



- How would a scaled-up program likely change the predicted effect size and costs?
- Do measurements of effects of scaled-up existing programs using non-experimental data provide similar results to experimental data, and if not, why not and which measurement provides a more accurate prediction?

Each section of the chart has a related section in this document. Each section of the document is designed to be relatively independent of the others so that readers can select the section from the chart they wish to read and study and use the PDF bookmarks to find the associated section. You can request hard copies of the chart in poster and fold out form by contacting Rena Subotnik (rsubotnik@apa.org)

Introducing multiple methods into RCTs is becoming a more frequent occurrence. However, a deep vein of literature that illustrates the process by which professionals design and utilize data for multiple methods does not exist. Typically, it takes almost 8-10 years from the inception of an RCT for a complete set of analyses to appear in journal publications. RCTs with multiple methods take even longer. So finding illustrative examples of RCTs using multiple methods requires looking at RCTs started in the late 1980s and early to mid-1990s. Fortunately, there were a few RCTs from this period that incorporated multiple methods, as well as some publications that illustrated the power of multiple methods RCTs.

In providing descriptive material associated with this chart, we draw primarily from three such RCTs. The first RCT is *Project STAR*, a class size reduction experiment from kindergarten to third grade conducted in Tennessee beginning in 1986. The second RCT is the *New Hope* intervention carried out in Milwaukee starting in 1994. New Hope aimed to lift individuals earning wages below the poverty level into higher paying, more stable jobs in such a way as to improve their life circumstances and that of their children. New Hope provided income supplements, health insurance coverage and paid day-care over a three year period. The third RCT, which began in 1994, is *Moving to Opportunity* (MTO). This program provided a chance for families living in public housing in very poor and risky neighborhoods to relocate to better neighborhoods. These three RCTs with multiple methods have long term follow-ups and a literature trail that is rich with examples of multiple methods within RCTs, as well as the use of such data to address the questions cited above. We provide brief descriptions of these RCTs in an appendix, while the bibliography to this document contains most of the publications spawned by these multiple methods RCTs.

Each of these RCTs with multiple methods was started between 1986 and 1994. Each of the experiments ran for 3 or 4 years and included long-term follow-ups after completion. Each also generated productive literature starting in the year or two after the end of the experiment, and continuing through 2008. We chose these three illustrative and multiple methods RCTs because they represent some of the best examples from the 1980s and 1990s, and were among the first to incorporate multiple methods into their design. However, each of these RCTs has flaws in design, implementation and analysis. Some of these flaws are inevitable parts of doing social experimentation. Others might be attributed to compressed time during planning and implementation or limited theoretical development and budgets. Hindsight is 20/20, yet reflecting on these flaws offers not only the opportunity to illustrate the complexity of designing

RCTs and using multiple methods in the real world, but also the opportunity to learn from and improve future multiple methods RCTs.

Finally, we have likely omitted articles, research and viewpoints that could have improved this description. We hope that this effort will be seen as a starting point for an expanded discussion of multiple methods RCT's in the research literature, and that a richer and more diverse set of perspectives will emerge in future discussions.

### **Background on the Use of RCTs and Multiple Methods**

There has been a long and messy debate in the social science community, and most recently in the educational research community, over the appropriate role and priority that should be given to experimentation with random selection in funding research and development (R&D). This debate is intertwined with at least two other long-running discussions. The first is between the role and priority of “qualitative vs. quantitative” evidence. The second debate is about how best to arrive at reliable predictions for large-scale social or educational programs. The latter argument involves two important questions: (1) whether and under what conditions results from non-experimental data can provide unbiased estimates of scaled-up effects of social and educational interventions/policies, and (2) whether and how smaller scale experimentation, which cannot be generalized outside its specific context, can contribute to making reliable predictions for large-scale programs. Turning small-scale interventions into large-scale interventions, or even implementing small-scale interventions in different contexts can be problematic because results from small-scale RCTs cannot be generalized beyond the specific experimental population and context. So using small-scale experimental programs as a way to identify and design efficient large-scale social and educational programs still remains an elusive goal. RCTs with multiple methods can help address these debates and critical issues.

These discussions have been intense in the educational research community because of recent strong funding support favoring experimentation. Intellectual leaders and researchers from several disciplines, the National Academy of Science and federal policymakers in the Department of Education encouraged the use of RCTs as a high priority in R&D funding beginning around 2003 (Borman, 2002; Boruch, 1997; Chalmers, 2003; Cook, 2002, 2003; Duncan & Magnuson, 2003; Feuer, Towne, & Shavelson, 2002; Mosteller & Boruch, 2002; Raudenbush, 2005; Shadish, Cook, & Campbell, 2002; Shavelson & Towne, 2002; Slavin, 2002; Towne, Shavelson, & Feuer, 2001). This movement generated heated deliberations centering on the role of experimental and non-experimental research and qualitative and quantitative evidence in R&D funding and in improving long term policies in education (see for instance, Eisenhart, 2005, 2006; Eisenhart & Towne, 2003; Howe, 1998, 2004; Maxwell, 2004). Angrist (2004) gives an interesting history of this line of argument and suggests a set of analytical techniques in evaluating RCTs that address some of inevitable flaws in design and execution. However, the debate between the utility of experimental and non-experimental evidence considerably pre-dates this most recent flare-up (see for instance, Chen & Rossi, 1983; Cook & Campbell, 1979; Cronbach & Shapiro, 1982; Heckman & Smith, 1995; Mosteller, 1995)

Multiple methods RCTs address many issues relevant to these debates. Multiple methods RCTs represent an evolution from “black box” experimentation -- whose only purpose is to measure the impact of a particular intervention -- into combining quantitative and qualitative research

methods that allow researchers to address a broader set of questions (see introduction above). This can considerably enhance the scientific and policy value of RCTs. Using multiple methods RCTs addresses the questions cited in the introduction above by:

- Incorporating methods of data collection, commonly referred to as “qualitative,” that become indispensable and powerful tools within a RCT for understanding why and how the effects (or lack of effects) of an intervention occur and why effects differ among participants.
- Using these qualitative data to explore and predict how results are sensitive to contextual effects, thereby improving predictions of effects in different and/or larger scale settings.
- Providing opportunities to compare and contrast experimental and non-experimental measurements and test hypotheses as to why such results differ, thereby potentially improving the methods and reliability of non-experimental analyses.
- Perhaps most important, focusing attention of the research not only on whether an intervention works, but why it works, thereby contributing to building more general theories that can improve predictions in all settings and better prioritize what future experimentation to fund.

Multiple methods RCTs are a partial response to a long recognized need for *theory driven* experimentation aimed not only at accurate measurements of interventions, but at accounting for why and how effects occur (Chen & Rossi, 1983; Cook, 2002; Cook & Campbell, 1979; Cronbach & Shapiro, 1982; Donaldson, 2007; Duncan & Magnuson, 2003; Heckman & Smith, 1995; Raudenbush, 2005; Romich, 2006; Walshe, 2007). Theory development is a critical complementary process to experimentation because successful theories can dramatically reduce the need for future experimentation and allow better priorities to be assigned to future experimentation. Without the parallel development of theories, the process of experimentation will not converge, but rather lead to choosing from an infinite number of possible experiments.

Multiple methods RCTs can also help address whether results of measurements using non-experimental data are reliable, why results may differ between experimental and non-experimental measurements and under what conditions non-experimental results are more reliable. For instance, contextual effects can explain differences between experimental and non-experimental measurements, and multiple methods within RCTs can expand the range of contextual factors that can be tested for their influence. In this and other ways, multiple methods RCTs can help sort and integrate the large body of non-experimental research with experimental research. In the end, scientific consensus requires that researchers explain and reconcile both experimental and non-experimental measurements. Multiple methods can not only help reconcile these measurements, but also improve the reliability of non-experimental measurements, thus helping to form scientific consensus.

Multiple methods RCTs may represent a significant advance that uses complementary approaches in the pursuit of scientific knowledge, similar to those described by Salomon (1991), by (1) helping to address persisting questions and arguments in the research community, (2) developing stronger social and educational theories, (3) reconciling experimental and non-experimental measurements, and, (4) enabling improved external validity and better predictions of social and educational policies. However, multiple methods RCTs will also have some

significant limitations, mostly due to their increased costs and complexity. More experience is needed to test whether their potential contributions can be realized. Thus, RCTs with multiple methods do not ensure termination of the methodological and R&D policy debates (Howe, 2004).

## **Box One- Motivating Policy/Research Questions**

### **What does theory suggest to be effective?**

The first question to undertake when thinking about conducting a multiple methods RCT is “What does theory suggest to be effective?” From the beginning, it is critical to think about the possible theories that might explain (1) why effects are expected, (2) why such effects might be different across participants, (3) whether results are sensitive to contextual factors, and, (4) how to design an experiment that provides the information needed to refine and improve the intervention and why and how such effects might change when scaled-up. Thinking through the alternate possible causal mechanisms and how and why such mechanisms might produce effects leads inevitably to the use of “multiple methods.”

In the present context, a theory can range from simple hypotheses to using much more complex sets of interacting causal mechanisms that provide an explanation of why and how the measured effects occur. Sometimes there is no direct link to a theory that might apply to a potential intervention. Rather, unique hypotheses or theories might need to be developed for specific interventions. Burton, Goodlad, and Croft (2006) and Tilley (2004) provide clear examples of the range of simpler hypotheses that might account for the results of their crime prevention experiments and why such hypotheses are critical in research. Romich (2006) suggests ways that social policy experiments can advance theory-based knowledge in child development. Kling, Liebman, and Katz (2007) explicitly state four hypotheses that might account for the mechanisms involved in neighborhoods producing impacts on adult labor market outcomes. Cohen, Raudenbush, and Ball (2003) present a theoretical approach to modeling the relationship between educational resources and achievement, focusing on a casual role for instruction.

Project STAR’s first publication suggested three ways that class size may affect achievement: enhancing teacher morale, improving the number and quality of student-teacher interactions or improving student engagement (Finn & Achilles, 1990). Evidence collected in the experiment included teacher surveys and logs and observational data suggesting greater 4<sup>th</sup> grade student engagement for those in smaller K-3 classes (Finn & Achilles, 1999). Multiple methods data collected from teacher aides also helped to address why teacher aides in larger classes did not have statistically significant effects over large classes with no aides (Finn & Achilles, 1999; Gerber, Finn, Achilles, & Boyd-Zaharias, 2001).

Investment in theories can have low payoff if the measurements that theories are developed to predict are not accurate. Perhaps the most important role of RCTs is to provide more accurate measurements that make development of increasingly refined theories productive (Heckman & Smith, 1995). Project STAR’s compelling experimental evidence spawned a rich theoretical literature directed toward understanding the causative mechanisms that created these effects. Research studies spanning several disciplines suggested hypotheses about classroom processes

and parental effects that might account for achievement gains in small classes (Blatchford, 2003, 2005; Blatchford, Bassett, & Brown, 2005; Blatchford, Bassett, Goldstein, & Martin, 2003; Blatchford, Goldstein, & Mortimore, 1998; Blatchford & Martin, 1998; Bonesrønning, 2004; Boozer & Cacciola, 2001; Bosker, 1998; Bosworth & Caliendo, 2007; Datar & Mason, in press; Finn, Pannozzo, & Achilles, 2003; Grissmer, 1999; Hattie, 2005; Lazear, 2001; Webbink, 2005). This literature is one of the best examples of theory development to explain an experimental effect. Such a literature can help specify what additional RCTs might be pivotal in deciding between theories, as well as eliminating many areas of experimentation that would not make any contributions.

New Hope was partly built on a hypothesis that working at least 30 hours a week over a three-year period (if supplemented by additional health, income and child care benefits) could lift individuals who were not working or whose earnings were below the poverty line into lives of more stable employment and increased wages. These outcomes would then improve participants' lives and the lives of their children over a longer term (Duncan, Huston, & Weisner, 2007). The primary initial experimental measures focused on labor force behavior, and the results showed statistically significant effects for the treatment group. However, the researchers were initially puzzled by several issues. The control group participants made substantial employment and wage gains that did not depend on their having received New Hope benefits, and these gains were much larger than the incremental gains of New Hope recipients. Moreover, many eligible for New Hope benefits did not use them, or used them only sporadically. These results were inconsistent with the project's theoretical framework, and if not for multiple methods data, New Hope would have only left unanswered questions and small contributions to theory and policy.

However, the multiple methods data collections in New Hope enabled substantial contribution to understanding and designing new policies for welfare, childcare, health, education and employment to improve the lives of the working poor. A refocusing on wider outcome measures for working mothers with children enabled the development of theories that helped explain (1) the experimental results, (2) why this pattern of results emerged, (3) why benefit utilization was much lower than expected, (4) why New Hope made the difference for some, but not others, (5) why control group participants made such large gains, (6) why effects for boys more than girls were particularly large and sustained in achievement and behavior, (7) why more flexible menus of benefits might have enhanced effects, (8) why and what kinds of targeting would have improved efficiency, and, (9) what key contextual factors and other issues present in Wisconsin would need to be addressed in any large-scale statewide or national interventions (Duncan et al., 2007; Huston, Duncan, Granger, Bos, McLoyd, & Mistry, 2001; Yoshikawa, Weisner, & Lowe, 2006).

Besides its contribution to policy related to the working poor, New Hope serves as perhaps the best current model for designing, utilizing and documenting multiple methods in RCTs, and in illustrating that simple theories will be inadequate in predicting the complexity and often chaotic lives of this population. This contribution to research methodology and theory building may be its most important and longest lasting legacy. Multiple methods were used extensively in the following ways: (1) a comprehensive set of outcome measures using surveys and testing, (2) in-depth interviews with participants, their children and the children's teachers, and, (3) an ethnographic study of 44 families during and after the experiment. Moreover, the design,

analyses and documentation of New Hope multiple methods data extended beyond academic journals with separate documents designed for researchers and policy audiences. For example, Yoshikawa et al. (2006) provide a volume of studies using multiple methods data to address research questions, while Duncan et al. (2007) direct their work to both policy and research audiences.

Researchers pursued the MTO experiment because results from scores of non-experimental studies suggested that living in poor neighborhoods may adversely affect a wide range of adult and child outcomes. The results, however, elicited concern about the high correlation of neighborhood characteristics with individual, family and school characteristics, and the strong possibility of selectivity bias in non-experimental measurements. Different theories about neighborhood effects also predicted opposite directions for the outcomes. Kling et al. (2007) state the theoretical hypotheses as follows:

It is hard to judge from theory alone whether the externalities from having neighbors of higher socioeconomic status are predominately beneficial (based on social connections, positive role models, reduced exposure to violence, and more community resources), inconsequential (only family, influences, genetic endowments, individual human capital investments, and the broader non-neighborhood social environment matter), or adverse (based on competition from advantaged peers and discrimination). (p. 84 )

For instance, Wilson (1996) articulated a theory about why unemployment was high and wages were low for inner city residents and the effect on neighborhoods of changes in job opportunities within and close to inner cities. The MTO experiment was designed to test, among other things, whether changing neighborhoods caused changing adult labor market outcomes, and which theory better predicted the outcomes. And further, whether those outcomes improved, worsened, or did not change.

When the basic MTO experimental results showed no effect on adult labor market opportunities or children's achievement from moving to better neighborhoods, researchers utilized multiple methods data to explore the reason for the null labor market effects (Kling et al., 2007; Turney, Clampet-Lundquist, Edin, Kling, & Duncan, 2006). In the process, they discovered that large effects were registered on adult mental health measures and behavioral measures for children. Thus, the use of multiple methods data not only helped to explain null results on the original variables of primary interest, but helped to validate some theories, while dismissing others. In addition, multiple methods data enabled the identification of important outcomes not included in the original objectives. Clampet-Lundquist, Edin, Kling, and Duncan (2006) provide another example of using multiple methods data to test four explicit hypotheses that might account for the positive behavioral effects shown for boys, but not for girls, in MTO.

### **What is known from previous research?**

Reviewing previous research has always been a difficult and nuanced task because of the almost universal disparity in outcomes present in previous studies, without a theory available to explain why such differences are present. The critical question is how to distinguish among the often multitude of studies that could be relevant and how to synthesize these studies in the most

meaningful way. For example, a long running debate in education (mainly from the 1980s to early 2000s) addressed the effect of additional resources on educational outcomes. Three research studies utilized differing techniques to select and weigh the value of studies (including meta-analysis) to arrive at contrasting conclusions (Greenwald, Hedges, & Laine, 1996; Hanushek, 1997, 2002; Krueger, 2002, 2003).

A major motivating factor for RCTs and, in particular, multiple methods RCTs, is to move future literature reviews toward a scientific and/or policy consensus on a given question. The absence of consensus in previous non-experimental studies has been a major motivating factor in moving toward experimentation. However, black-box experimentation alone may not create either research or policy consensus due to the lack of generalizability of such experiments to different and larger scale settings, and the inevitable flaws present in most social and educational experimentation. Consensus will require being able to explain why the current set of both experimental and non-experimental measurements differ. Black-box experimentation alone usually fails to provide evidence for why experimental and non-experimental measurements are different, but multiple methods data can provide considerable help in addressing these differences.

