to locate, etc. Forget about the former and worry a lot about possible biases from the latter.

One method for handling nonresponse is to 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 an analysis of the complete sample of families not deliberately excluded from the study. Form a dichotomous variable equaling 1 if a case is response and 0 otherwise and then predict this dichotomous variable with a logit or probit regression using birth certificates and screening interview data. Obtain predicted values of the dichotomy for each response case and then use the inverse of those predicted values as the weight. Comparisons of weighted vs. unweighted models provide an indication of whether nonresponse bias is present. Other methods of modeling nonresponse (e.g., Rubin's propensity scores) could be used as well.

Not all of the papers are honest about attrition. For example, "Familial Factors Associated With Characteristics Of Nonmaternal Care For Infants" (1997) states that the attrition rate from 1 month to 15 months is only 6%, which is technically true, but misleading given the enormous amount of attrition that had already occurred before the 1-month interview.

Effects sizes and costs

A key policy question regarding childcare quality involves effect sizes and whether the benefits of increased quality are worth the costs. This is a hard question to answer since it is difficult to quantify costs and, particularly, benefits. Nevertheless, some steps could be taken in these papers to assist the green-eyeshaders amongst us.

Several of the more recent papers use the top vs. bottom quartile method for comparing effect sizes for quite different quality variables. This does not provide the reader with a good sense of how large the average quality difference is between the two groups. The papers state that the breakpoints for the bottom and top quartiles are roughly one-standard deviation below and above the mean – a two-standard deviation difference. But the average child in the bottom quality quartile is well below the 25th percentile, while 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 three standard deviation “treatment.” Viewed in terms of what you get with a three-standard-deviation “treatment,” the effects sizes are pretty small. Vandell and Wolfe (2000) provide some translation of NICHD-study effects sizes in more useful standard-deviation terms.

A very useful datum for these analyses is the average per-hour cost of childcare for children in the top and bottom quartile, plus, if available, some indication of the fraction of each group of children in subsidized care. Differences in (unsubsidized) hourly rates provide an order-of-magnitude indication of differences in the resources needed to provide those two levels of care. We hope that future work with these data will take on some of these effect size and cost issues. Again, Vandell and Wolfe (2000) provide a good discussion of some of these issues.

Appendix

Here we present comments on the “Selection and Attrition...” paper from Jay Belsky, Doug Besharov, Peg Burchinal, Jean Layzer, and Sarah Friedman, along with responses from Duncan and Gibson.

CAUSAL INFERENCE

Burchinal: From our perspectives as developmentalists, much of the controversy between developmentalists and economists regarding selection effects is due to fundamental differences in analytic traditions. Psychologists regard the experimental design with random assignment to treatment and control groups as the gold standard against which all other designs are compared (Kirk, 1982). Most psychologists believe that it is not possible to infer cause and effect from observational studies, with the caveat that overwhelming evidence across studies can be interpreted causally such as with studies linking smoking to cancer and lung disease. It is likely that random assignment designs are not possible in economics. Instead, some economists believe causality can be inferred from observational data when analysis models accurately reflect theoretical models. They argue that statistical methods can produce accurate descriptions of true relationships between predictor and outcome variables based on the assumption that the theory correctly delineates the complete prediction equation for the dependent variable.

Duncan/Gibson: We do not understand this point. To economists, it appears as though psychologists consistently disclaim aspirations toward causal inference when their articles use observational data, but then they go right ahead and use words like “influence,” “impact” and “affect” to describe their results, and often draw strong policy conclusions based on their analyses. The NICHD study-based literature is certainly no exception to this. What is the goal of the NICHD quality studies if not to draw policy conclusions regarding the effects of quality on children’s development? We could cite dozens of examples in the NICHD-study-based papers of causal statements. Policy conclusions require causal models and not mere correlations. Like economists, psychologists should aspire to causal inference in all of their empirical work; life is too short to do otherwise.

Economists also revere random experiments, particularly when the experimental condition approximates the policy intervention of interest, the experimental study sample approximates the population to which the policies will be applied, and neither take-up problems in the experimental group nor attrition from the experimental and control groups compromise the experimental impact estimates.

But when randomized experiments are not available, economists attempt in their data collection designs and analyses to approximate random assignment conditions. As detailed in Duncan et al. (2000; the “Endogeneity Problem” paper), this sometimes involves orienting one’s data collection around a “natural experiment” – a situation in which a legal or macro-structural change has produced variation in the policy variable of interest. Lacking a natural experiment, economists turn to econometric methods to best approximate experimental conditions in their analyses of observational data. As you point out in your paper “Family Selection and Child Care Experiences: Implications for Studies of Child Outcomes,” these methods (e.g., instrumental

variables, change models) sometimes sacrifice statistical power for bias reduction. This is certainly true, but illustrates the tradeoffs economists are willing to consider to approximate experimental conditions (e.g., eliminate bias in the estimate of the policy variables of interest) in their analyses of nonexperimental data.

SELECTION

Belsky: With respect to the issue of selection, we have a difference of opinion between developmentalists and economists. Developmentalists have been critical of our work for OVERcontrolling, using too many background variables. While we can argue forever about this, my real problem with the Duncan argument is that he wants us to control for stuff that could be affected by child care. We have done this somewhat, but it is very problematical. Let's take an analogy: smoking. I want to see the effects of smoking on health. I know you started smoking at age 18, so it makes all the sense in the world to control for physical health PRIOR TO THE ONSET of smoking so that we can then see, perhaps, how smoking affects, let's say, exercise behavior, having partialled out the effect of health prior to smoking on exercise behavior. But Greg would have us control for health AFTER the onset of smoking. But what if smoking affects health? Then we underestimate smoking's affect, potentially on exercise behavior, by controlling for THE EFFECT OF smoking ON health when we control for health measured after the onset of smoking; and since health can affect exercise, we can mis-estimate, actually underestimate, the effects of smoking on physical exercise if, in actuality: smoking-->health after the onset of smoking--->exercise behavior.

You have to recall that our study began with kids at birth. So most of our controls adjust for where families stood when life began, as this could not be affected by later child care usage. But Greg would have us control for stuff that comes AFTER child care usage because it is family stuff. But if this family stuff is itself affected by child care usage, then we are controlling away child care effects. That is, if child care---->family stuff---->child development, then by controlling for family stuff that is measured subsequent to child care usage one ends up under estimating effects of child care. Perhaps more importantly, I think there is good reason to believe, though I cannot share it now, that adding controls affects our results very little. In our work we have tried to balance arguments for adding more controls with those for including fewer. This invariably leaves everyone unhappy, developmentalists because we risk over controlling and economists because we risk undercontrolling. And Aristotle stressed the wisdom of the golden mean! Go figure.

Burchinal: Many psychologists and sociologists are concerned about whether control variables in regression or path analyses may serve as mediators. Within the child care literature, there is growing concern that “family selection factors” may mediate part of child care effects on child outcomes. For example, there is some evidence that a factor like maternal sensitivity to the child is influenced by the child care experience (NICHD ECCRN, 1999a). Child care experiences could influence parenting in at least two ways. The child in higher quality care may elicit more sensitive parenting because of enhanced cognitive or social skills acquired through child care experiences or the parent learn more sensitive caregiving practices from child care caregivers. Similarly, the child in lower quality care may be viewed as more difficult because of behaviors learned in the child care environment. If the direction of effects between some family selection

variables and child care quality is bidirectional, then use of those selection variables as covariates will result in underestimating the association between child care quality and child outcomes due to the overlapping variance.

Duncan/Gibson: We agree completely with the point that one should not control for selection factors that may have been affected by prior childcare experiences. We do not know about the timing of all of the NICHD study measures, so we may have included some measured too late in the children's lives to be safely considered as exogenous. They should NOT be included in the analysis, since controls for endogenous predictors may increase rather than reduce bias to the quality-of-care coefficients of interest.

Burchinal: I worry that the underlying belief the magnitude of the family selection problem is based on believing that parents are able to identify good quality child care. I think this belief is not well validated empirically. For example, the Cost, Quality, and Child Outcomes (CQO) study included assessments of parent's values and ratings of child care quality. Debby Cryer developed a parental version of the Early Childhood Environmental Rating Scale (ECERS) and the Infant-Toddler Environmental Rating Scale (ITERS). Parents were asked to rate how important each dimension rated on the ECERS or ITERS was to them. They were also asked to rate their child's classroom on multiple items assessing those dimensions. Results indicated that parents almost uniformly valued the dimensions of child care (e.g., health, safety, teacher-child interactions) very highly. The mean scores on each dimension were close to the maximum score (3; 1=not important at all, 2=somewhat important, 3=extremely important). In addition, almost all parents indicated they thought their child was receiving very good child care. The mean ratings of child care by the parents were above 6 (1=very poor quality, 7=excellent quality). However, the ratings by CQO staff indicated that most of the care was not as good as parents believed. The mean scores on the ECERS was about 4.25 and on the ITERS was about 3.25. Therefore, this suggests that parents may not be very good at knowing whether their child was receiving good quality care (Cryer & Burchinal, 1997).