Four important reasons why previous results from various studies exploring the same phenomena can differ are (1) methodological bias, (2) the presence of contextual effects, (3) differences in the characteristics of the population studied, and, (4) structural or other changes in programs/interventions during scale-up such that predictions from smaller scale programs have little predictive accuracy. Addressing potential bias requires a thorough knowledge of the strengths and weaknesses inherent in the various methodologies used in previous work. A relatively new strategy groups studies into the following categories: experimental, quasi-experimental, “natural” experiments and non-experimental. Webbink (2005) provides an example of this type of review. However, within each of these categories there is usually wide variation in quality. Thus, simple categorization can be misleading. Duncan and Gibson-Davis (2006) and Duncan, Magnuson, and Ludwig (2004) provide advice on how to critique and interpret non-experimental results. Cronbach and Shapiro (1982) and Heckman and Smith (1995) provide critiques of experimental studies. Cook, Shadish, and Wong (2005) compare and contrast experimental and quasi-experimental results. Rosenzweig and Wolpin (2000) and O’Connor (2003) both provide perspectives from developmental psychology and economics on the strengths and weaknesses inherent in “natural experiments.”

Multiple methods RCTs should be designed to address and settle issues that prevent the establishment of consensus in a literature review. For instance, multiple methods RCTs can provide evidence on contextual effects that help reconcile previous disparate results. It is also possible to design multiple methods RCTs that incorporate a non-experimental measurement. For example, although Project Star did not incorporate a non-experimental measurement component, two later studies chose non-experimental samples from Tennessee and compared and contrasted experimental and non-experimental measurements of particular outcomes. Krueger (1999) compared STAR experimental findings with results estimated non-experimentally from the variation in large class sizes that show similar experimental and non-experimental results. Using propensity scoring, Wilde and Hollister (2007) show significant differences comparing

experimental and non-experimental results with a sample of Tennessee students outside the STAR experiment.

### **Consider relevance for the populations of interest.**

Utilizing multiple methods in RCTs can be viewed as road testing a prototype intervention and be initiated in the planning and design stage in the form of a “mini-efficacy trial.” The purpose is partly to obtain feedback from the population of interest about their attitudes, reactions or predictions, as well as to suggest changes to a particular intervention. Another rationale for an efficacy trial would be to determine the groups that should be targeted for inclusion in a study.

Techniques such as focus groups, interviews and surveys of a sample of participants might be appropriate for the planning stage of a multiple methods RCT. Focus groups allow for exploring the appropriateness of the intervention, an array of participant reactions, potential new design features and more. Interviews can allow a two-way conversation focusing on these same topics. Surveys can be less expensive where larger samples are required, but they lack the flexibility for unstructured feedback. Brock, Doolittle, Fellerath, and Wiseman (1997) and Poglinco, Brash, and Granger (1998) provide an example of using multiple methods in an efficacy trial on a small population of potential participants in New Hope during the extensive pre-planning for the major study.

## **Box Two- Desirability/Feasibility of an RCT Study**

### **Desirability**

**Is intervention X well-enough developed to warrant a controlled study, or are efficacy studies needed first to clarify constructs and establish the basic efficacy of the proposed intervention?**

The development of a theory, the review of literature and the use of multiple methods in the planning and design stage provide information for making a decision whether to proceed first with a small scale efficacy trial or a larger, and more formally structured RCT. Since RCTs, especially multiple methods RCTs, are substantially more costly and require much more planning than efficacy trials, conducting efficacy trials prior to multiple methods RCTs is likely to become the rule rather than the exception. Efficacy trials not only establish viability for an intervention, and provide potential re-design and re-targeting insights to make it more effective, but also allow field testing for multiple methods data collections and eventual re-design. Multiple methods may be as important to efficacy trials as they are to structured RCTs (see section **Consider relevance for the populations of interest** in Box One above). For instance, Duncan et al. (2007, pp 23-26) describe their 50 participant pilot project and adjustments made in the later intervention as a result of the pilot. According to the researchers, the pilot project seemed crucial to understanding the population of interest and matching the program benefits to that population.

**Are the results generated likely to be worth the expense?**

This question can be viewed in two different contexts: the R&D or scientific context and the policy context. R&D or scientific questions address whether a particular intervention is a sound use of R&D funds. In this context, the normal scientific criteria used in peer reviews are relevant to evaluate proposed RCTs or other research approaches. An increasingly important question will be how a multiple methods RCT will contribute to testing a particular theory or set of theoretical hypotheses. Without multiple methods, it will often be difficult to provide a strong argument about why an RCT would contribute to theory. If the only outcome is the measurement of the intervention effect -- even if done with a sound design -- the contribution to theory will often be minimal. A sound design for a multiple methods RCT that addresses the questions stated in the introduction can significantly enhance the scientific value of a research proposal. In contrast, it may be increasingly difficult to make the scientific case for black-box RCTs due to their limited contributions to theory and inability to provide explanations for disparate results from previous studies.

With regard to the policy context, the researcher must determine the value of the intervention -- if successful -- to society. The set of questions that arise in this context are not only costs versus benefits to society that would result from a successfully scaled-up intervention, but also the chances that a successful small-scale intervention could be widely scaled-up without a significant deterioration of effects. These questions are more difficult to address with black-box RCTs than for multiple methods RCTs. The latter methods provide much more information about potential scale-up issues arising from contextual effects, how to target an intervention to the population showing larger effects and how to re-design the intervention to make it more effective. Again, black-box RCTs have less policy value compared to a feasible multiple methods RCT.

Perhaps more than any other single publication, Kling, Liebman and Katz, (2005) should be read by those researchers and policymakers who question the scientific and/or policy value of funding multiple methods in RCTs. We cite this article in the introduction, but the article goes on to elaborate in more detail how in-depth interviews influenced their research.

Our qualitative fieldwork had a profound impact on our MTO research. First it caused us to re-focus our quantitative data collection strategy on a substantially different set of outcomes. In particular, our original research design concentrated on the outcomes most familiar to labor economists: the earnings and job training patterns of MTO adults and the school experiences of their children. Our qualitative interviews led us to believe that MTO was producing substantial utility gains for treatment families, but primarily in domains such as safety and health that were not included in our original data collection plan. *In our subsequent quantitative work, we found the largest program effects in the domains suggested by the qualitative interviews* [italics added].

Second, our qualitative strategy led us to develop an overall conceptual framework for thinking about the mechanisms through which changes in outcomes due to moves out of high poverty areas might occur. *Our conversations with MTO mothers were dominated by their powerful descriptions of their fear that their children would become victims of violence if they remained in high poverty housing projects* [italics added]... This fear appeared to be having a significant impact on the overall sense of well-being of these mothers, and it was so deep-seated that their entire daily routine was focused on keeping

their children safe. .... We hypothesized that the need to live life on the watch may have broad implications for the future prospects of these families.

Third, our fieldwork has given us a deep understanding of the institutional details of the MTO program. This understanding has helped us to make judgments regarding the external validity of our MTO findings, particularly regarding the relevance of our results to the regular Section 8 program. *In addition, this understanding has prevented us from making some significant errors in interpreting our quantitative results* [italics added].

Fourth, *by listening to MTO families talk about their lives, we learned a series of lessons that have important implications for housing policy. For many of the things we learned, it is hard to imagine any other data collection strategy that would have led to these insights* [italics added]. (pp. 244-245)

## **Feasibility**

### **Are the factors of interest amenable to experimental manipulation and control in the real world?**

Can a particular intervention be successfully tested in an experimental framework? Social and educational experiments inevitably depart from ideal experimental conditions. These departures can sometimes seriously mitigate the scientific advantages that are inherent in ideal experiments. Gueron (2002), who had 30 years of experience at the Manpower Demonstration Corporation (which pioneered large scale social welfare experimentation), provides the best resource for understanding the complexity of actually doing an RCT. Her article covers most of the real-world constraints and limitations that can compromise the internal validity of such efforts. Gueron (2003, 2007) also provides unique perspectives on both the difficulty and the utility of doing social welfare experiments.

Heckman and Smith (1995) examine a group of social experiments that measured the impacts of job training programs conducted in the 1980s and early 1990s. One of their conclusions is that significant deviations from experimental conditions destroyed much of the scientific value of the results. These deviations were often peculiar to particular experiments, but their article identifies and characterizes many of the vulnerabilities inherent in social experiments. As a result, it is important to assess the susceptibility of a proposed intervention to the potential vulnerabilities that have plagued social experimentation such as those described by Heckman and Smith.

Another potential vulnerability in experimentation is that the “signal to noise ratio” will turn out to be too small for successful measurement. In any intervention, there is an unbiased effect size (i.e., the signal), which in the absence of any noise (i.e., bias, errors and uncertainties created by a finite sample), would emerge from an RCT. Yet, there are always conditions that will introduce “noise.” Some of this noise can be predicted and limited by the selection of sampling parameters. However, such an analysis can only take account of sampling uncertainty, and cannot fully take into account the inevitable other sources of noise from random sources and flaws present in social experiments.

In order to have successful RCTs, it is necessary for the signal to emerge clearly from the noise (a favorable signal to noise ratio). Since the amount of noise is always uncertain, a researcher would optimally like to create an intervention with a large signal. Two factors often control and limit the strength of the signal in an experiment. The first is that the costs of the experiment will usually increase with larger variations in the treatment. For example, the costs to the state of Tennessee for Project STAR was about \$13 million to sustain class size differences of 24 vs. 16 pupils per class over four years. The costs were nearly proportional to the size of the class size reduction. Although smaller reductions would cost much less, it's unlikely that a class size reduction of 2-4 students (a small signal) per class would have produced such definitive results.

The second factor limiting signal strength is that large variations can generate political problems arising from inequitable treatment of test and control participants (Gueron, 2002). As a result of such large differences in class sizes maintained over four years, parents of pupils assigned to large classes started lobbying for their children to participate in the experimental classes. Some parental opposition was mitigated by randomly dividing the large classes into two groups with teacher aides in one group, leaving a smaller group of children without benefit. However, about 15 percent of the children assigned to large classes appeared in small classes over the course of the experiment –likely due to parental pressure (Finn & Achilles, 1999). Thus, large signals may also cause some compromise in the integrity of the experiment.

As can be seen, it's important to consider potential reactions of participants in the control group. For instance, Duncan et al. (2007, pp 42-43) describe the reactions of some New Hope participants to being assigned to the control group. Randomization was clearly described to participants from the beginning of their potential involvement. The researchers explained that although some participants would not receive New Hope's additional benefits, no participants would lose any benefits as a result of the experiment. Nevertheless, some participants who "lost" the lottery were disgruntled and painted a negative picture of the program to the researchers.

### **Can an RCT be conducted without encountering ethical constraints?**

Any social or educational experiment will have to balance the potential benefits of carrying out the experiment with the possible costs to participants. This balancing is ultimately evaluated by Human Subjects Review Boards who have the independence and authority to protect research participants. However, because this balancing is often difficult and not straight-forward, studies involving any significant costs to participants or possible ethical issues need early review by such Boards.

Generally, the limitations on experiments imposed by ethical considerations and review boards would be predicted to cause some compromises from ideal experimental conditions. Gueron (2002) provides a real-world perspective on the ethical issues that arose in carrying out over two decades of social welfare experimentation, as well as some necessary compromises that sometimes limit internal and external validity. Often, these compromises can be addressed analytically in a way that still maintains the advantage inherent in random assignment. Angrist (2004) illustrates an analytical methodology that addresses non-compliance in treatment groups and crossovers from control groups. These techniques allow some flexibility in service denial or compelling participation without undue compromise in measuring and interpreting effects from random assignment experiments.

**Is it likely that the study would gain the necessary cooperation and enough recruits to be assigned randomly to treatment conditions?**

Project STAR was mandated by the Tennessee legislature. Schools with at least three kindergarten classes were invited to participate. (This eliminated smaller schools from consideration.) About 100 schools volunteered to partake in the randomization of children to classes, and 79 schools were selected to participate. Project STAR was conducted prior to the need for parental permission for children's involvement in research, so during the experiment virtually all students entering kindergarten in these schools took part, as did all students entering these schools in grades 1-3. Given the compressed time available for planning Project STAR, had the project been conducted more recently, it is possible that parental permission would have been a significant obstacle to carrying out the experiment. This likely would have introduced limitations on external validity and potential selectivity bias.

Unlike Project STAR, many experiments do not have a state mandate for participation. In assessing whether sufficient numbers of participants can be recruited, there are three main considerations. The first is whether sufficient volunteers can be obtained for the participation "lottery." The second is whether a sufficient number of lottery "winners," those assigned to the experimental condition, will actually utilize the benefits offered. The third consideration is how different (selective) the volunteers and those using benefits will be from the actual population of interest. Both MTO and New Hope encountered some difficulty not only in recruiting "volunteers" for the lottery, but also having participants (compliers) in the treatment group actually take advantage of the benefits offered.

Duncan et al. (2007, pp. 36-41) describe a year-long effort to recruit 1,357 participants to New Hope, and the subsequent effort to understand why some participants did not take full advantage of the project benefits. Although New Hope participants ended up approximately matching the racial/ethnic characteristics of a national sample of working poor individuals, participants' choice to take part likely made them somewhat different from those who did not volunteer based on other characteristics. Those who became eligible for New Hope benefits ended up using those benefits less than expected. Fortunately, the original sample was large enough (1,357) to allow a focus on working mothers (745) – the group that utilized benefits most often and accounted for much of the program effects. Because the subgroups of primary interest often emerge only after initial analysis, larger samples offer some insurance against this problem of lower than expected utilization.

MTO was carried out in 5 cities and recruited individuals from public housing in high poverty areas. Individuals who volunteered may have been more motivated to move out of public housing. Over a 4 year period, over 4,000 families volunteered for the program. Of those who were randomly assigned to the treatment, about 47 percent actually moved (Kling et al., 2007). Clearly, self-selection of volunteers into the lottery group and further selectivity in the treatment group has the potential to bias experimental results and limit generalizability. Heckman and Smith (1995) provide further examples of this kind of selection bias in job training experiments.

**Will funding be sufficient to support an RCT design with adequate statistical power?**

Multiple methods RCTs can be significantly more costly than “black box” experimentation. For instance, larger sample sizes are often required to measure whether effects differ across participants with different characteristics. Further, the additional data collection required in multiple methods RCTs can add significant costs. In the longer term, the benefits derived from multiple methods RCTs can be substantial because the derived theories can more efficiently guide future experimentation. In the short run, however, they will increase R&D costs. These additional costs may represent a significant barrier to proposing multiple methods - especially between black-box RCTs and multiple methods RCTs - in the absence of firm guidance from funding agencies about their priorities.

A “power” analysis is the usual method used to determine the number of participants required to measure different effect sizes with various degrees of certainty. It is always difficult to incorporate the wide range of possible factors that can introduce additional uncertainty and bias into any social experiment. Yet, failure to incorporate these factors can make experiments too weak to accurately measure desired effects. Often it is the case that experimental effects are widely and unpredictably different across participants, and the major contribution of the study is to examine what is causing the effects in a sub-population of interest. Sample sizes larger than those dictated by power analysis provide some assurance that such effects can be studied.

New Hope had a board of directors whose responsibility was partly to garner necessary funding to carry out the project. Although funding was obtained to initiate the program, additional funding for various research components was added during and after the experiment. In the end, over 50 different foundations, government agencies and businesses provided financial support for New Hope. The original funding to support a target population of about 1,200 allowed the experiment to start. However, additional funding supported a “family” study that incorporated multiple methods into the data collection. About two-thirds of the participants of the family study were individuals with children -- making the sample adequate for studying this population (Duncan et al., 2007). The later funding of this sample proved crucial in making the entire experiment so valuable. Multiple methods RCTs will be more likely than black box experimentation to provide “unexpected” results and/or discover unanticipated opportunities that would require added funding to exploit. In Project STAR, MTO and New Hope, later funding to refocus study objectives, explore hypotheses in more detail, or do longer term follow-ups was crucial to their scientific and policy utility.

Occasionally, funding can be generous. Project STAR was funded by the state of Tennessee as part of a compromise that delayed the institution of smaller class sizes statewide until the completion of the study (Ritter & Boruch, 1999). The \$13 million needed to fund the study was a small proportion of the costs of implementing a policy of smaller classes statewide. This experiment represented a compromise between implementing an expensive state-wide program and funding an experiment. In such circumstances, the sums needed for experimentation looked small to legislators, but large to researchers. This sum supported very large sample sizes (over 6,000 in the kindergarten cohort) and extensive multiple methods data collection. Project STAR encountered much more difficulty raising the smaller amounts of funding required for long-term follow-ups.