Empirically, the magnitude of association between family characteristics and child care quality tend to be small to moderate. The moderate associations are more likely when care by relatives is included with care by unrelated adults. "Selection effects" in care by relatives, of course, also include other sources of variance such as common genetic and environmental influences.

The CQO study with over 700 children in 4 states in center child care (Peisner-Feinberg & Burchinal, 1997) reported modest correlations between quality measures in care by unrelated adults and maternal education ($r = .24$), parent's progressive attitudes about child rearing ($r = .22$), a measure of the home environment ($r = .15$), and ethnicity ($r = .06$).

The NICHD Study of Early Child Care has modest to moderate correlations with the 36m caregiver positive ratings of the home environment ($r = .20$ for care by nonrelatives and $r = .46$ for care by relatives), observed maternal sensitivity in interactions with child ($r = .26$ for care by nonrelatives and $r = .34$ for care by relatives), maternal education ($r = .19$ for care by nonrelatives and $r = .30$ for care by relatives), caregiving attitudes ($r = .12$ for care by nonrelatives and $r = .25$ for care by relatives) and family income ($r = .20$ for care by nonrelatives and $r = .27$ for care by relatives).

The correlations between child characteristics and observed care quality were smaller - maternal rating of child temperament ($r = -.01$ for care by nonrelatives and $r = -.08$ for care by relative), maternal ratings of behavior problems ($r = -.08$ for care by nonrelatives and $r = -.02$ for care by relatives).

Overall, these child and family characteristics accounted for 11% of the variance in the child care quality composite score in observations of nonrelative care and for 26% of the variance in observations of relative care.

I believe that we must consider family selection effects, but don't believe that we have sufficient evidence to suggest that the issue is a critical as stated in the critique.

Duncan/Gibson: There are many ways in which "selection" factors might bias the quality estimates, only some of which actually depend on parents making knowledgeable decisions. In our case of the ill-tempered child getting kicked out of care, the "selection" process is the center's reacting to the child and the parents have nothing to do with it. In the case where geographic constraints on availability limit high-end quality options, the "selection" process is constrained by community-level factors beyond (save residential mobility) the control of the parents. In short, selection bias can easily arise even when parents are ill-informed.

We are not surprised that there is little variation in how parents rate the quality of their child care since it is difficult for parents to admit that they might value cost or convenient location above quality. Similarly, we are not surprised that this does not correlate well with "true" assessments of quality, as we know that, parental ratings to the contrary, there is much variability in the actual quality of child care. In our view, the best way of addressing selection factors is to attempt to do so directly in the analysis.

Besharov: I think that the most likely direction of selection bias is to overstate the impact of quality, for all the reasons given in Greg's paper. (The reasons why it might understate the impact of quality are not as persuasive to me, and, even if valid, I think would be countered by selection effects in the other direction.)

I also think that it helps to take a step back in considering the issue. It helps to have a theory, not unexamined assumptions (which I shared until recently, by the way.) Let's ask the following question: What do we think should effect child outcomes--and how?

I start with the assumption that 25-50% of a child's future is determined at birth (call it heredity, or something more PC like temperament). Then I assume that parents (even those working full-time) are responsible for another 25-50% of child outcomes. That does not leave much for other influences. Assume that 30% (or even 40%) of the variance is caused by other factors--and then generously give child care half of that figure, say 20%.

Well, it seems to me that there would have to be tremendous VARIATION IN THE IMPACT of child care for it to be noticed. But I don't think anyone has ever documented variations large enough to effect this formulation. Hence, my view of "good enough" child care.

Duncan/Gibson: We are very hesitant to draw a priori conclusions about the direction of selection bias. Take the example of research on neighborhood influences on children's development. Conventional wisdom has it that nonexperimental studies that relate observed (e.g., census-based) neighborhood characteristics to child outcomes OVERSTATE the impact of neighborhoods. Why? Because unmeasured positive characteristics of parents probably increase the quality of chosen neighborhoods AND child outcomes.