### **Will I know afterwards what conditions are necessary for the intervention to be effective?**

One of the major vulnerabilities of black-box experimentation is that it often provides little power for predicting effects in different contexts, generalizing to different populations or scaling up programs. One of the major advantages of multiple methods experimentation is an increased ability to estimate how impacts might change in different contexts, populations and scaled-up versions. Perhaps the major reason for doing multiple methods RCTs is the ubiquity of contextual effects in social and educational interventions, and the need to develop theories that can make better predictions that apply to different contexts and populations.

The California class size reduction initiative, which was partly motivated by Project STAR (and a huge one-time budget surplus in California), is being used as the poster child for the lack of predictability from contextual and scale-up effects. Project STAR was not a small scale experiment, but a fully scaled-up experiment involving 79 schools and over 12,000 children (Finn & Achilles, 1999). Much was learned from Project STAR that could guide policy and implementation in other states. Three important lessons were: (1) 3-4 years of small classes, starting with kindergarten, were needed for long-term effects, (2) effects were much larger for minority and disadvantaged children, and (3) the teachers in Project STAR were not newly recruited, but were drawn from the pool of existing experienced teachers.

California mandated class size reductions statewide moving from around 30 to 20 pupils per class in grades K-3 beginning in 1996 (Bohrnstedt & Stecher, 2002). The legislation passed one month before the start of school, along with strong monetary incentives for immediate implementation, and a set of rules governing implementation. These rules stated that implementation should begin in 1<sup>st</sup> grade, and completed for all students before reduction at second grade occurred. Likewise, 2nd grade implementation had to be completed before either kindergarten or 3rd grade classes were reduced. These rules meant that kindergarten classes were not reduced until much later in the process, and that most children in the first few years had less than 3-4 years of consecutively smaller classes. The program was not targeted to minority or disadvantaged children, but rather all children in grades K-3. The reduction placed a huge immediate demand on a teacher labor market unable to spike the supply of teachers in the short run. Newly recruited teachers were often inexperienced and lacked certification, and classroom space was less than ideal. An evaluation concluded that the reductions had small effects on achievement in the short term (Bohrnstedt & Stecher), and cited these contextual effects as a possible explanation for the results. Jepsen & Rivlin (2002) did a longer term evaluation and found somewhat larger effects.

Unlike Tennessee, California did not prudently phase in the program beginning in kindergarten so that all children would experience three to four years of smaller classes, nor did they target the intervention to minority and disadvantaged children. This failure to slowly phase in the program led to shortages of teachers and classroom space, and smaller short-term effects from children receiving only 1-2 years of smaller classes. Another unintended side effect of the California initiative was that the number of combination classes (more than one grade taught in a classroom) increased, and analysis of effects for these children showed negative results (Sims, 2004). Haste in implementation also failed to ensure that sufficient data were collected and

available to provide unbiased measurements of short-term effects. For instance, comparable test scores were not available for the years prior to the experiment, and thus the evaluation lacked a critical source of comparative evidence. An unintended consequence of the rush to small classes and not targeting treatment to specific populations was that many high quality teachers in central city schools left for the suddenly available jobs in suburban schools. Inner city schools not only had to recruit teachers to reduce class size, but also had to fill additional vacancies caused by those moving to suburban schools. These changes meant that it was impossible to predict California effects from Project STAR effects, or to provide unbiased measurements of the short-term effects of small classes in California.

In the case of MTO, there were no significant effects on the primary measures of employment, wages and children's achievement. Turney et al. (2006) explored with multiple methods data what might explain the null labor force effects, and under what conditions positive effects might have been expected.

Duncan et al. (2007, chap. 5) used multiple methods data to investigate why the impact on some New Hope participants was high, while for others it was low. Part of this difference was predicted by the context of participants' lives. For some, their lives were dominated by serious obstacles like domestic abuse or addiction that prevented them from taking advantage of New Hope benefits. For these participants, the conditions necessary for effective intervention would have involved addressing those issues. For other participants, mainly men and women without families, utilization of benefits was low and many were able to make significant labor market gains without New Hope. New Hope women with children had conditions in their lives that allowed them to make the most of New Hope benefits including improved and reliable child care and healthcare, which enabled them to make gains for themselves and their children.

## **Box Three- Employing Multiple Methods in Designing and Implementing RCTs**

### **Considerations for Internal Validity**

**What factors led to intervention X working? Failing?**

**What factors led to intervention X working for some groups and not others?**

Project STAR's first publication hypothesized that reduced class size would affect achievement in at least one of three ways: (1) by enhancing teacher morale, (2) by increasing student-teacher interactions, and, (3) by increasing student engagement (Finn & Achilles, 1990). Later analysis of their teacher and classroom data at 4th grade supported the student engagement hypotheses over the other two (Finn et al., 2003). In its aftermath, Project STAR appears to have spawned a rich theoretical literature and set of research studies spanning several disciplines that suggest hypotheses and theories about classroom processes and parental effects that might account for achievement gains in small classes (Blatchford, 2003, 2005; Blatchford, Basset, & Brown, 2005; Blatchford, Bassett, Goldstein, et al., 2003; Blatchford, Goldstein, et al., 1998; Blatchford & Martin, 1998; Bonesrønning, 2004; Boozer & Cacciola, 2001; Bosker, 1998; Bosworth &

Caliendo, 2007; Datar and Mason, in press; Finn et al., 2003; Grissmer, 1999; Hattie, 2005; Webbink, 2005). This literature serves as an example of the development of theories to explain an experimental effect, and what such theories can look like. For instance, this work lent some support to the hypothesis that increased time spent by teachers with individual students in small classes might explain part of the intervention effect.

Project STAR also found larger short-term effects for minority and low income children (Finn & Achilles, 1990, 1999; Krueger, 1999). At 8<sup>th</sup> grade, the reported long-term effects were somewhat mixed on whether there were differential achievement effects for minority and low income students (Finn & Achilles, 1999; Krueger & Whitmore, 2001; Nye et al. 2000a; Nye, Hedges, & Konstantopoulos, 2002; 2004a). But minority and low income students were significantly more likely than similar students in large classes to take college admission tests, graduate from high school and enroll in advanced courses (Finn, Gerber, Achilles, & Boyd-Zaharias, 2001; Finn, Gerber, & Boyd-Zaharias, 2005; Krueger & Whitmore, 2001). Evidence does suggest that teachers spend more time involved in one-on-one interactions with students in small classes (Blatchford et al., 2003). Grissmer (1999) suggests that short-term differential effects may be due to increased individual teacher time devoted to minority and low income students in smaller classrooms that compensate for lack of parental time with children on school related topics. The lack of class size effect for more advantaged students may be due to shifts in parental time and resources in response to class size. For instance, parents may devote more time when class sizes are larger. Datar and Mason (in press) and Bonesrønning (2004) explore whether increased class size influenced types of parent behaviors. Ideally, Project STAR would have collected mixed methods data from parents to assess whether parental time spent with children on school topics is different across racial and SES groups, and whether parental time and resources change in response to class size changes.

Duncan et al. (2007), Yoshikawa et al. (2006) and Weisner (2005) provide the richest examples in the literature of how data collected with multiple methods can be used to provide explanations and refine theoretical hypotheses about results. Duncan et al. (chap. 5) explain through the use of multiple methods data why some participants took more advantage of benefits and made larger gains in income or employment than others. For instance, among women with children, the data suggest that about 20 percent of families had problems (drugs, alcohol, domestic abuse, etc) that could not be addressed by the New Hope benefits offered. Another significant portion of participants eligible for benefits were not constrained in their economic life by factors that the benefits could address. For instance, no employment or income effects were measured for women without children partly because two of the key benefits of child care and health insurance for children were not barriers to employment for these women.

Duncan et al. (2007, pp. 77-79), drawing from Yoshikawa et al. (2006) and Huston et al. (2001), also use multiple methods data from New Hope to summarize why behavior changes resulting from New Hope interventions were different on the part of boys and girls. They suggest that existing higher levels of risk to boys, especially in poor neighborhoods, may have led parents to favor providing more and better day care and after school activities (as indicated by higher enrollment for boys than girls) than for girls.

The Moving To Opportunity (MTO) experiment produced no significant effects on participants' earning and labor force behavior or achievement scores of their children (Kling, Liebman et al., 2007; Sanbonmatsu, Kling, Duncan, & Brooks-Gunn, 2006). Turney et al. (2000) used multiple methods data to develop hypotheses as to why earnings and employment did not change much in response to the intervention. Unanticipated effects that were large and significant occurred for mental health measures of adults and children's behavior (Kling, Liebman et al., 2007). For instance, Clampet-Lundquist et al. (2006) used multiple methods data in the MTO experiment to try and explain differences in behavioral effects for boys and girls.

In some of these instances, the researchers collected data well into and after the experiment that were not part of the original design to further explore causal mechanisms and differential effect sizes by group. Although some multiple methods can be built into the design of RCTs, it may also be efficient to institute a flexible response capability that allows for introducing new multiple methods data collections in response to early RCT findings, especially if such findings are unexpected. In some cases, it is even possible to follow-up with participants long after the end of the experiment to explore theoretical hypotheses. For instance, in Project STAR, while follow-up measurements included effects on high school graduation, taking college entrance exams and advanced courses, no follow-up has thus far tried to collect data that would try to explain what differences between individuals in small and large class sizes might explain these long-term effects.

**Include collection of baseline demographic and other measures to confirm that randomization was accomplished**

Randomization will still leave differences in average characteristics of treatment and control groups. Establishing baseline characteristics of the treatment and control groups can identify those characteristics where average differences do exist. If there are such differences, it would be important to include variables for those characteristics in equations estimating treatment effects. Although Project STAR shows that the average demographic characteristics of treatment and control groups were similar, it would have been desirable to collect baseline test score data at the beginning of kindergarten. In New Hope, researchers collected tracking data on work, benefit usage and supplementary income and state benefit utilization from the beginning of the experiment and showed balance between test and control groups. However, the first extended comprehensive survey was not conducted until two years after random assignment – when more detailed analysis of the results of randomization could be checked.

**Use interviews or surveys to learn how subjects experienced the intervention.**

In Project STAR, researchers annually captured the experiences of K-3 teachers and teacher aides with time logs and a year-end survey that asked about their experiences in small and large classes. Researchers also asked 4th grade teachers to assess the learning behaviors of each child in the experiment. Gerber et al. (2001) offer an analysis of time logs for teacher aides in the experiment and Finn and Achilles (1999) supply an analysis of teachers' assessments of learning behaviors at 4<sup>th</sup> grade for each child. The latter data reveal that teachers' had more positive perceptions of their students' learning behaviors if their students had been in smaller classes in previous grades. Both data collections and their analyses were invaluable in developing theories

and hypotheses about why regular sized classes with teacher aides did not have significant achievement effects, and why small classes did have effects. It would have also been valuable to capture data from children and their parents about their experiences in small and large classes. In addition, it would have been beneficial to collect data from control and experimental groups in high school and beyond that could help explain why long-term effects persisted and particularly large positive differences in high school graduation occurred for minority groups.

New Hope collected interview data with participants in year two, and long-term follow-up five and eight years after initiation. More importantly, researchers initiated an intensive sub-study during the experiment using the 745 participants with children to look at family effects and effects on children's school performance (standardized assessments in mathematics and reading) and behavior. This opportunistic sub-study provided in-depth information about how parents and children experienced and changed as a result of the intervention. This sub-study incorporated the design and fielding of new surveys that included detailed information from parents about their children and also included interviews with children. In addition, children's teachers completed surveys to report on performance and behavior. Finally, researchers incorporated a unique ethnographic study that targeted 44 families. This study employed repeated visits and open-ended home interviews from 1998-2001 and again in 2004. Yoshikawa et al. (2006) provide 13 analyses of the ethnographic data by addressing a number of questions that illustrate the power of such data in understanding the lives and experiences of poor working mothers and their children. Duncan et al. (2007) present an interpretation of results using ethnographic data as follows:

The New Hope offer made a big difference for some people, but it was not a good fit for others. Some parents refused to entrust their children to the care of someone other than a family member. Many parents worked evenings and weekends, when few child-care centers or licensed homes were available. The child-care subsidy was therefore of little use to them. (p. 13)

Two books and recent articles offer rich perspectives on the issues and analyses of specific RCTs with mixed methods data. Yoshikawa et al. (2006) provide a detailed set of analyses and interpretations of the mixed methods data from New Hope that incorporate the ethnographic data. This volume is probably the single best resource for illustrating the value of mixed methods data collection within a specific RCT. Weisner (2005) supplies a wider set of examples of RCTs using mixed methods and the issues and analyses linked to inclusion of such data. These volumes consider the issues involved in the entire process from idea origination, to design of data collection instruments, to their analyses, and to interpretation and integration with the other data from the RCT. Yoshikawa, Weisner, Kalil, and Way (2008) provide a recent perspective on using multiple methods in developmental science and the range of methodological choices available in implementing mixed methods. Gardenhire and Nelson (2003) provide an assessment of the challenges and benefits of qualitative data in four RCTs including New Hope.

A unique use of the ethnographic data allowed Duncan et al. (2007) to gather information about the lives and experiences of three participants in New Hope in a way that illustrates indelibly the complexity of the lives of poor families and their children. Since the life experiences of poor families differs dramatically from the lives of those who set policies and do research,

understanding these lives remains a significant barrier to better policy outcomes and research questions. Mixed methods data of the type collected and analyzed in New Hope can help bridge this “cultural” gap by increasing appreciation for the lives of those targeted by interventions. Such knowledge leads to improved theories, better design of future interventions and better choice among candidate interventions.

**Check measured outcomes for indications that intervention X worked better for some groups than others.**

Each individual has a unique genetic endowment and follows a unique environmental trajectory. Environment effects are also largely expressed through gene-environment interactions (Rutter, 2002). Both individual uniqueness and interactional dynamic makes it unlikely that interventions will have identical effects across participants in any social or educational intervention. Exploring whether effects are different across groups is critical because the cost-effectiveness or benefit/cost ratios of interventions can be made more favorable by targeting interventions on those groups with larger effects (Grissmer, 2002).

Finn and Achilles (1999), Krueger (1999, 2002), Krueger and Whitmore (2002), and Nye et al (2000a, 2002, 2004a) contain analyses of class size effects for Project STAR by income, race and achievement level. These analyses of short-term effects from Project STAR have always found larger effects for minority and low income students. In the longer term, reported differential effects were more mixed for 8<sup>th</sup> grade achievement, but were strongly significant for high school graduation and levels of college entrance test taking (Finn et al., 2005; Krueger & Whitmore, 2001).

Duncan et al. (2007, chap. 5) explain through multiple methods data why some participants utilized benefits more and made larger gains in income or employment than others. They differentiate income and labor force effects for men without children, women without children and women with children. For instance, among women with children, the data suggested that about 20 percent of families had problems that could not be addressed by the New Hope benefits offered (drugs, alcohol, domestic abuse, etc). Another significant portion of participants eligible for benefits did not become engaged with the program for a variety of reasons. No employment or income effects were measured for women without children partly because one of the key New Hope benefits was day care, which has been shown to be a critical barrier to employment for women with children. The barriers for women without children were different and often not addressed by New Hope benefits.

Clampet-Lundquist et al. (2006), using MTO multiple methods data, and Huston et al. (2001), using New Hope multiple methods data, provide two excellent examples of different effects of interventions on children’s behavior by gender. Huston et al. try to explain why the achievement and behavior of male children of New Hope participants improved more than female children, and why the girls’ school behavior actually deteriorated. A viable hypothesis emerged from the mixed methods data that boys experienced greater risk in poor neighborhoods – leading parents to use extra resources to protect their boys against such risk. For instance, in New Hope, researchers found that more boys participated in after school programs with academic and recreational activities than girls.

Clampet-Lundquist et al. (2006) use MTO data to assess why behavior and mental health measures improved for girls, but not for boys, who relocated into higher income neighborhoods. This article provides an excellent example of formulating four competing theoretical hypotheses that might explain these results, and uses multiple methods data -- including a new in-depth interview of teens -- to test these hypotheses.

**Use more intensive interviews, case studies, and ethnographic research to investigate reasons for variability of effects within and between groups.**

Several chapters/articles are helpful in thinking about how to design and conduct multiple methods data collections in RCTs (Brock, 2005; Brock et al., 1997; Cooper, 2005; Cooper, Brown, Azmitia, & Chavira, 2005; Datta, 2005; Duncan & Raudenbush, 1999, 2001; Fricke, 2005; Gibson-Davis & Duncan, 2005; Goldenberg, Gallimore, & Reese, 2005; Greene, 2005; Harkness, Hughes, Muller, & Super, 2005; Huston, 2005; Weisner, 2002; Weiss, Kreider, Mayer, Hencke, & Vaughan, 2005).

In the New Hope study, researchers used extensive interviews and ethnographic data to develop theories and hypotheses about the reasons for differential effects (Duncan et al., 2007, Huston et al., 2001; Weisner, 2005; Yoshikawa et al., 2006). For instance, Duncan et al. (chap. 5) create a new categorization that distinguishes participants by “potential obstacles” of utilizing New Hope benefits – partly based on interview and ethnographic data. For families with substantial barriers (drug and alcohol abuse, arrest records, presence of developmentally impaired children, domestic abuse, etc), New Hope did not offer much to target such barriers, and intervention effects on these families were small or non-existent. On the other hand, some participants in both test and control groups proved their abilities to accomplish New Hope objectives without New Hope benefits, leading to small or non-existent overall effects. The largest effects were for families “poised to profit” from the specific benefits offered by New Hope. For instance, the child care benefit allowed some parents to upgrade the quality of their day care substantially. This illustrates that mixed methods data allow analysis of categories that go far beyond the usual gender, race and income categories.

Huston et al. (2001) explain why the achievement and behavior of boys improved more than for girls for New Hope participants, and why girls’ school behavior actually deteriorated for New Hope participants. In both cases a viable hypotheses emerged from the mixed methods data that placed boys at existing greater risk in poor neighborhoods, leading parents to favor using extra resources to protect boys against such risk. For instance, in New Hope, it was found that boys more often participated in after school programs with academic and recreational activities than girls.

Bos, Duncan, Gennetian, and Hill (2007) provide an example of employing in-depth interview data to highlight the fear associated with threats to safety in the lives of poor families, especially single-parent families. Bos et al. state, “In the qualitative sub-study, parents appeared to worry more about their boys than about their girls, especially when they reached early adolescence. There was experimental evidence that New Hope’s child care supports were more likely to be used for boys than for girls. Mothers often said that their boys were vulnerable, and they used any resources they had to counteract negative influences. As one mother said, ‘It’s different for

girls. For boys, it's dangerous. [Gangs are] full of older men who want these young ones to do their dirty work. And they'll buy them things and give them money' (p.12 ) New Hope boys were more likely than girls to be in organized after-school programs where they received help with homework and had opportunities for recreation (Duncan et al., 2007). The larger impact on boys may be explained by the fact that from the parents' perspectives, boys had much more to gain from in intervention than girls.

In addition, there are several other examples from the literature on using ethnographic, interview and other mixed methods data to investigate why effects occur and can change across participants (Bernheimer, Weisner, & Lowe, 2003; Datta, 2005; Lowe & Weisner, 2004).

Clampet-Lundquist et al. (2006) use MTO data to assess why behavior and mental health measures improved for girls, but not for boys who relocated into higher income neighborhoods. This article provides an excellent example of formulating four competing theoretical hypotheses that might explain these results, and uses multiple methods data -- including a new in-depth interview of teens -- to test these hypotheses. Kling, Ludwig and Katz, (2005) supply a testimonial to the value of the in-depth interviews collected in MTO, and illustrate how such data were used in the article. We cite from the introduction to this article:

Our qualitative fieldwork had a profound impact on our MTO research. First it caused us to re-focus our quantitative data collection strategy on a substantially different set of outcomes. In particular, our original research design concentrated on the outcomes most familiar to labor economists: the earnings and job training patterns of MTO adults and the school experiences of their children. Our qualitative interviews led us to believe that MTO was producing substantial utility gains for treatment families, but primarily in domains such as safety and health that were not included in our original data collection plan. *In our subsequent quantitative work, we found the largest program effects in the domains suggested by the qualitative interviews* [italics added].

Second, our qualitative strategy led us to develop an overall conceptual framework for thinking about the mechanisms through which changes in outcomes due to moves out of high poverty areas might occur. *Our conversations with MTO mothers were dominated by their powerful descriptions of their fear that their children would become victims of violence if they remained in high poverty housing projects* [italics added]... This fear appeared to be having a significant impact on the overall sense of well-being of these mothers, and it was so deep-seated that their entire daily routine was focused on keeping their children safe... We hypothesized that the need to live life on the watch may have broad implications for the future prospects of these families.

Third, our fieldwork has given us a deep understanding of the institutional details of the MTO program. This understanding has helped us to make judgments regarding the external validity of our MTO findings, particularly regarding the relevance or our results to the regular Section 8 program. *In addition, this understanding has prevented us from making some significant errors in interpreting our quantitative results* [italics added].

*Fourth, by listening to MTO families talk about their lives, we learned a series of lessons that have important implications for housing policy. For many of the things we learned, it is hard to imagine any other data collection strategy that would have led to these insights [italics added] (pp. 244-245).*

## **Does intervention X remain effective when different outcomes are used?**

**Include multiple quantitative outcome measures to assess different aspects of the desired outcomes (e.g., specialized outcome measures aligned with the purpose of intervention X as well as more general measures such as standardized test scores).**

**Use case studies, interviews and observations to detect unanticipated/unmeasured outcomes.**

The experience in Project STAR, New Hope and particularly MTO, as well as other RCTs with long term follow-up studies suggests that the effect of social and educational interventions are (1) unlikely to be confined to a single outcome measure or to a single generation, especially in the long term, (2) some outcomes are likely to be unpredictable and/or unanticipated, especially in the long term, and, (3) some of the primary measures often chosen by researchers can have small and/or null effects (e.g. achievement, labor force measures), while unanticipated measures, usually health and behavioral measures, can have large and significant effects.

It is also important to remember that the importance of effects depends not only on their effect size, but also on their contribution to long-term benefits. For instance, the Abecedarian and Perry Preschool experiments originally employed IQ and achievement test scores to measure academic performance. Although the participants showed improvement on these measures, most of the benefits flowed either in the form of lower levels of grade retention and special education placement, or changes in behavior that resulted in less involvement with the criminal justice system (Karoly et al., 1998). Other important unanticipated effects included examples of generational effects like lower levels of addictive behavior and teen pregnancy (Karoly et al., 1998; Karoly, Kilburn, & Cannon, 2005; Masse & Barnett, 2002; C. T. Ramey, Campbell, Burchinal, Skinner, Gardner, & S. L. Ramey, 2000; Reynolds, Temple, Robertson, & Mann, 2002; Reynolds et al., 2007; Schweinhart, 2004). Failure to measure the full range of effects can result in significant underestimation of the benefits of an intervention. Although using multiple methods does not guarantee that all effects will be measured, these data provide the best opportunity to capture unanticipated outcomes and develop stronger theories that can better predict the full range of outcomes.

In Project STAR, the outcome measures used in the short-term were standardized test scores, grade retention and special education placement. These effects were certainly significant, but the effect sizes ranging from .2 to .3 generally would not be expected to have such large impacts on high school graduation or signing up for college entrance examinations. The achievement gains in elementary school translated into significantly higher secondary school graduation rates and increased levels of taking college entrance tests (Finn et al., 2001, 2005; Krueger & Whitmore, 2001).

New Hope's original objective was to move families out of poverty through more stable and higher paying jobs and better health care. However, a new set of outcome measures were introduced when the supplemental parent-child study started during the second year of the experiment. This study assessed, among other measures, changes in parenting practices, children's school performance (as rated by teachers and through standardized testing) and children's behavior (Duncan et al., 2007; Huston et al., 2001).

The original objectives of MTO involved improvements in income and labor force behavior and children's performance in school. In general, the experiment showed no significant effects for any of these measures (Goering & Feins, 2003; Katz, Kling, & Liebman, 2001; Ladd & Ludwig, 1997; Rosenbaum & Harris, 2001; Sanbonmatsu et al., 2006). However, in-depth interviews alerted researchers to refocus their analyses on mental health, criminal behavior and children's conduct measures, which showed large effects (Browning & Cagney, 2003; Kling, Ludwig and Katz, 2005; Kling, Liebman and Katz, 2007; Leventhal & Brooks-Gunn, 2003b; Ludwig, Duncan, & Hirschfield, 2001).

**Are all of the components of intervention X necessary for it to work, or are some unnecessary? Are some needed components missing?**

**Plan to measure the various intervention components; build in case studies to learn which components mattered to different subjects and to generate hypotheses about other components that might have made intervention X more effective.**

Determining whether all components are necessary and whether some components mattered more to some participants than to others is critical since simplifying and targeting an intervention can significantly reduce costs. For instance, Project STAR analyses asked whether four years of smaller classes were required to affect achievement, or whether similar effects would occur with fewer years of intervention. Smaller class sizes are costly, and if each year did not make contributions to the effect, significant cost savings would be possible. Hanushek (1999) suggests that most of the achievement effects occurred in the first year. However, Krueger (1999), Finn et al. (2001), and Nye, Hedges, and Konstantopoulos (2001) suggest that 3 or 4 years of small classes are required for sustained, long term achievement effects. It would have been desirable in Project STAR to systematically vary the class size rather than aim for reductions of 8 pupils per class across all schools. Perhaps most of the achievement gains were due to reductions of 6 rather than 8 pupils. If so, then in future class size reduction initiatives, significant cost saving would be possible.

Project STAR analyses showed higher effects on achievement from smaller classes through grade three for minority and disadvantaged students (Finn & Achilles, 1999; Krueger, 1999). However, Finn et al. (2001), Nye et al (2000a, 2002, 2004a) and Krueger and Whitmore (2002) suggest that the effect sizes declined somewhat by 8<sup>th</sup> grade and the larger effects for minority and disadvantaged students were mixed at 8th grade. Krueger and Whitmore (2001) and Finn et al. show significantly larger effects on high school graduation and college entrance test taking for minority and disadvantaged students who participated in the experimental small classes. Finishing high school requires more than direct cognitive gains. Other developmental skills such

as social skills and behavioral and emotional skills play important roles in completing education and in labor force success. This suggests that Project STAR may have affected children's social, behavioral or emotional trajectories as well as their cognitive trajectories. Finn and Achilles, (1999) argue that improved classroom behavior may partially account for achievement gains, but ideally a wider range of developmental measures would have been included in kindergarten through third grade and in the longer term follow-ups.

This pattern of larger long-term effects for measures other than direct achievement measures seems to be emerging as a consistent finding from several early interventions of long duration. For instance, the Perry Preschool and Abecedarian projects had significant effects on many behavioral measures such as reduced involvement with the criminal justice system, even though achievement gains leveled off or declined in the longer term (Karloly, Greenwood, et al, 1998; Karoly, Kilburn, et al., 2005; Masse & Barnett, 2002; Ramey et al., 2000; Reynolds, Temple, Ou, et al. 2007; Reynolds, Temple, Robertson, et al., 2002; Schweinhart, 2004).

New Hope found the utilization of three key benefits widely divergent across participants, and that the largest effects were for those participants whose particular life circumstances "fit" the particular menu of offered benefits. For instance, the cost of day-care was a prime benefit for mothers with children -- especially those with multiple children -- so men and women without children could not take advantage of this lucrative benefit. Also, many children had significant health or disability problems, and the health insurance benefit provided coverage for these kinds of issues. Perhaps one of the major lessons arising from New Hope is the need to characterize the diverse needs of a population before designing the benefit package, and offering a wider and more flexible menu that would address a broader range of issues (Duncan et al., 2007). As an example, about 20 percent of participants had more severe problems linked to drugs, alcohol or domestic abuse. For these families, other interventions were needed before they could take advantage of the New Hope benefits (chap. 3-4).

### **Are the treatment effects sustained over time?**

**Plan extended follow-ups, particularly of treatment group members, using both quantitative and qualitative data (e.g., achievement data, case studies, interviews).**

Project STAR has followed participants through high school. Achievement data were collected at 8<sup>th</sup> grade. Then, at the end of high school, data were collected on how many college entrance tests were taken as well as high school graduation rates. The analyses suggest that the size of achievement effects declined somewhat at 8<sup>th</sup> grade, and earlier differential effects for minority and disadvantaged effects were mixed (Krueger & Whitmore, 2002; Nye et al, 2000a, 2002, 2004a). However, Finn et al. (2001) and Krueger and Whitmore (2001) show large effects on high school graduation and levels of taking college entrance tests, with much larger effects for minority and disadvantaged students.

In general, the long-term effects of New Hope and MTO tended to be small or non-existent for direct labor force measures such as income and employment, but were somewhat larger for selected behavioral and school performance of participants' children by gender and school subject (although MTO did not directly measure the effects on children's achievement) (Kling et

al., 2007; Sanbonmatsu et al., 2006). Also, adult and children's mental health measures showed positive long-term effects (Leventhal & Brooks-Gunn, 2003b; Kling et al.).

New Hope followed up with interviews two years and five years after the experiment ended to determine long-term effects on participants and their children. Duncan et al. (2007, chap. 11) report that the larger and most persistent effects were on the children -- particularly the achievement and behavior of the boys. This is another example of the importance of measuring generational effects. It is possible that smaller, but persisting improvements in the lives of parents, particularly single mothers, can generate larger and longer lasting effects in the next generation. Thus far, the long-term effects on the children of the individuals that participated in interventions like Perry Preschool or Abecedarian have not been measured.

Clampet-Lundquist et al. (2006) use data from follow-up interviews with MTO participants four to seven years after project initiation to explore differential effects on behavior changes in boys and girls. They also carried out an additional data collection with a sub-sample of teens focusing on a theory based set of hypotheses directed at explaining gender differences in outcomes. Clampet-Lundquist et al.'s article is an excellent example of adding a multiple methods data collection five years into the experiment to further test hypotheses generated by the original data collection. The original analysis suggested no differences in risk behavior for boys in the treatment and control groups, but in fact, girls exhibited better mental health and lower risk behavior. The added in-depth interviews with teens, together with the original follow-up data, suggested specific viable explanations for the differences in outcomes between genders.

## **Considerations for External Validity**

**How do contextual effect factors affect the impact of intervention X?**

**Use case studies, administrative data, interviews and observations to document contextual factors (e.g., local policy environment, resources, cultural concerns, history) and how they might interact with intervention X.**

Researchers face substantial obstacles in translating successful small-scale experiments into successful large-scale programs [See Schneider and McDonald (2007, vol. I-II) for an excellent review of research on scaling-up.] Experiments only provide results with predictive validity if the conditions and context of the experiment can be duplicated in other settings. However, experimental conditions can never be perfectly duplicated. In fact, the conditions in experimentation (a high degree of control of conditions, personnel selected by researchers, etc) that are necessary to make experiments successful from a scientific perspective, often guarantee smaller effects in scaled-up real world settings. And, contextual effects seem to be ubiquitous in social and educational interventions. One of the key advantages of multiple methods in RCTs is to provide information that can better predict how results might change in different contexts, conditions and scales.

Although Project STAR was carried out on a large scale in "real world" conditions, results from the experiment cannot be assumed to transfer to different populations in different schools under

different conditions. The relatively smooth implementation of STAR in 79 Tennessee schools stands in stark contrast to the state-wide class size reductions in California plagued by teacher shortages and limited space (Bohrnstedt & Stecher, 2002).

In Project STAR, each of 79 schools represented a separate experiment because each school included at least one randomly assigned small class, a large class and a large class with teacher aides. But, the context was different across schools enabling the researchers to explore contextual effects. Teachers were also randomly assigned to classrooms to enable research on teacher effects in classrooms. Two simple and important examples of contextual effects in STAR are that minority and disadvantaged students experienced higher achievement effects, and that all students in small classes for 1-2 years, rather than 3-4 years, experienced no sustained effects (Finn & Achilles, 1999). Thus, STAR effects would be predicted to vary by student characteristics, and by the number of years of small classes between K-3. However, STAR data have also been used to explore more complex contextual effects.

For instance, Nye et al. (2004b) and Peevely, Hedges, and Nye (2005) explored the effect on achievement gains of teacher experience, education and salary and classroom composition. They suggested that teacher experience effects are larger in math than reading, and that lower SES classrooms have larger variance in score gains due to teachers than higher SES classrooms. Dee (2004) suggests that students who have same race teachers have higher score gains. Nye et al. (2000b) analyzed contextual effects of class composition and school location and concluded that class composition and location do not change effect size significantly.

New Hope was a small-scale experiment conducted in an economic and welfare policy environment (Wisconsin) that was not typical of other states. State economic conditions produced a labor market that was quite favorable to finding and maintaining employment. For instance, Duncan et al. (2007) point out that the control group also generated substantial gains in employment and income, making the explanation of treatment effects challenging. However, the New Hope participants made even larger gains in employment and income than controls. Wisconsin also was at the forefront of welfare reform such that generalizations of New Hope to other states became problematic. However, the rich multiple methods data enabled researchers to do more than speculate on how a program like New Hope might be redesigned and scaled up nationally (Duncan et al., chaps. 6-7).

Interesting neighborhood contextual factors were identified in MTO. Turney et al. (2006), using multiple methods data from in-depth interviews with participants, identified barriers to employment. These barriers may explain how, in spite of relocation to better neighborhoods in Baltimore, participants did not experience employment and earnings effects. Identifying such barriers helps to delineate conditions in other cities that might be needed for earning and employment effects to occur.