But when Jens Ludwig investigated this issue using experimental data from the Baltimore Moving to Opportunity program, he found a negative selection story. The key factor driving residential mobility was MISbehavior on the part of children – parents are more likely to move in response to a child getting into trouble. A naïve regression approach would make it look like better neighborhoods are not associated with better kids. But the children in families that moved did indeed behave better in the new neighborhoods than they did in the old ones.

More to the childcare point, we found in New Hope, a randomized, anti-poverty program, a negative selection story for boys being placed into care – it appears as if mothers with misbehaving boys were most likely to put them into structured settings such as after-school programs. Since we can't rule out such processes at work in childcare choices, it seems unwise to prejudge the direction of bias.

Burchinal: As Duncan/Gibson point out, the number of covariates included in analysis models in the NICHD ECCRN papers has varied. The general strategy used in the first papers involved selecting as family selection factors the family and child factors that significantly correlated with both child care and child outcomes. Other “child” and “family” variables were also usually included in analysis. While the full battery of potential covariates has never been included, the influential covariates have been included in all analyses. Measures of maternal education, family income, or maternal IQ (PPVT) and measures quality of the home environment (HOME or rated maternal sensitivity) have been included in all papers, I believe. The other family measures do not add substantially to predicting our cognitive outcomes or maternal ratings of social outcomes or change meaningfully the estimated child care parameters in those models. Nevertheless, the ECCRN needs to understand that other covariates could be included and their inclusion is important to other disciplines: the issue has been discussed and is now understood. We plan to address this concern as we write new papers.

Duncan/Gibson: That is all we ask. And variables unimportant in simple analyses sometimes become more important in more complete models.

Burchinal: There seems to be differences between economists and statisticians regarding the issue of including correlated analysis variables. Statisticians (Mosteller & Tukey, 1977; Neder & Wasserman, 1981) warn about the impact on parameter estimates of including correlated predictors. The impact ranging from modest suppression to inestimable standard errors. While classic colinearity, which produces inestimable standard errors, is rarely a problem within child care research, it is not uncommon for faulty conclusions to be drawn from analyses that included correlated child care measures. For example, caregiver education and training tend to be highly correlated because most levels of training variables are tied to achievement of educational

degrees. Similarly, group size and ratio are clearly linked and are identical in many non-center child care settings. It is likely that analyses that predicted either child care quality or child outcomes from these four child care variables would obtain nonsignificant parameter estimates, even if the block of 4 variables accounted for a significant portion of the variance.

Duncan/Gibson: We can certainly imagine situations where selection variables and/or quality measures are too highly correlated to draw reliable inference about their separate effects. As Mark Appelbaum observed in comments about the multicollinearity issues: What God has brought together, let no man put asunder!

However, in the absence of evidence that multicollinearity is indeed blowing standard errors up to unacceptable levels, we believe that one should include as many exogenous selection factors as possible. We would also not worry much if only one of two highly correlated selection factors could be included, since the included one is doing most of the needed bias adjustment for the excluded one. However, one needs to point out that it is unclear which of the two correlated predictors is responsible for the effect of the included one.

Jean Layzer: I find myself in almost total agreement with Greg and his co-author. As you know, we persuaded the government that we should not look at child outcomes in the Low-Income Child Care Study, because of the selection bias problem. We have been criticized for this by Lynn Kagan and Bruce Fuller, whose own study does look at child outcomes and falls into the same old trap. I think Greg is exactly correct when he points out that even the adjustment made in NICHD underestimates the effect of the unmeasured variables.

Indeed, Lee Cronbach, as long ago as 1976, argued that you cannot use these adjustments to arrive at an unbiased estimate of the treatment effect. The paper "Analysis Of Covariance: Angel Of Salvation Or Temptress And Deluder?" was an Occasional Paper published by Stanford but undoubtedly published elsewhere and bears rereading on this point: "specification errors as well as errors of measurement have an attenuating effect... Since the parameters of the specification errors are unknown, no correction procedures can be counted on to provide an unbiased estimate of the treatment effect". His conclusion would match Greg's I think. "The solution is not to abandon realistic social science but to make less presumptuous claims regarding the result". Amen!