### **How close are the measured outcomes to outcomes of interest?**

**When designing the study, interview key stakeholders to determine the relevance/appropriateness of the outcome measures proposed for the study.**

The ultimate stakeholder in social and educational experimentation is the American taxpayer. For these stakeholders, a commonly used criterion is that the long-term benefit to society (measured in monetary terms) must at least exceed societal costs, and hopefully have a rate of return that justifies government borrowing. However, Karoly and Bigelow (2005) suggest that many government programs would not meet this criterion, and offer an alternate criterion -- that a particular early childhood program only needs to have a benefit/cost ratio higher than other government programs.

Near-term stakeholders in Project STAR included the Tennessee legislature, Tennessee teachers and parents of K-3 students. The Tennessee legislature authorized Project STAR, which clearly indicated that raising achievement was a prime objective. But, the impetus for smaller class sizes at the policy level was due to pressure from parent and teacher groups. Stakeholders were directly involved in specifying the intervention, as well as setting objectives [See Ritter and Boruch (1999) for a history of STAR.] However, while the initial focus was on immediate achievement, the most important outcomes for society showed up in the long-term follow-ups where researchers were able to register substantial gains in high school graduation and college entrance behavior.

The evolution of New Hope had a long history, dating from over 15 years prior to the initiation of the experiment (Duncan et al., 2007, chaps 1-2). A clear objective was to provide evidence for how to design an improved welfare system. Besides federal and state policymakers, stakeholders included the business community and welfare participants themselves. The broadened objectives in New Hope transformed from a primary emphasis on adult labor force outcomes to a strong emphasis on behavioral outcomes for both parents and children, as well as schooling outcomes for boys. For the children's outcomes, the dual emphasis on school achievement and behavior both in and out of school proved to be an important and persisting generational result of the New Hope experiment.

The federal Department of Housing and Urban Development was the sponsor of MTO. Its clear purpose was to determine how important neighborhoods were to adult outcomes, so that better housing policies could be promoted. However, MTO expanded from a primary emphasis on improvements in labor market measures, which showed no statistically significant effects, to measures of mental health, parenting and children's behavior (Kling et al., 2007; Sanbonmatsu et al., 2006). No achievement effects were found, but some effects on children's behavior and mental health of adults and children were significant (Kling et al.; Leventhal & Brooks-Gunn, 2003b). One of the key findings from MTO is that non-experimental research had overestimated neighborhood effects. In fact, neighborhood effects are smaller, involve a broader range of outcomes and are more complex than previously thought (see, for instance, Booth & Crouter, 2001; Duncan & Raudenbush, 1999, 2001; Kling et al.; Leventhal & Brooks-Gunn, 2003a; Turney et al., 2006).

**How would resource constraints affect the institutionalization of intervention X if it were found to be effective?**

**Build collection of cost data into the study and conduct cost-effectiveness and cost benefit analysis.**

Levin and McEwan (2000, 2002) provide a good introduction to conducting either cost-effectiveness or cost-benefit analyses and distinguishing between them. A cost-effectiveness analysis compares the proposed intervention to alternate interventions that focus on a single common outcome (e.g., higher achievement). Cost-benefit analyses use a single intervention to see if long-term monetary benefits exceed costs from all outcomes.

There are several good examples of conducting analyses incorporating costs and benefits. Karoly et al. (1998) compare the costs and benefits of Perry Preschool and a nurse visiting program. Karoly & Bigelow (2005) analyze the costs and benefits of universal preschool in California. Masse & Barnett (2002) estimate the costs and benefits from the Abecedarian Project. Reynolds et al. (2002) perform a cost-benefit analysis of a Chicago early childhood intervention. Lynch (2004) summarizes several cost-benefit analyses of early childhood programs. Grissmer (2002) contains a cost-effectiveness analysis of four options for improving achievement.

Such analyses cannot usually yield reliable predictions for scaled-up programs in different contexts. Although it is usually assumed that effect sizes can change with context, costs as well as effects are sensitive to context, so costs measured in experimental settings may change dramatically in large-scale settings. Brewer, Krop, Gill, & Reinhardt, (1999) illustrate the variance in the cost of class size reductions depending on location (cost of living differences), the specific rules used to implement such reductions, the availability of space, the hiring practices of teachers, the pay scales of teachers and the characteristics of the students targeted for smaller class size. For instance, implementing class size reductions in inner cities carries higher space and teacher costs, but also leads to larger effect sizes.

Moreover, scaling up small-scale interventions to large-scale public sector programs can carry several additional cost considerations. For instance, because such programs depend on forming a successful political coalition for passage, powerful stakeholders can lobby for wider eligibility in experimental groups, which can lead to higher average costs and lower effects. Gordon (2004) suggests that federal allocations of Title I funding (for low income students) to local governments gets partially diverted by local governments for alternative non-educational uses. Finally, programs are rarely fully funded. Duncan et al. (2007, chap. 8) provide analyses of cost-benefits for New Hope and a discussion of the implications for expanding New Hope to a larger state or national program.

**How do the details of the intervention and the controls imposed by the study design differ from the real world conditions under which intervention X might be implemented?**

**Collect and report descriptive data that will allow policymakers to assess the similarity of the sample population and setting to those in other situations to which they might want to generalize results.**

Project STAR was a large-scale intervention involving over 12,000 students in grades K-3 in mostly large suburban and urban schools in Tennessee. Tennessee children in STAR included disproportional numbers of minority and disadvantaged participants compared to all Tennessee students, and Tennessee students are disproportionately more disadvantaged than U.S. students

(Grissmer, 1999). The larger effects for minority and disadvantaged children mean that average effects can change markedly as the composition of students change across states. Tennessee implemented Project STAR almost entirely in suburban and inner city schools. The costs and effects may change for rural schools (where recruiting teachers may be more difficult) or in states with higher or lower costs of living than in Tennessee. Tennessee also utilized experienced teachers rather than hiring new teachers, although the experiment did not provide for specific preparation or instruction directed at teachers working with smaller classes.

Researchers learned many lessons from Project STAR that could guide policy and implementation in other states. Three important lessons were: (1) 3-4 years of small classes were needed for long-term effects, (2) effects were much larger for minority and disadvantaged children, and (3) the teachers in Project STAR were not newly recruited, but were drawn from the pool of existing experienced teachers. Project STAR did spur class size reductions in many states beginning in the 1990s, which extended into the next decade. In general, these reductions were more often directed to schools and districts with larger proportions of minority and disadvantaged students. Grissmer, Flanagan, Kawata, and Williamson (2000) and Grissmer and Flanagan (2006) use state NAEP scores to assess the effects of class size reductions and other initiatives across states from 1990-2005. They conclude that consistency exists with the average effect of such class size reductions and the results from Project STAR.

California was the notable exception in successfully building off of Project STAR. California mandated sizable class size reductions statewide for all students in grades K-3 beginning in 1996 (Bohrnstedt & Stecher, 2002). The short time between mandate and implementation, and the fact that all children in grades K-3 were eventually involved, left school districts throughout the state unprepared for hiring the necessary additional teachers and finding the needed classroom space. Unlike Tennessee, California did not prudently phase in the program beginning in kindergarten so that all children would experience 3-4 years of smaller classes, nor did they target the intervention to minority and disadvantaged children. This failure to phase in the program more slowly generated shortages in teachers and classroom space. The initiative led to smaller short-term effects due to including more advantaged children and children receiving only 1-2 years of smaller classes. Such haste also failed to ensure that sufficient data were collected and available to provide unbiased measurements of short-term effects. For instance, comparable test scores were not available for the years prior to the experiment, so the evaluation lacked a critical source of comparative evidence. An untended consequence of the rush to small classes and lack of targeting was that many better quality teachers in central city schools left for the suddenly available jobs in suburban schools. Inner city schools not only had to recruit teachers needed to reduce classes, but also had to fill additional vacancies caused by those moving to suburban schools. These changes meant that it was impossible to predict California effects from Project STAR effects, or to provide unbiased measurements of the short term effects of small classes in California.

New Hope was a small-scale program with volunteer participants. Thus, expansion to a large-scale program would mean incorporating those populations who did not volunteer, as well as all the cost and effectiveness issues associated with scaling up from small experimental programs (Quint, Bloom, Black, & Stephens, 2005; Schneider & McDonald, 2007). Duncan et al. (2007,

chap. 8) provide analyses of cost-benefits for New Hope, and a discussion of the implications and uncertainty involved in scaling New Hope to a state or national program.

## **Box Four- Role of Multiple Methods in Providing a Deeper Understanding of Study Findings**

**In addition to using quantitative measures to assess outcomes, use data from case studies, interviews, surveys, and/or observations to interpret the observed outcomes (e.g., how intervention was experienced and responded to by subjects in differing circumstances).**

Perhaps the main reason why collecting multiple methods data in RCTs is necessary is that each participant in any social science RCT has a unique genetic and developmental history, and many of the forces shaping development involve gene X environment interactions (Rutter, 2002). So the a priori expectation should be of differential effect size across participants. If theories are to be ultimately successful in predicting behavior, they must in some way take account of and incorporate this wide diversity inherent in study subjects. Data from multiple methods can be seen as a start to understanding this uniqueness and diversity and exploring ways of identifying groups with similar enough paths and responses that enable more efficient targeting and more accurate predictions. These individual paths and responses to interventions can only be captured by multiple methods data.

New Hope collected an extremely rich set of multiple methods data, and researchers have used these data to try and understand several emerging issues. These issues include why control participants who did not receive New Hope benefits made large labor market gains; why the incrementally larger gains made by participants receiving New Hope benefits were statistically significant, but of modest size; and why many participants eligible for New Hope benefits did not utilize their benefits or utilized them only sporadically. In addition, the differential effects for boys on behavior and achievement were puzzling.

The 12 analyses contained in Yoshikawa et al, (2006), using New Hope data, provide outstanding examples of: (1) how multiple methods data can address unexpected utilization and results, (2) how to make these types of interventions and similar policies more effective, and, (3) in general, how results are dependent on context. Huston et al, (2001) and Duncan et al. (2007) provide examples of incorporating the analyses of multiple methods data into academic and policy publications. A volume by Weisner (2005) contains 12 chapters that illustrate the value of multiple methods data to address research questions not necessarily embedded in RCTs. As such, it provides material that is helpful in learning how such methods have been used across different research areas, what kinds of methods have been employed, and how these data have contributed to testing hypotheses and theories about why and how behavioral effects occur.

Perhaps more importantly, these analyses illustrate the complexity and uniqueness of the lives of poor, working mothers, and why it is difficult to design interventions and policies that could have very large impacts for high proportions of such women. For instance, Lowe, Weisner, and Geis (2003) provide a picture of the challenge of finding day care for the children of poor,

working mothers and the problem with “one size fits all” benefit packages. An important lesson drawn from New Hope is the need for more extensive and flexible benefit options to address the complexity and diversity in the lives of poor working mothers and their children (Duncan et al., 2007)

In Project STAR, Finn et al. (2003) utilize observational data, teacher surveys and interviews to address the question of why small classes work. Gerber et al. (2001) also employ teacher and teacher aide logs, surveys and interviews to address why the effects from adding teacher aides to classrooms did not have large effects.

Clampet-Lundquist et al. (2006) employed follow-up interview data in MTO to develop and test hypotheses as to why the experience of changing neighborhoods was different for girls than boys, and why girls fared better than boys in new neighborhoods. Turney et al. (2006) use interview data from MTO participants to explain why moving to higher income neighborhoods had no effects on employment, income or welfare utilization.

### **Use data from case studies and/or interviews to illustrate findings in a compelling manner.**

There are two primary audiences for the findings of RCTs: researchers and policymakers. Researchers tend to be concerned about whether effects occur and how big the effects were compared to alternatives. However, in order to develop theories and improve research designs, researchers will increasingly have to address questions such as *why* effects occur. The constraints on normal academic publications often preclude the longer page length required to address such questions. In complex RCTs with extensive multiple methods, results need to be communicated through edited books or longer summary publications. Yoshikawa et al. (2006) provide an indispensable resource for researchers designing mixed methods RCTs and communicating their results to other researchers. This volume is entirely directed toward using multiple methods data to address key issues in explaining the pattern of New Hope results, particularly the differential effects across adults and children. Weisner (2005) provides examples for researchers from a wider range of studies.

Policymakers also need publications that illustrate findings in a compelling manner and address concerns specific to their roles. Policymakers must develop political support for any new program, and so both legislators and the public need to be convinced of the merits of a program. Policymakers will face questions from legislators and the public about why a particular program will work, how it can be targeted to achieve larger effects and what it will cost. In such contexts, relating stories about how individual participants responded, why it worked for some participants and not others and how lives were changed can be effective methods of communicating for policymakers. Policymakers also need research translated into “readable” and compelling prose. Duncan et al. (2007) provides an outstanding example of communicating the results of mixed methods analyses to a more general audience, including policymakers. This volume wraps the basic results of the intervention around a compelling narrative that illustrates how and why the intervention worked in individual cases, how one could change the design to obtain a more effective and efficient intervention, how to set eligibility rules and in what contexts this

intervention might or might not work, as well as the potential risks of moving to a large-scale program.

**Examine quantitative and qualitative results to determine whether additional hypotheses (e.g., about additional outcomes, modifications to the intervention) might be pursued in subsequent studies or different stages of the current RCT.**

RCTs are usually envisioned as having a fairly unchangeable design that includes well-defined planning, implementation and analysis stages. This format is dictated largely by the federal proposal process. Such a research design is generated in accordance with pre-existing theories and a fixed set of outcome measures, yet are problematic when RCTs show unexpected results or have unexpected outcomes. Although follow-up RCTs might be designed to address these issues, it may be more efficient and timely to use resources to expand data collection either during the RCT or in longer term follow-up. In fact, multiple methods RCTs are directed toward answering a more complex set of questions than a black-box RCT, and the chances of unexpected results are correspondingly higher. Puzzling differential outcomes and results require new theories. This kind of research likely requires a more flexible and opportunistic research funding process that is able to respond to unexpected findings.

Both New Hope and MTO might be described as having an evolving and opportunistic research strategy that responded to emerging research findings with additional and expanded multiple methods data collections. These data collections were targeted toward identifying a wider set of outcomes, framing and testing new emerging hypotheses, and providing explanations of unexpected effects.

MTO was the first large-scale RCT designed to explore the effect of neighborhoods on adults' economic outcomes and children's schooling outcomes by randomly assigning volunteer participants to differential access to higher income neighborhoods. However, no significant effects were found on adult employment and income across the five locations and no effects were found on children's schooling outcomes (Kling, Liebman and Katz, 2007; Sanbonmatsu et al., 2006). The null effects from these primary measures resulted in a redirection of the study to determine if there were effects on the mental and physical health of adults and children, and on the incidence of risky behavior among youth.

These data came from a follow-up survey about seven years after the initiation of MTO to each participating adult and up to two children per household that included a much wider set of outcome measures, and also explored possible explanations for the null effects on economic outcomes. Kling et al. (2007) provide a summary of these results that suggests large and significant positive effects on adult mental health measures, but no effects for physical health measures. Young females experienced positive effects on physical and mental health and lower incidence of risky behavior. However, male youth showed no effects or offsetting negative effects on each of these measures. Kling et al. also provide some hypotheses to explain the null effects on adult economic measures. Turney et al. (2006) use the long term follow-up and an additional in-depth interview with 67 participants to explain why moving to higher income neighborhoods had no effects on employment, income or welfare utilization. Clampet-Lundquist

et al. (2006) also use this follow-up interview and an additional in-depth interview of 86 teens to develop and test hypotheses as to why girls fared better than boys when parents moved to a higher income neighborhood. Multiple methods data collections were critical in instigating the redirection of data gathering and interpretation. An article by Kling, Liebman and Katz, (2005) provides unusual and compelling testimony that highlighted the value of qualitative, in-depth interviews and eventually changed the research strategy for the MTO project. These interviews helped to identify mechanisms driving the outcomes and offered insights into interpreting results. Perhaps more than any other single publication, Kling, Liebman and Katz, (2005) should be read by those researchers and policymakers who question the value of embedding multiple methods in RCTs. We cite directly from the introduction to Kling, Liebman and Katz, (2005), but the article goes on to elaborate on how the in-depth interviews conducted influenced their research:

Our qualitative fieldwork had a profound impact on our MTO research. First it caused us to re-focus our quantitative data collection strategy on a substantially different set of outcomes. In particular, our original research design concentrated on the outcomes most familiar to labor economists: the earnings and job training patterns of MTO adults and the school experiences of their children. Our qualitative interviews led us to believe that MTO was producing substantial utility gains for treatment families, but primarily in domains such as safety and health that were not included in our original data collection plan. *In our subsequent quantitative work, we found the largest program effects in the domains suggested by the qualitative interviews* [italics added].