If you combine this warning about the inability to achieve a completely unbiased estimate of the treatment effect, with the question about the meaning of the size of the effect measured, you would be much more cautious in your statements and you wouldn't necessarily base policy decisions on these studies, as opposed to regarding them as adding some more information to the knowledge base.

Duncan/Gibson: We wish we had Cronbach's flair for titles.

Friedman: I read your critique of our publications and Jay's responses to you. The way I see things is that our study is a study by developmental psychologists (no economists, sociologists or demographers submitted applications in response to the RFA and I am a developmental psychologist myself). We have designed a study that is superior to what was out there in the

field when we started working. Our findings have been published in very prestigious developmental journals, which means we have met the high standards of peer review. I believe that we have enriched the scientific literature in very meaningful way and I think we have findings that policy makers and parents can trust.

At the same time, I am great believer in the importance of dialogue across disciplines and the enrichment of knowledge through such dialogue. So, I am very glad that you were interested in studying our papers in detail and that you are raising questions. But stopping with questions is not good enough. Neither is debating the issues without examining the data in different ways. So, in my opinion, the best approach would be for someone to analyze the data of Phase I using these analytical techniques. As you know, any researcher who wants the Phase I data set can get it by placing an application with RTI. They can find it in the study web site:
<http://public.rti.org/secc/>

Duncan/Gibson: We agree on the importance of enrichment through interdisciplinary dialogue. We tried to point out in our critique the considerable strengths, particularly construct validity of the quality measures, of the NICHD study. One way of viewing our critique is that it argues for co-equal weight being attached to the issues of external validity and endogeneity bias.

DIFFERENCES BETWEEN UNADJUSTED AND ADJUSTED EFFECT SIZES AS AN INDICATION OF LIKELY BIAS

Burchinal: In the interpretation of the difference between unadjusted and adjusted parameter estimates or effect sizes, I believe it is problematic to consider the difference between unadjusted and adjusted estimates as indicating the degree of bias. It seems likely that at least some of the difference could represent some of true covariance between quality and outcome that also covaries with selection factors. It might not, but it might—it is confounded so one simply can't tell.

Duncan/Gibson: The Altonji et al. (2000) result cited in our comments rests on some fairly strong assumptions – e.g., that included selection factors are a random subset of all possible selection factors. There are certainly a number of situations where these assumptions will not hold. The difference between unadjusted and adjusted quality estimates is a useful diagnostic rather than a definitive indicator.

ALTERNATIVE APPROACHES TO THE SELECTON BIAS PROBLEM

Burchinal: The alternative approaches provide viable methods for reducing parameter bias given that one assumes an additive model for predicting child outcomes. They, however, tend to have considerably less power to detect child care-child outcome associations. The instrumental variable approaches discards much of the true variance in observed quality measures when that variance is not associated with factors such as state regulations or availability of care. The fixed-effect regression approach provides a good alternative when the within-cluster correlation for repeated factor is small, but very poor power when it is large.

Overall, I believe that a primary difference between the analytic approaches adopted by economists and developmentalists involves the relative importance placed on bias and power. Economists seem to believe that they can make causal inferences from observational data if they use statistical methods that take into account omitted variable biases. I believe that their datasets often are large, so power has not been a substantial concern. Developmentalists are trained that only random assignment permit causal inferences. Parameter estimates are viewed as inherently biased when estimated from observational data. Efforts are made to identify potential confounds and consider them if possible. Their datasets often are small, so power has always been a substantial issue.

Duncan/Gibson: We agree that instrumental variables (IV) approaches are seldom satisfying and that the tradeoffs between bias and power are nasty. They arise in change models as well. The greater measurement error of a change-based dependent variable does not, under reasonable assumptions, impart bias, but it does reduce the precision (i.e., increase the standard errors) of the estimates. We could well imagine IV and change model situations where the standard errors are too large to draw any conclusions about effect sizes. But, provided standard errors are acceptably small, we would always opt for the procedure producing the least biased estimate.

More generally, we believe that durable knowledge of childcare quality effects comes from a number of different approaches converging on the same story. There is substantial value in the existing studies. Our goal in writing our critique was to try to get people to push the data further to see if the in-hand results were robust to what we perceive to be important potential threats to causal inference.