Second, our qualitative strategy led us to develop an overall conceptual framework for thinking about the mechanisms through which changes in outcomes due to moves out of high poverty areas might occur. *Our conversations with MTO mothers were dominated by their powerful descriptions of their fear that their children would become victims of violence if they remained in high poverty housing projects* [italics added]..... This fear appeared to be having a significant impact on the overall sense of well-being of these mothers, and it was so deep-seated that their entire daily routine was focused on keeping their children safe. .... We hypothesized that the need to live life on the watch may have broad implications for the future prospects of these families.

Third, our fieldwork has given us a deep understanding of the institutional details of the MTO program. This understanding has helped us to make judgments regarding the external validity of our MTO findings, particularly regarding the relevance of our results to the regular Section 8 program. *In addition, this understanding has prevented us from making some significant errors in interpreting our quantitative results* [italics added].

Fourth, *by listening to MTO families talk about their lives, we learned a series of lessons that have important implications for housing policy. For many of the things we learned, it is hard to imagine any other data collection strategy that would have led to these insights* [italics added]. (pp. 244-245)

The original focus of New Hope was also on adult economic measures: increased employment and wages and less welfare dependency. However, partly due to funding opportunities, and partly due to lower utilization of benefits by adults without children, the experiment increasingly

focused on outcomes for mothers -- about 71 percent of the total sample. Researchers introduced new data collection measures that focused on a wider set of adult health, parenting and other behavioral measures, as well as measures to link performance in schools to health and behavioral measures of the participants' children (Duncan et al., 2007; Huston et al., 2001).

Bos et al. (2007) provide an example of employing qualitative data that, like the MTO in-depth interviews, highlights the fear associated with threats to safety in the lives of poor families -- especially single parent families. Bos et al. states "In the qualitative sub-study, parents appeared to worry more about their boys than about their girls, especially when they reached early adolescence. There was experimental evidence that New Hope's child care supports were more likely to be used for boys than for girls. Mothers often said that their boys were vulnerable, and they used any resources they had to counteract negative influences. As one mother said, 'It's different for girls. For boys, it's dangerous. [Gangs are] full of older men who want these young ones to do their dirty work. And they'll buy them things and give them money' (p.12 ) New Hope boys were more likely than girls to be in organized after-school programs where they received help with homework and had opportunities for recreation (Duncan et al., 2007). The larger impact on boys may be explained by the fact that from parents' perspectives, boys had much more to gain from the intervention than girls.

## **Box Five- Explore implications for research, policy and next steps.**

### **Relate outcomes of the current study to findings from prior research.**

**Do the results of this RCT confirm or contradict results from other studies of similar interventions? Consider results from RCTs and other types of studies (e.g., quasi-experimental, correlational and ethnographic).**

**What factors may account for differences in results between this RCT and previous studies? Take account of variations in study design, characteristics of participants, outcome measures, settings, times, and fidelity of implementation.**

Ideally, a literature review (as outlined in the discussion of Box One of this guide - "What is Known from Previous Research") is available and can serve as the basis for integrating the new results with outcomes from previous studies. A major motivation for conducting a multiple methods RCT is to make it more likely that future literature reviews will generate a scientific and/or policy consensus. So, the integration of the present results with the previous literature review should focus on if and how the present results settle disparities existing in the previous literature, and/or raise new questions and issues that must be the subject of future research.

Multiple methods RCTs are likely to emerge as the primary method to explain both direct and indirect disparities in past measurements. The direct contribution arises from being able to measure contextual effects, measure differential effects across participants, and eliminate many non-experimental sources of bias. Each of these contributions will help reconcile existing disparities in the literature. The indirect contribution will come from building stronger theories.

Theories are successful only to the extent that they can accurately predict the results of many measurements.

In this context, it is important to realize that consensus usually emerges only when the disparate results from previous research can be reasonably reconciled or explained by viable theories. Consensus is generally not achieved by any single “gold standard” experiment alone. Project STAR probably comes the closest to a “gold standard” intervention. But, Project STAR also provided many explanations for the disparity in previous measurements by showing differential effects (larger effects for minority and disadvantaged students), necessary components (3-4 years required for sustained effects), the presence of strong pressure for interference and selectivity (pupils assigned to large classes often made their way to smaller classes), and absence of strong teacher and school contextual effects. All of these helped to explain some of the disparities in previous measurements. Ehrenberg, Brewer, Gamoran, and Willms (2001) and Grissmer (1999) provide examples of integrating Project STAR results with previous class size measurements.

This integration of new multiple methods RCTs results with previous literature requires a thorough knowledge of the strengths and weaknesses of analysis using non-experimental data, data from natural experiments, quasi-experiments and experimental measurements. Three important explanations for why previous results may differ include: (1) measurement bias, (2) the presence of contextual effects, or (3) differences in the characteristics of the population studied. Since the potential for bias usually differs by research methods, one strategy for reviewing studies is to group them by method into experimental, quasi-experimental, “natural” experiments and non-experimental. Webbink (2005) provides an example of this type of review. However, within each of these categories, there is usually wide variation in quality, so that simple categorization often can be misleading. Because of this, the review must assess quality of studies within each category

There is substantial literature that helps with conducting such a critique. Duncan and Gibson-Davis (2006), Duncan and Magnusson (2003), and Duncan et al. (2004) discuss the way that experimental methods address measurement bias issues in non-experimental data. They also argue for the potential of natural experiments. Cronbach and Shapiro (1982) and Heckman and Smith (1995) provide critiques of experimental studies. Cook et al. (2005) compare and contrast experimental and quasi-experimental results. O’Connor (2003) and Rosenzweig and Wolpin (2000) provide a developmental psychology and economics perspective on both advantages and disadvantages inherent in “natural experiments.”

Other useful resources that summarize and interpret schooling effects for adolescents from eight experiments in welfare reform policy include Gennetian, Duncan, Knox, Vargas, Clark-Kauffman, and London (2004). Leventhal and Brooks-Gunn (2003a), Oakes (2004), and Kling et al. (2007) discuss the difficult issues involved in measuring “neighborhood” effects, and contrast findings from experimental and non-experimental studies. Two articles provide some more direct comparison of experimental and non-experimental results using Project STAR data (Krueger, 1999; Wilde & Hollister, 2007).

### **Respond to finding positive effects:**

## **Consider policy implications.**

**Decide whether further scale-up is needed. If so, decide whether replicate studies are needed before going to scale, and what would be the cost, cost-effectiveness and cost-benefit of going to scale.**

The value of positive results from RCTs that do not collect multiple methods data can be degraded significantly if (1) results cannot be generalized to different populations, (2) contextual effects cannot be identified, (3) weaknesses in the design of intervention cannot be identified and improvements suggested, and, (4) the issues in scaling-up to larger programs cannot be addressed. Only multiple methods data can be used to address these issues, and the next steps after garnering positive results is to carefully assess these issues using the multiple methods data. Each of these four issues should be addressed in analyses and publications before proceeding to make decisions on the next steps. Perhaps more importantly, a publication needs to address the implications of the results on our understanding of why and how researchers achieved the effects.

New Hope provides an outstanding example of the additional work and documentation required after obtaining positive effects for both adults and children. Duncan et al. (2007) provide a summary of the many analyses undertaken and target policy-makers who might be interested in undertaking a state-wide or national program. They were able to address these policy issues only because of New Hope's comprehensive multiple methods data collections.

Probably the most difficult challenge is making predictions of costs and effects for a scaled-up program from data/results originally collected in small-scale programs. Schneider and McDonald (2007, vol. I-II) provide a comprehensive assessment of the issues involved in scaling-up education programs. In general, small-scale experiments should not be used as the basis for a major program implementation unless compelling cases can be made that effects and costs will not change in different context or at different scales. In education, the Success for All intervention comes with an interesting history of moving from smaller to larger scale and evaluating the results experimentally (Borman & Hewes, 2002; Borman, Slavin, Cheung, Chamberlain, Madden, & Chambers, 2005, 2007). Efforts to gradually increase implementation of Success for All across many different types of schools allow many of the contextual hypotheses and scaling issues to be tested. Slavin (2002, 2008), Chatterji (2005, 2008) and Briggs (2008) also bring useful perspectives to the question of synthesizing research evidence and deciding when programs should be recommended for wider implementation.

The results from Project STAR had a major impact on class size policies throughout the nation from 1995-2007. This influence was partly due to its experimental design, large sample and the transparency of its findings to policymakers. It also benefited from widespread public support and belief in smaller class sizes, especially when coupled with expanding state revenues. With the exception of California, states implemented smaller classes in a way that took account of Project STAR's findings. Smaller classes were often targeted to minority and disadvantaged children and reductions were usually made for 3-4 years in early grades.

Project STAR benefited from not facing many of the scale-up issues inherent in many other educational interventions. Project STAR was already operating at a large scale -- in 79 schools. Implementation only required finding additional teachers and more classroom space. In general, outside California, implementation was gradual and targeted enough to allow for careful identification of teachers and space. Scale-up was also easier because class size effects in Tennessee were achieved with no additional teacher training. Providing quality training is often a key issue in scale-up. However, it is also possible that teacher training could have enhanced the STAR effects, and could be the focus of additional research.

### **Responding to finding marginal or no effect:**

Null or marginal findings in experiments are often more important in making scientific progress than large effect sizes, particularly when null effects are found where current theories and understanding would have predicted large effects in a given RCT. Such null effects directly undermine current theories and understanding, and present an opportunity to develop new theories and discard old ones. One of the most important experiments in physics was the Michelson and Morley (1887) experiment that measured whether light propagating at right angles held the same speed. The null result paved the transition from Newtonian mechanics to Special Relativity. Multiple methods data are crucial in helping to reject a current theory and move to a new theory by generating an understanding of why assumptions in the old theory are untenable, and what alternative theory might better explain the new results. It is also important to publish null results in the literature since they provide crucial information for individuals involved in theory building. The current bias toward publication of studies with significant effects can meaningfully impair the work of theory development.

Moving from public housing to a higher income neighborhood in MTO was hypothesized to improve adult job opportunities, employability and income, and children's schooling outcomes due to the accessibility of better schools and achieving better parent outcomes. However, researchers found no significant effects on adult employment and income of children's schooling outcomes across the five locations in the experiment (Kling, Liebman & Katz, 2007; Sanbonmatsu et al., 2006). These null effects focused research and additional data collections on teasing out the flaws in the theories that predicted positive effects (Kling, Liebman & Katz, 2005, 2007; Sanbonmatsu et al. Turney et al., 2006). For instance, Turney et al. conducted in-depth interviews with 67 participants in Baltimore to explore why the economic outcomes were insignificant. They state in the study abstract:

The voucher group did not experience employment or earnings gains in part because of human capital barriers that existed prior to moving to a low-poverty neighborhood. In addition, employed respondents in all groups were heavily concentrated in retail and health care jobs. To secure or maintain employment, they relied heavily on a particular job search strategy – informal referrals from similarly skilled and credentialed acquaintances who already held jobs in these sectors. Though experimentals were more likely to have employed neighbors, few of their neighbors held jobs in these sectors and could not provide such referrals. Thus controls had an easier time garnering such referrals (p.137)

Project STAR found experimentally that having teacher aides in grades K-3 had no consistent and significant effect on achievement. Multiple methods data indicated that aides spent only about 25-30% of their time on direct instructional tasks, with the remaining time spent on administrative or non-instructional interactions with students. However, even aides who spent more time on instruction did not lead to effects on achievement. Another hypothesis was that administrative and non-instructional time by teacher aides allowed teachers to be more effective, leading to achievement gains (Gerber et al., 2001). However, multiple methods data suggested that teachers' perceptions of their ability to manage time, cope with student misbehavior or engage students in the learning process was no different for teachers with or without aides (Gerber et al.). Managing aides demands additional teacher time that may reduce teacher productivity. So, the productivity of aides must exceed the possible lost teacher productivity from managing teacher aides to register net gains. These data pointed to an emerging hypothesis that teacher aides had no specific training or educational background that would prepare them for the job. It is also possible that training might be needed in order to help teachers utilize aides effectively. Both of these hypotheses could be pursued through future research.

## References

- Angrist, J. D. (2004). American education research changes tack. *Oxford Review of Economic Policy*, 20(2), 198-212.
- Bernheimer, L. P., Weisner, T. S., & Lowe, E. D. (2003). Impacts of children with troubles on working poor families: Mixed method and experimental evidence. *Mental Retardation*, 41(6), 403-419.
- Blatchford, P. (2003). A systematic observational study of teachers' and pupils' behaviour in large and small classes. *Learning and Instruction*, 13(6), 569-595.
- Blatchford, P. (2005). A multi-method approach to the study of school class size differences. *International Journal of Social Research Methodology*, 8(3), 195-205.
- Blatchford, P., Bassett, P., & Brown, P. (2005). Teachers' and pupils' behavior in large and small classes: A systematic observation study of pupils aged 10 and 11 years. *Journal of Educational Psychology*, 97(3), 454-467.
- Blatchford, P., Bassett, P., Goldstein, H., & Martin, C. (2003). Are class size differences related to pupils' educational progress and classroom processes? Findings from the institute of education class size study of children aged 5-7 years. *British Educational Research Journal*, 29(5, 'In Praise of Educational Research'), 709-730.
- Blatchford, P., Goldstein, H., & Mortimore, P. (1998). Research on class size effects: A critique of methods and a way forward. *International Journal of Educational Research*, 29(8), 691-710.

- Blatchford, P., & Martin, C. (1998). The effects of class size on classroom processes: 'It's a bit like a treadmill - working hard and getting nowhere fast!' *British Journal of Educational Studies*, 46(2), 118-137.
- Bohrnstedt, G. W., & Stecher, B. M. (Eds.). (2002). *What we have learned about class size reduction in California (Capstone Report, Class Size Reduction (CSR) Research Consortium)*. Palo Alto, CA: California Department of Education.
- Bonesrønning, H. (2004). The determinants of parental effort in education production: Do parents respond to changes in class size? *Economics of Education Review*, 23(1), 1-9.
- Booth, A., & Crouter, A. C. (Eds.). (2001). *Does it take a village?: Community effects on children, adolescents and families*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Boozer, M. A., & Cacciola, S. E. (2001). *Inside the 'black box' of Project STAR: Estimation of peer effects using experimental data (Discussion Paper No. 832)*. New Haven, CT: Economic Growth Center.
- Borman, G. D. (2002). Experiments for educational evaluation and improvement. *Peabody Journal of Education*, 77(4), 7-27.
- Borman, G. D., & Hewes, G. M. (2002). The long-term effects and cost-effectiveness of success for all. *Educational Evaluation and Policy Analysis*, 24(4), 243-266.
- Borman, G. D., Slavin, R. E., & Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2005). The national randomized field trial of success for all: Second-year outcomes. *American Educational Research Journal*, 42(4), 673-696.
- Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of success for all. *American Educational Research Journal*, 44(3), 701-731.
- Boruch, R. F. (1997). *Randomized experiments for planning and evaluation: A practical guide*. Thousand Oaks, CA: Sage Publications.
- Bos, H., Duncan, G. J., Gennetian, L. A., & Hill, H. D. (2007). *New Hope: Fulfilling America's promise to "make work pay" (Discussion Paper No. 16)*. Washington, DC: Brookings Institution Press.
- Bosker, R. J. (1998). The class size question in primary schools: Policy issues, theory, and empirical findings from the Netherlands. *International Journal of Educational Research*, 29(8), 763-778.
- Bosworth, R., & Caliendo, F. (2007). Educational production and teacher preferences. *Economics of Education Review*, 26(4), 487-500.

- Brewer, D. J., Krop, C., Gill, B. P., & Reichardt, R. (1999). Estimating the cost of national class size reductions under different policy alternatives. *Educational Evaluation and Policy Analysis, 21*(2), 179-192.
- Briggs, D. C. (2008). Comments on Slavin: Synthesizing causal inferences. *Educational Researcher, 37*(1), 15-22.
- Brock, T. (2005). Viewing mixed methods through an implementation research lens: A response to the New Hope and Moving to Opportunity Evaluations. In T. S. Weisner (Ed.), *Discovering successful pathways in children's development: Mixed methods in the study of childhood and family life* (pp. 317-325). Chicago: University of Chicago Press.
- Brock, T., Doolittle, F., Fellerath, V., & Wiseman, M. (1997). *Creating New Hope: Implementation of a program to reduce poverty and reform welfare*. New York: Manpower Demonstration Research Corporation [MDRC].
- Browning, C. R., & Cagney, K. A. (2003). Moving beyond poverty: Neighborhood structure, social processes, and health. *Journal of Health and Social Behavior, 44*(4), 552-571.
- Burton, P., Goodlad, R., & Croft, J. (2006). How would we know what works? Context and complexity in the evaluation of community involvement. *Evaluation, 12*(3), 294-312.
- Chalmers, I. (2003). Trying to do more good than harm in policy and practice: The role of rigorous, transparent, up-to-date evaluations. *The ANNALS of the American Academy of Political and Social Science, 589*, 22-40.
- Chatterji, M. (2005). Evidence on "what works": An argument for extended-term mixed-method (ETMM) evaluation designs. *Educational Researcher, 34*(5), 14-24.
- Chatterji, M. (2008). Comments on Slavin: Synthesizing evidence from impact evaluations in education to inform action. *Educational Researcher, 37*(1), 23-26.
- Chen, H. T., & Rossi, P. H. (1983). Evaluating with sense: The theory-driven approach. *Evaluation Review, 7*(3), 283-302.
- Clampet-Lundquist, S., Edin, K., Kling, J. R., & Duncan, J. G. (2006). *Moving at-risk youth out of high-risk neighborhoods: Why do girls fare better than boys? (Working Paper No. 509)*. Princeton, NJ: Princeton University, Industrial Relations Section.
- Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis, 25*(2), 119-142.
- Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24*(3), 175-199.
- Cook, T. D. (2003). Why have educational evaluators chosen not to do randomized experiments? *The ANNALS of the American Academy of Political and Social Science, 589*(1), 114-149.

- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin.
- Cook, T. D., Shadish, W. R., & Wong, V. C. (2005). Within-study comparisons of experiments and non-experiments: Can they help decide on evaluation policy? Unpublished manuscript.
- Cooper, C. R. (2005). *Developmental pathways through middle childhood: Rethinking contexts and diversity as resources*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Cooper, C. R., Brown, J., Azmitia, M., & Chavira, G. (2005). Including Latino immigrant families, schools, and community programs as research partners on the good path of life. In T. S. Weisner (Ed.), *Discovering successful pathways in children's development: Mixed methods in the study of childhood and family life* (pp. 359-385). Chicago: University of Chicago Press.
- Cronbach, L. J., & Shapiro, K. (1982). *Designing evaluations of educational and social programs (1st ed.)*. San Francisco: Jossey-Bass.
- Datar, A., & Mason, B. Do reductions in class size “crowd out” parental investment in education? *Economics of Education Review*, In Press, Corrected Proof. Available online February 5, 2008.
- Datta, L. (2005). Mixed methods, more justified conclusions: The case of the Abt Evaluation of the Comer Program in Detroit. In T. S. Weisner (Ed.), *Discovering successful pathways in children's development: Mixed methods in the study of childhood and family life* (pp. 65-83). Chicago: University of Chicago Press.
- Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *The Review of Economics and Statistics*, 86(1), 195-210.
- Donaldson, S. I. (2007). *Program theory-driven evaluation science: Strategies and applications*. New York: Lawrence Erlbaum Associates.
- Duncan, G. J., & Gibson-Davis, C. M. (2006). Connecting child care quality to child outcomes: Drawing policy lessons from non-experimental data. *Evaluation Review*, 30(5), 611-630.
- Duncan, G. J., Huston, A. C., & Weisner, T. S. (2007). *Higher ground: New Hope for the working poor and their children*. New York: Russell Sage Foundation.
- Duncan, G. J., & Magnuson, K. A. (2003). The promise of random-assignment social experiments for understanding well-being and behavior. *Current Sociology*, 51(5), 529-541.
- Duncan, G. J., Magnuson, K. A., & Ludwig, J. (2004). The endogeneity problem in developmental studies. *Research in Human Development*, 1(1), 59-80.
- Duncan, G. J., & Raudenbush, S. W. (1999). Assessing the effects of context in studies of child and youth development. *Educational Psychologist*, 34(1), 29-41.

- Duncan, G. J., & Raudenbush, S. W. (2001). Neighborhoods and adolescent development: How can we determine the links? In A. Booth & A. C. Crouter (Eds.), *Does it take a village? Community effects on children, adolescents, and families* (pp. 105-136). Mahwah, NJ: Lawrence Erlbaum Associates.
- Ehrenberg, R. G., Brewer, D. J., Gamoran, A., & Willms, J. D. (2001). Class size and student achievement. *Psychological Science in the Public Interest*, 2(1), 1-30.
- Eisenhart, M. (2005). Hammers and saws for the improvement of educational research. *Educational Theory*, 55(3), 245-261.
- Eisenhart, M. (2006). Qualitative science in experimental time. *International Journal of Qualitative Studies in Education*, 19(6), 697-707.
- Eisenhart, M., & Towne, L. (2003). Contestation and change in national policy on "scientifically based" education research. *Educational Researcher*, 32(7), 31-38.
- Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. *Educational Researcher*, 31(8), 4-14.
- Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal*, 27(3), 557-577.
- Finn, J. D., & Achilles, C. M. (1999). Tennessee's class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis*, 21(2), 97-109.
- Finn, J. D., Fulton, D., Zaharias, J., & Nye, B. A. (1989). Carry-over effects of small classes. *Peabody Journal of Education*, 67(1, Project STAR and Class Size Policy), 75-84.
- Finn, J. D., Gerber, S. B., Achilles, C. M., & Boyd-Zaharias, J. (2001). The enduring effects of small classes. *Teachers College Record*, 103(2), 145-183.
- Finn, J. D., Gerber, S. B., & Boyd-Zaharias, J. (2005). Small classes in early grades, academic achievement, and graduating from high school. *Journal of Educational Psychology*, 97(2), 214-223.
- Finn, J. D., Panno, G. M., & Achilles, C. M. (2003). The "why's" of class size: Student behavior in small classes. *Review of Educational Research*, 73(3), 321-368.
- Fricke, T. (2005). Taking culture seriously: Making the social survey ethnographic. In T. S. Weisner (Ed.), *Discovering successful pathways in children's development: Mixed methods in the study of childhood and family life* (pp. 185-221). Chicago: University of Chicago Press.
- Gardenhire, A., & Nelson, L. (2003). *Intensive qualitative research: Challenges, best uses, and opportunities* (MDRC Working Paper). New York: MDRC.

- Gennetian, L. A., Duncan, G., Knox, V., Vargas, W., Clark-Kauffman, E., & London, A. S. (2004). How welfare policies affect adolescents' school outcomes: A synthesis of evidence from experimental studies. *Journal of Research on Adolescence*, 14(4), 399-423.
- Gerber, S. B., Finn, J. D., Achilles, C. M., & Boyd-Zaharias, J. (2001). Teacher aides and students' academic achievement. *Educational Evaluation and Policy Analysis*, 23(2), 123-143.
- Gibson-Davis, C. M., & Duncan, G. (2005). Qualitative/quantitative synergies in a random-assignment program evaluation. In T. S. Weisner (Ed.), *Discovering successful pathways in children's development: Mixed methods in the study of childhood and family life* (pp. 283-303). Chicago: University of Chicago Press.
- Goering, J., & Feins, J. D. (Eds.). (2003). *Choosing a better life: Evaluating the Moving To Opportunity social experiment*. Washington, DC: Urban Institute Press.
- Goldenberg, C., Gallimore, R., & Reese, L. (2005). Using mixed methods to explore Latino children's literacy development. In T. S. Weisner (Ed.), *Discovering successful pathways in children's development: Mixed methods in the study of childhood and family life* (pp. 21-46). Chicago: University of Chicago Press.
- Gordon, N. (2004). Do federal grants boost school spending? Evidence from Title I. *Journal of Public Economics*, 88(9-10), 1771-1792.
- Greene, J. (2005). A reprise on mixing methods. In T. S. Weisner (Ed.), *Discovering successful pathways in children's development: Mixed methods in the study of childhood and family life* (pp. 405-419). Chicago: University of Chicago Press.
- Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research*, 66(3), 361-396.
- Grissmer, D. W. (1999). Conclusion: Class size effects: Assessing the evidence, its policy implications, and future research agenda. *Educational Evaluation and Policy Analysis*, 21(2, Special Issue: Class Size: Issues and New Findings), 231-248.
- Grissmer, D. W. (2002). Cost-effectiveness and cost-benefit analysis: The effect of targeting interventions. In H. M. Levin & P. J. McEwan (Eds.), *Cost-effectiveness and educational policy* (pp. 97-108). Larchmont, NY: Eye on Education.
- Grissmer, D. W., & Flanagan, A. (2006). *Improving the achievement of Tennessee students: An analysis of the National Assessment of Educational Progress*. Santa Monica, CA: RAND Corporation.
- Grissmer, D. W., Flanagan, A., Kawata, J. H., & Williamson, S. (2000). *Improving student achievement: What state NAEP test scores tell us*. Santa Monica, CA: RAND Corporation.
- Gueron, J. M. (2002). The politics of random assignment. In F. Mosteller, & R. F. Boruch (Eds.), *Evidence matters* (pp. 15-49). Washington, DC: Brookings Institution Press.

- Gueron, J. M. (2003). Fostering research excellence and impacting policy and practice: The welfare reform story. *Journal of Policy Analysis and Management*, 22(2), 163-174.
- Gueron, J. M., (2007). Building evidence: What it takes and what it yields. *Research on Social Work Practice*, 17(1), 134-142.
- Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis*, 19(2), 141-164.
- Hanushek, E. A. (1999). Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects. *Educational Evaluation and Policy Analysis*, 21(2), 143-163.
- Hanushek, E. A. (2002). Evidence, politics and the class size debate. In L. Mishel & R. Rothstein (Eds.), *The class size debate* (pp. 37-65). Washington, DC: Economic Policy Institute.
- Harkness, S., Hughes, M., Muller, B., & Super, C. M. (2005). Entering the developmental niche: Mixed methods in an intervention program for inner-city children. In T. S. Weisner (Ed.), *Discovering successful pathways in children's development: Mixed methods in the study of childhood and family life* (pp. 329-358). Chicago: University of Chicago Press.
- Hattie, J. (2005). The paradox of reducing class size and improving learning outcomes. *International Journal of Educational Research*, 43(6), 387-425.
- Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *The Journal of Economic Perspectives*, 9(2), 85-110.
- Howe, K. R. (1998). The interpretive turn and the new debate in education. *Educational Researcher*, 27(8), 13-20.
- Howe, K. R. (2004). A critique of experimentalism. *Qualitative Inquiry*, 10(1), 42-61.
- Huston, A. C. (2005). Mixed methods in studies of social experiments for parents in poverty. In T. S. Weisner (Ed.), *Discovering successful pathways in children's development: Mixed methods in the study of childhood and family life* (pp. 305-315). Chicago: University of Chicago Press.
- Huston, A. C., Duncan, G. J., Granger, R., Bos, J., McLoyd, V., Mistry, R., et al. (2001). Work-based antipoverty programs for parents can enhance the school performance and social behavior of children. *Child Development*, 72(1), 318-336.
- Jepsen, C., & Rivkin, S. G. (2002). *Class size reduction, teacher quality, and academic achievement in California public elementary schools*. San Francisco: Public Policy Institute of California.
- Karoly, L. A., & Bigelow, J. H. (2005). *The economics of investing in universal preschool education in California (Monograph No. 349)*. Santa Monica, CA: RAND Corporation.

- Karoly, L. A., Greenwood, P. W., Everingham, S. S., Hoube, J., Kilburn, M. R., & Rydell, C.P., et al. (1998). *Investing in our children: What we know and don't know about the costs and benefits of early childhood interventions (Monograph Report No. 898)*. Santa Monica, CA: RAND Corporation.
- Karoly, L. A., Kilburn, M. R., & Cannon, J. S. (2005). *Early childhood interventions: Proven results, future promise (Monograph No. 341)*. Santa Monica, CA: RAND Corporation.
- Katz, L. F., Kling, J. R., & Liebman, J. B. (2001). Moving to opportunity in Boston: Early results of a randomized mobility experiment. *The Quarterly Journal of Economics*, 116(2), 607-654.
- Kling, J. R., Liebman, J. B., & Katz, L. F. (2005). Bullets don't got no name: Consequences of fear in the ghetto. In T. S. Weisner (Ed.), *Discovering successful pathways in children's development: Mixed methods in the study of childhood and family life (pp. 243-281)*. Chicago: University of Chicago Press.
- Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, 75(1), 83-119.
- Kling, J. R., Ludwig, J., & Katz, L. F. (2005). Neighborhood effects on crime for female and male youth: Evidence from a randomized housing voucher experiment. *The Quarterly Journal of Economics*, 120(1), 87-130.
- Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics*, 114(2), 497-532.
- Krueger, A. B. (2002). Understanding the magnitude and effect of class size on student achievement. In L. Mishel & R. Rothstein (Eds.), *The class size debate (pp. 7-35)*. Washington, DC: Economic Policy Institute.
- Krueger, A. B. (2003). Economic considerations and class size. *The Economic Journal*, 113(485), F34-F63.
- Krueger, A. B., & Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR. *The Economic Journal*, 111(468), 1-28.
- Krueger, A. B., & Whitmore, D. M. (2002). Would smaller class sizes help close the black-white achievement gap? In J. E. Chubb & T. Loveless (Eds.), *Bridging the achievement gap (pp. 11-46)*. Washington, DC: Brookings Institution Press.
- Ladd, H. F., & Ludwig, J. (1997). Federal housing assistance, residential relocation, and educational opportunities: Evidence from Baltimore. *American Economic Review*, 87(2), *Papers and Proceedings of the Hundred and Fourth Annual Meeting of the American Economic Association*, 272-277.
- Lazear, E. P. (2001). Educational production. *The Quarterly Journal of Economics*, 116(3), 777-803.

- Leventhal, T., & Brooks-Gunn, J. (2003a). Children and youth in neighborhood contexts. *Current Directions in Psychological Science, 12*(1), 27-31.
- Leventhal, T., & Brooks-Gunn, J. (2003b). Moving to opportunity: An experimental study of neighborhood effects on mental health. *American Journal of Public Health, 93*(9), 1576-1582.
- Levin, H. M., & McEwan, P. J. (2000). *Cost-effectiveness analysis: Methods and applications (2nd ed.)*. Thousand Oaks, CA: Sage Publications.
- Levin, H. M., & McEwan, P. J. (Eds.). (2002). *Cost-effectiveness and educational policy*. Larchmont, NY: Eye on Education, Inc.
- Lowe, E. D., & Weisner, T. S. (2004). ‘You have to push it—who's gonna raise your kids?’: Situating child care and child care subsidy use in the daily routines of lower income families. *Children and Youth Services Review, 26*(2), 143-171.
- Lowe, E. D., Weisner, T. S., & Geis, S. (2003). *Instability in child care: Ethnographic evidence from working poor families in the New Hope Intervention (Next Generation Working Paper No. 15)*. New York: MDRC.
- Ludwig, J., Duncan, G. J., & Hirschfield, P. (2001). Urban poverty and juvenile crime: Evidence from a randomized housing-mobility experiment. *The Quarterly Journal of Economics, 116*(2), 655-679.
- Lynch, R. G. (2004). *Exceptional returns: Economic, fiscal, and social benefits of investment in early childhood development*. Washington, DC: Economic Policy Institute.
- Masse, L. N., & Barnett, W. S. (2002). A benefit-cost analysis of the abecedarian early childhood intervention. In H. M. Levin & P. J. McEwan (Eds.), *Cost effectiveness and educational policy (pp. 157-173)*. Larchmont, NY: Eye on Education.
- Maxwell, J. A. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher, 33*(2), 3-11.
- Michelson, A. A., & Morley, E. W. (1887). On the relative motion of the earth and the luminiferous ether. *American Journal of Science, 34*(203), 334-345.
- Moses, M. S. (2002). The heart of the matter: Philosophy and educational research. *Review of Research in Education, 26*(1), 1-21.
- Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children, 5*(2, *Critical Issues for Children and Youths*), 113-127.
- Mosteller, F., & Boruch, R. F. (Eds.). (2002). *Evidence matters: Randomized trials in education research*. Washington, DC: Brookings Institution Press.