FIXED EFFECTS (CHANGE) MODELS

Belsky: Another issue involves what goes by the name of fixed-effects. Those in favor of fixed effects analyses say we should control for time 1 effects of child functioning when assessing child care effects on child functioning at time 2. This is EXACTLY the way to proceed if one's question is how does child care at a narrowly specified point in time between time 1 and time 2 affect CHANGE in child functioning from time 1 to time 2. But if your developmental question is not about change between time 1 and 2 because you are interested, at least first off, at examining the effects of cumulative child care history (up to time 1 or time 2), then by controlling for time 1 child functioning you fall into the same problem outlined above: You have controlled away effects of child care.

Ultimately, we are not talking about mutually exclusive strategies, but logically sequenced ones. To our group, the first question has been how is children's functioning at a certain point in time affected by child care experience up through THAT point in time. Once we get those issues sorted out, one can then proceed to say, hey, if child care through two years of age have these effects on two year olds, and child care through 3 years has these effects on 3 year olds, and child care through 4.5 years has these effects on 4.5 year olds (forthcoming soon), how does child care between 2 and 3 affect change in functioning between 2 and 3, and how does child care between 3 and 4.5 affect child functioning between 3 and 4.5. Personally, I don't think these subtle developmental-process issues are the pressing ones that parents and policy-makers want to

know about. That is why we have not followed this strategy. But open-minded people can honestly disagree on this issue. And one can certainly get on with the analyses that focus upon change after the analyses focusing upon the effects of cumulative child care history.

A last analogy: My first question about a mutual fund that I am thinking of investing in is probably, what is its average performance over the past 5 years, 10 years, and 20 years; and indeed this is the first kind of info that funds routinely supply to the would be investor. But a later, second order question might be something like, how does the fund perform, relative to its peers, during a bear market or a bull market? You see the distinction I am trying to make here? I think it would be wrong to say that if you first answer, or even only answer, the first set of questions regarding 5, 10 and 20 year performance you are doing things wrong, whereas if you supply info first or only on the second question (about bear and bull market performance) you are doing things right. Rather it is a case where the two sets of questions are asking for different information. Which should be offered up first and is most important? I suspect that most investment counselors would say the first set of info; they might even tell you that it is all you need to know about. Perhaps we are too much like a mutual fund! God knows we have plenty of investigators in our portfolio.

Duncan/Gibson: Like you, we value causal estimates of long-run developmental impacts more than estimates of short-run impacts. And, in our proposed change model, we hate to see the first 24 months of data "wasted" for control purposes rather than as part of a more comprehensive birth-to-school-entry look at quality impacts. But we view the 24- to 36-month change analysis as a more convincing (less biased) approach to the quality-outcome estimation problem than the approaches taken in existing studies. We should strive for unbiased long-run studies, but not ignore opportunities for more convincing short-run studies.

STABILITY OF CARE

Burchinal: Stability of care is an important dimension, but this study did not measure it well. Not counted were changes within a setting such as the changing of the teacher or switching classrooms. It also is highly correlated with amount of care ($r=.80$ in younger ages. I believe). Including both variables in the same analysis will result in failure to detect associations with either variable.

Duncan/Gibson: It is also important to keep in mind why the arrangement is not stable. If, for example, the instability of arrangements is due to child temperament, then this underscores our point that prior temperament needs to be taken into account and that failure to do so will impart an upward bias to the estimates.

SAMPLING AND ATTRITION

Burchinal: I believe that it is problematic to view the NICHD SECC sample as representative, even of the original catchment area. The refusal rates, when examined closely, were too high. Nevertheless, it provides a diverse sample for examining longitudinal issues of child care and

child development. Attrition has been nonrandom as well. We are in the process of computing multiple imputations of missing data. We hope to examine the reported associations in the published papers when those imputed data are available. To date, we have used simple imputation data to test whether reported results obtain when missing data is imputed. All regression results obtain. There are some minor differences that emerge when we compare quality of care across settings. In particular, I believe the quality of relative care is not as high when missing data are imputed.

It also seems that most samples examined in the child care literature are not representative. The NLSY is not a statistically representative sample of children in the US (it was a representative sample of the adults when recruited). The early child data from the NLSY was clearly biased toward low income, low education families. More recent waves of data are less biased, but substantial bias exists unless sample weights are considered in analysis. The National Household of Educational Survey is representative, but only when sampling weights are considered.

Duncan/Gibson: There are many instances where the NICHD study distributions (e.g., of centers meeting minimum quality standards, of simple correlations between mother's schooling and quality of care) are cited as estimates of something useful. Given the nature of sample selection and attrition, we have no idea what meaningful population these descriptive statistics refer to. (In contrast, regression-based estimates that control for dimensions of sample selection and attrition are indeed useful).

With a little work, one could come up with weights that would enable an analyst to make representative statements about catchment-area hospitals. It may also be possible to use the comparative characteristics of the sampled and nonsampled hospitals to place the sampling frame in a broader demographic perspective. The NLSY children are not a representative sample of children, but ARE, when weighted, a reasonably representative national sample of children born to mothers age 14-21 in 1979. That makes its descriptive statistics quite useful, since there is a known, if not ideal, population of inference. If the National Household Educational Survey is representative only when weighted, then all descriptive statistics based on those data should be weighted.

With regard to imputation of missing data, we assume that you mean missing data for cases for which partial information has been obtained. That can be very useful, although, as you know, tricky. Weighting for missing CASES can be thought of as another form of imputation, and, given the level and nature of attrition, probably as or more important to do.

Belsky: About attrition: I think Greg is on much firmer ground here. Our attrition is clearly selective and we lose those most at risk. Many in our group believe that this leads to us underestimating effects of quality, as those most at risk benefit the most from better quality care. I, personally, remain to be convinced, because my colleagues cling to this hypothesis whatever the data say.

Duncan/Gibson: It is hard for us to predict the likely direction of bias caused by the attrition patterns. You know that you lost more low- than high-SES families, but that could either increase or decrease the estimated impacts of quality depending on the nature of the

nonresponse. All the more reason to investigate with weighting or propensity scores - not a cure-all, but it would help.

EFFECT SIZES

Belsky: With regard to effects sizes, I think I have made this point before. What is more important from a public policy perspective--a small effect that affects huge numbers of children or a big effect that affects small numbers of children? Until the issue of effect sizes are contextualized by putting them in the context of the number of kids in child care year in and year out, so soon, if not already, our elementary schools will be such that virtually half of the kids in any school have been in nonmaternal care arrangements since their opening months of life, then I don't think talking about effect sizes and policy consequences make much sense. My analogy--I like these, as you can tell--is with inflation. What's 1% more inflation. Nothing, if you are speaking about buying a single big Mac, but across a trillion dollar economy, it is big bucks. Effects on individual children may be small, but effects on society may be something else entirely. I don't know, but neither do those who say, hey, these effects are small, nothing to worry about.

Burchinal: We believe that even “modest” effects sizes provide reasonable evidence of meaningful associations between child care quality and child outcomes based on developmental theory and problems with attenuation. It is unreasonable to expect that child care experiences will account for 5% or more of the variance in outcomes for very young children when we know that factors such as sensitivity in parenting often account for less than 10% of variance in analyses that also adjust for family demographics (for example see NICHD ECCRN, 1999b). In addition, it is unlikely that child care experiences will account for 5% of the variance in preschool outcomes as they did in the CQO analysis if extensive family selection measures had been collected and included in those analyses. Cohen (1988) defines as modest effect sizes either regression coefficients that correspond to partial correlations of $r \geq .10$ or to standardized differences between group means of $d \geq .30$. Use of effect sizes rather than statistical significance as evidence of associations between child care quality and child outcomes will be more robust indicators, especially when power to detect associations is lowered by practices such as using analytic methods that dramatically decrease reliability, include family measures with bidirectional effects, or include moderately to highly correlated measure of child care experiences.

Duncan/Gibson: On effect sizes, we agree completely that one cannot jump to policy conclusions from estimated effect sizes. But, unlike Burchinal, we do not regard Cohen-type considerations of “small” and “large” effect sizes, or evidence on fraction of variance explained, as crucial for the policy debate on childcare quality.

Economist argue that it is the quality impacts relative to the cost that matter here. Small effect sizes that are inexpensive to bring about may well be worth it, while big effects from very expensive interventions may not be. It is quite possible to have a cost-effective intervention for a policy variable that accounts for only a few percent of variance explained. The welfare-to-work literature is filled with examples of cost effective job search programs that have relatively small

impacts but cost very little. Thus, reliance on Cohen-type effect size rules, in the absence of cost considerations, appear misguided to us.