- Nye, B. A., Hedges, L. V., & Konstantopoulos, S. (2000a). Do the disadvantaged benefit more from small classes? Evidence from the Tennessee class size experiment. *American Journal of Education*, 109(1), 1-26.
- Nye, B. A., Hedges, L. V., & Konstantopoulos, S. (2000b). The effects of small classes on academic achievement: The results of the Tennessee class size experiment. *American Educational Research Journal*, 37(1), 123-151.
- Nye, B. A., Hedges, L. V., & Konstantopoulos, S. (2001). The long-term effects of small classes in early grades: Lasting benefits in mathematics achievement at grade 9. *The Journal of Experimental Education*, 69(3), 245-257.
- Nye, B. A., Hedges, L. V., & Konstantopoulos, S. (2002). Do low achieving students benefit more from small classes? Evidence from the Tennessee class size experiment. *Educational Evaluation and Policy Analysis*, 24(3), 210-217.
- Nye, B. A., Hedges, L. V., & Konstantopoulos, S. (2004a). Do minorities experience larger lasting benefits from small classes? Evidence from a five-year follow-up of the Tennessee class size experiment. *Journal of Educational Research*, 98(2), 94-100.
- Nye, B. A., Konstantopoulos, S., & Hedges, L. V. (2004b). How large are teacher effects? *Educational Evaluation and Policy Analysis*, 26(3), 237-257.
- Oakes, J. M. (2004). The (mis)estimation of neighborhood effects: Causal inference for a practicable social epidemiology. *Social Science & Medicine*, 58(10), 1929-1952.
- O'Connor, T. G. (2003). Natural experiments to study the effects of early experience: Progress and limitations. *Development and Psychopathology*, 15, 837-852.
- Peevely, G., Hedges, L. V., & Nye, B. A. (2005). The relationship of class size effects and teacher salary. *Journal of Education Finance*, 31(1), 101-109.
- Poglinco, S. M., Brash J., & Granger, R. C. (1998). *An early look at community service jobs in the New Hope demonstration*. New York: MDRC.
- Quint, J., Bloom, H. S., Black, A. R., & Stephens, L. (with Akey, T. M.). (2005). *The challenge of scaling up educational reform: Findings and lessons from First Things First (Final Report)*. New York: MDRC.
- Ramey, C. T., Campbell, F. A., Burchinal, M., Skinner, M. L., Gardner, D. M., & Ramey, S. L. (2000). Persistent effects of early childhood education on high-risk children and their mothers. *Applied Developmental Science*, 4, 2-14.
- Raudenbush, S. W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher*, 34(5), 25-31.
- Reynolds, A. J., Temple, J. A., Ou, S. R., Robertson, D. L., Mersky, J., Topitzes, J. W., et al. (2007, March-April). Effects of a preschool and school-aged intervention on adult health and

- well-being: Evidence from the Chicago Longitudinal Study. Paper presented at the biannual meeting of the Society for Research on Child Development, Boston.
- Reynolds, A. J., Temple, J. A., Robertson, D. L., & Mann, E. A. (2002). Age 21 cost-benefit analysis of the Title I Chicago child-parent centers. *Educational Evaluation and Policy Analysis, 24*(4), 267-303.
- Ritter, G. W., & Boruch, R. F. (1999). The political and institutional origins of a randomized controlled trial on elementary school class size: Tennessee's Project STAR. *Educational Evaluation and Policy Analysis, 21*(2), 111-125.
- Romich, J. L. (2006). Randomized social policy experiments and research on child development. *Journal of Applied Developmental Psychology, 27*(2), 136-150.
- Rosenbaum, E. & Harris, L. E. (2001). Residential mobility and opportunities: Early impacts of the Moving to Opportunity Demonstration Program in Chicago. *Housing Policy Debates, 12*(2), 321-346.
- Rosenzweig, M. R., & Wolpin, K. I. (2000). Natural "natural experiments" in economics. *Journal of Economic Literature, 38*(4), 827-874.
- Rutter, M. (2002). Nature, nurture, and development: From evangelism through science toward policy and practice. *Child Development, 73*(1), 1-21.
- Salomon, G. (1991). Transcending the qualitative-quantitative debate: The analytic and systemic approaches to educational research. *Educational Researcher, 20*(6), 10-18.
- Sanbonmatsu, L., Kling, J. R., Duncan G. J., & Brooks-Gunn, J. (2006). Neighborhoods and academic achievement: Results from the Moving to Opportunity Experiment. *Journal of Human Resources, 41*(1), 649-691.
- Schneider, B. L., & McDonald, S. K. (Eds.). (2007). *Scale-up in education*. Lanham, MD: Rowman & Littlefield.
- Schweinhart, L. J. (2004). *The High/Scope Perry Preschool Study through age 40*. Ypsilanti, MI: High/Scope Educational Research Foundation.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education*. Washington, DC: National Academy Press.
- Sims, D. P. (2004). Unintended consequences of education and housing reform incentives. (Master's Thesis, Massachusetts Institute of Technology). Retrieved from Proquest Dissertation Express. (AAT 0807684)

- Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, 31(7), 15-21.
- Slavin, R. E. (2008). Perspectives on evidence-based research in education--what works? Issues in synthesizing educational program evaluations. *Educational Researcher*, 37(1), 5-14.
- Tilley, N. (2004). Applying theory-driven evaluation to the British crime reduction programme: The theories of the programme and of its evaluations. *Criminal Justice*, 4(3), 255-276.
- Towne, L., Shavelson, R. J., & Feuer, M. J. (Eds.). (2001). *Science, evidence, and inference in education: Report of a workshop*. Washington, DC: National Academy Press.
- Turney, K., Clampet-Lundquist, S., Edin, K., Kling, J. R., & Duncan, G. J. (2006). Neighborhood effects on barriers to employment: Results from a randomized housing mobility experiment in Baltimore. In G. Burtless & J. R. Pack (Eds.), *Brookings-Wharton papers on urban affairs: 2006* (pp. 137-187). Washington, DC: Brookings Institution Press.
- Walshe, K. (2007). Understanding what works--and why--in quality improvement: The need for theory-driven evaluation. *International Journal for Quality in Health Care*, 19(2), 57-59.
- Webbink, D. (2005). Causal effects in education. *Journal of Economic Surveys*, 19(4), 535-560.
- Weisner, T. S. (2002). Ecocultural understanding of children's developmental pathways. *Human Development*, 45(4), 275-281.
- Weisner, T. S. (2005). *Discovering successful pathways in children's development: Mixed methods in the study of childhood and family life*. Chicago: University of Chicago Press.
- Weiss, H. B., Kreider, H., Mayer, E., Hencke, R., & Vaughan, M. (2005). Working it out: The chronicle of a mixed methods analysis. In T. S. Weisner (Ed.), *Discovering successful pathways in children's development: Mixed methods in the study of childhood and family life* (pp. 47-64). Chicago: University of Chicago Press.
- Wilde, E. T., & Hollister, R. (2007). How close is enough? Evaluating propensity score matching using data from a class size reduction experiment. *Journal of Policy Analysis and Management*, 26(3), 455-477.
- Wilson, W. J. (1996). *When work disappears: The world of the new urban poor* (1st ed.). New York: Knopf.
- Yoshikawa, H., Weisner, T. S., Kalil, A., & Way, N. (2008). Mixing qualitative and quantitative research in developmental science: Uses and methodological choices. *Developmental Psychology*, 44(2), 344-354.
- Yoshikawa, H., Weisner, T. S., & Lowe, E. D. (Eds.). (2006). *Making it work: Low-wage employment, family life, and child development*. New York: Russell Sage

## Appendix – Project Descriptions

### Project Star

The multi-district Tennessee STAR experiment randomly assigned students in a single cohort of kindergarten students (the 1986-87 entering cohort) in 79 participating schools that each had at least three kindergarten classrooms. The students were assigned to one of three groups: (1) large classes (approximate mean of 22–24 students) *with a teacher aide*, (2) large classes *without a teacher aide*, or (3) small classes (approximate mean of 15–16 students). Kindergarten teachers in each school were also assigned randomly to one of these types of classrooms. The structure of this design meant that each school constituted a separate random experiment.

The sample of entering students across 328 kindergarten classes was approximately 6,500 students. Those students entering at kindergarten were scheduled to maintain their treatment through first, second, and third grade. However, the groups changed in significant ways over the four years of the experiment due to sample attrition and new students entering schools during kindergarten, 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> grade. Students entering participating schools after kindergarten entrance were also randomly assigned to one of the three groups, but these late entering students came from schools outside the sample and likely had been in larger classrooms prior to entering a participating school. Many students originally in participating schools also moved away after 1, 2 or 3 years. As a result, over the course of the experiment almost 12,000 children participated.

Students leaving the sample and those entering later generally had lower scores on standardized math and reading tests than those who began and remained in the original entering cohort. The 12,000 study subjects therefore consisted of some who remained in the sample all four years, some who entered the school later and had fewer than 4 years in an assigned treatment group, and some who left the sample after kindergarten entry and completed fewer than four years in a treatment or control group. There were also some crossovers (about 15 percent) who switched at some point from one treatment group to another.

In addition to mathematics and reading tests administered in each grade, teachers and aides completed questionnaires and time logs to document their perceptions and experiences. In the Grade 4 follow-up study, researchers collected behavior data in addition to achievement scores. Grade 4 teachers rated each pupil who had been in STAR using a 28-item Student Participation Questionnaire. This instrument assesses specific learning behaviors ("engagement behaviors") judged by educators to be important in the classroom. The instrument yields reliable, valid measures of the effort students allot to learning, initiative-taking in the classroom, and non-participatory behavior (disruptive or inattention).

Researchers also conducted follow-up measurements with the students at 4<sup>th</sup> and 8<sup>th</sup> grade and after high school. At 4<sup>th</sup> and 8<sup>th</sup> grade, they collected reading and math assessment data. In high school, measurements included college entrance test taking (SAT and ACT)) and whether students completed high school.

The major results from Project STAR include the following:

- Students assigned to smaller classes all four years had statistically significant achievement gains of about .15 to .40 standard deviations above the mean of students assigned to the two large class groups (with and without teacher aides).
- Gains in reading were not significantly different from gains in mathematics.
- The effect of teacher aides in large classes was small and positive, but statistically insignificant when compared with large classes without aides.
- Effect sizes were much larger for minority and disadvantaged students in small classes.
- Students assigned to small classes for 3-4 years had much higher gains than those in small classes for only 1-2 years.
- Significant effects persisted through 8<sup>th</sup> grade for students who had participated in smaller K-3 classes, though effect sizes declined somewhat from effect sizes at 3<sup>rd</sup> grade, and effects had greater persistence through 8<sup>th</sup> grade if students had more years in smaller K-3 classes.
- Students in small classes in K-3 had higher high school graduation rates, and increased incidence of testing linked to college applications (SAT and ACT).

## **New Hope**

New Hope was an ambitious project based on two simple, yet widely held principles: (1) people who are willing to work full-time should be able to do so, and (2) they should not be poor as a result. The program was designed to improve the lives of low-income individuals and families by providing several benefits for parents who worked full time: an earnings supplement to raise their income above poverty, subsidized health insurance, and subsidized child care. The program also offered access to wage-paying community service jobs for people who could not find full-time work.

New Hope was run as a demonstration project from 1994 to 1998 in two inner-city areas of Milwaukee, Wisconsin by the New Hope Project, Inc., a local community-based organization. The researchers targeted New Hope at two geographic areas with high levels of poverty, thus allowing a more detailed analysis of program context than would be possible in a program that served a wide geographic area.

The program had only four eligibility requirements: that an applicant live in one of the two targeted service areas, be age 18 or older, be willing and able to work at least 30 hours per week, and have a household income at or below 150 percent of the federally defined poverty level. Participation was voluntary, and adults were eligible regardless of whether they had children or whether they received public assistance.

Persons who met these criteria were eligible to receive three years of these benefits or services:

- Help in obtaining a job, including access to a time-limited, minimum-wage community service job (CSJ) if full-time employment was not otherwise available;
- A monthly earnings supplement that when combined with federal and state Earnings Income Credit (EICs) brought most low-wage workers' incomes above the poverty level;
- Subsidized health insurance, which gradually phased out as earnings rose;
- Subsidized child care, which also gradually phased out as earnings rose.

The study enrolled over 1,300 low-income adults who volunteered to participate. Half the applicants were randomly assigned to a program group that was eligible to receive New Hope's benefits, and the other half were randomly assigned to a control group that was not eligible for the enhanced benefits. From the total sample of 1,357 people, 745 people had at least one child between the ages of one and ten at the time of enrollment.

New Hope program data provide information on parents' use of the program's services, as well as their job status, hours worked and earnings. State administrative records provide data on employment and receipt of welfare and food stamp benefits. Researchers collected primary data in the form of interviews and surveys at two, five and eight years from the beginning of the experiment. In-person surveys revealed information on families' receipts of New Hope benefits, job histories, parents' employment and earnings, family functioning and parent-child relations. For up to two "focal" children in each family, the surveys also collected information from parents, teachers and children on school performance, psychological well-being and behavior problems. At the five year interview stage, children took standardized tests. Parents were also asked about stress levels, depression and their hopes for the future. Parents and children reported on parent-child relationships, children's experience in child care and activities outside school.

In order to better understand the detailed dynamics and contexts of family life, fieldworkers drew an ethnographic sample of 44 families from the total sample of participants. They gave these families — half of whom were in the New Hope group and half of whom were in the control group — periodic in-depth interviews from the third year to the final year of the New Hope program (1998-2001) and again in 2004. The ethnographic data include extensive field notes as well as focused interviews covering a wide range of topics, including, for example, parents' experiences with New Hope, family routines, work experiences, family relationships, child care arrangements and goals. Unlike surveys, these open-ended interviews and conversations allowed participants the opportunities to tell their stories. Families did not shy away from talking about difficult issues — domestic abuse, drugs and alcohol, family conflicts and health problems. In addition to conducting interviews, the ethnographic fieldworkers participated in family routines and events including lunches, dinners, birthday parties and trips to the mall.

Key New Hope differences between treatment and control groups included:

- There was varying impact of New Hope on work and earnings across study sub-groups.
  - For individuals working little or not at all at the beginning of the program, New Hope led to more work and higher earnings during the four years of operation, but did not have persisting significant effects after the program ended;
  - For individuals already working full-time at the beginning of the program, the program showed no effects on long term work or income;
  - The effects on work and earnings were significant and persisted after the program ended for some individuals whose barriers to employment were addressed by New Hope benefits (e.g. child care or health insurance);
  - For women without children at the beginning of the program, there were no work or earnings effects;
  - For men without children at the beginning of the program, there were boosts to work and income, but only sporadically; and

- Partly due to income supplements, New Hope reduced poverty substantially during and modestly after the end of the program.
- New Hope child care subsidies increased children’s participation in center-based child care and after school programs.
- New Hope insurance benefits led to fewer episodes of unmet medical and dental needs, and some improvement in adult mental and physical health.
- New Hope improved children’s school performance, especially in reading.
- For boys, New Hope led to increased positive social behaviors and reduced behavior problems, and increased engagement in school and higher education.
- For girls, New Hope had mixed effects. Parents reported improvements in their daughters’ positive behaviors, but teachers reported worse behavior for those same girls at school.

## **Moving to Opportunity (MTO)**

The MTO demonstration program was designed to assess the impact of providing families living in subsidized housing in high poverty neighborhoods with the opportunity to move to neighborhoods with lower levels of poverty. Families were recruited for the MTO program from public housing developments in Boston, Baltimore, Chicago, Los Angeles and New York. Researchers primarily targeted housing developments located in census tracts with 1990 poverty rates of at least 40 percent. The average poverty rate in these tracts in 1990 was 67 percent.

Program eligibility requirements included residing in a targeted development, having very low income that met the Section 8 income limits of the public housing authority, having a child under eighteen, and being in good standing with the housing authority. Participants volunteered to be part of the study. Families that volunteered for the program were more disadvantaged than their public housing counterparts who did not join MTO. MTO families were more likely than nonparticipating families to receive welfare and to be headed by women who were young and unemployed.

Volunteering families initially living in public housing were assigned by lottery to three groups:

- *Control group*. Received no new assistance, but continued to be eligible to stay in public housing.
- *Section 8 group*. Received traditional Section 8 voucher that enabled movement from public housing to subsidized rental housing without geographic restriction.
- *Experimental group*. Received Section 8 voucher, restricted for one year to a census tract with a poverty rate of less than 10 percent.

From 1994-1997, 4,248 eligible families were randomly assigned to one of these three groups. Families in the treatment groups had 4–6 months to find qualified housing and move, using a MTO voucher. Forty-seven percent of the experimental group families and 59 percent of the Section 8 group families used the program housing voucher to “lease-up,” or move to a new apartment.

Baseline interviews with heads of households were conducted from 1994 to 1999, before random assignment and relocation of movers. The structured interviews focused on demographic

information for householders and children and data from householders on labor force and welfare benefits characteristics.

Researchers supplemented MTO baseline surveys with state administrative earnings and welfare data. In 2002, four to seven years after enrollment, researchers surveyed all of the household heads in the experiment, as well as school-aged children and teens in each family. They collected more comprehensive measures related to economic self-sufficiency, and mental and physical health outcomes, as well as a broader range of mediating factors to potentially illuminate the mechanisms by which residential neighborhoods may affect economic and health outcomes. In addition, there were several specialized data collections conducted with sub-samples of participants. These included a subsample of children who were administered achievement tests. Researchers obtained juvenile arrest records, as well as “qualitative” interview data through in-depth personnel interviews and telephone conversations with teens and adults.

Results from the analyses of these data included the following differences between the three groups:

- There were no significant effects among the three groups on measures of work or earnings;
- There were no significant effects among groups on children’s achievement;
- There were significant positive effects on some measures of adult and child mental health for the experimental group;
- Boys in the experimental group fared no better or worse on measures of risk behavior than boys in the control group;
- Girls in the experimental group had improved mental health and lower risk behavior than girls in the control group; and
- Adults in the experimental group had reductions in obesity, but no effects on other physical health measures.