Burchinal: [Cost of care is an] important issue that we may be able to explore to some degree. Unfortunately, all we have parent's report and care provider's report regarding fees paid. There are substantial subsidies within the child care system that would be ignored.

Besharov: One comment on costs and effects: Sure a small social cost across many people may make a difference, although the legal concept of de minimus comes to mind. But an equally important point is that "quality" child care can be many TIMES more expensive than good enough care. The issue is not just the cost/benefit ratio--but also the opportunity cost of resource allocation. Would the world be better off with a much broader EITC or better high schools or . . . (you fill in the blank) or more Head Start for the middle class? I doubt it.

Duncan/Gibson: Good point, but one that is already taken into account in the cost-benefit framework. The benefit/cost ratio exceeds one only if the benefits of a given investment exceed the benefits of the best alternative use of the money.

Layzer: I wanted to say something about effect sizes, since we have been thinking about this in the context of family support. Tom Cook, under the influence of Greg D., pointed out a long time ago that a small effect size for some measures could have great policy importance -- for example an increase of 2% in income or a decrease in the number of children retained in grade. When we talk about small effects on children's cognitive or social-emotional development, we have to deal with what size effect is educationally significant, since that is what might have economic and thus policy significance. A small effect size (.2 of a standard deviation) translates into a difference of 3 points on a test with a mean of 100 and a standard deviation of 15 points. This doesn't seem very important, but it is a bit more complicated. If the kids are a homogeneous group, i.e., educationally subnormal, raising their scores 3 points might lift them above a threshold (even then I would question its significance, but there is room for argument) and thus might have policy significance.

Duncan/Gibson: The causal arrow of influence between Tom Cook and Greg D. runs from Cook to Duncan.

References

- Allison, Paul (1990) "Change Scores as Dependent Variables in Regression Analysis" in C.C. Clogg (ed.) Sociological Methodology, Oxford: Basil Blackwell, 93-114.
- Altoni, Joseph, Elder, Todd and Taber, Christopher (2000) "Selection on Observable and Unobservable Variables: Assessing the Effectiveness of Catholic Schools," mimeo, Northwestern University.
- Blau, David (1999). The Effects of Child Care Characteristics on Child Development, Journal of Human Resources, 34, 786-822.
- Cryer, D, & Burchinal, M. (1997). "Parents As Child Care Consumers. Early Childhood Research Quarterly, 12, 35-58.
- DuMouchel, William and Duncan, Greg J. (1983) "Using Sample Survey Weights to Compare Various Linear Regression Models" Journal of the American Statistical Association, 78, 535-543.
- Duncan, Greg, Magnuson, Katherine and Ludwig, Jens (2000) "The Endogeneity Problem in Developmental Studies," mimeo, Northwestern University.
- NICHD Early Child Care Research Network. (1996) "Characteristics of Infant Child Care: Factors Contributing to Positive Caregiving" Early Childhood Research Quarterly, 1, 269-306.
- NICHD Early Child Care Research Network. (1997) "Familial Factors Associated with Characteristics of Nonmaternal Care for Infants" Journal of Marriage and the Family, 59, 389-408.
- NICHD Early Child Care Network. (1998) "Early Child Care and Self-Control, Compliance and Problem Behavior at 24 and 36 Months" Child Development, 69, 1145-1170.
- NICHD Early Child Care Network. (1999a) "Child Outcomes When Child-Care Center Classes Meet Recommended Standards for Quality" American Journal of Public Health, 89, 1072-1077.
- NICHD Early Child Care Network. (1999b) "Effect Sizes from the NICHD Study of Early Childcare" Paper presented at the Biennial Meeting of the Society for Research in Child Development, April, Albuquerque, NM.
- NICHD Early Child Care Network. (2000) "Poor Children in Working Families: Parent Employment and Poverty as Predictors of Developmental Outcomes", mimeo.

NICHD Early Child Care Network (In press) “The Relation of Child Care to Cognitive and Language Development” Child Development

Peisner-Feinberg, E., & Burchinal, M. (1997). Concurrent relations between child care quality and child outcomes: The Study of Cost, Quality, and Outcomes in Child Care Center. Merrill-Palmer Quarterly, 43, 451-477.

Vandell, Deborah and Wolfe, Barbara (2000) Child Care Quality: Does It Matter and Does It Need to be Improved, Washington, D.C.: U.S. DHHS, Office of the Assistant Secretary for Planning and Evaluation.