

5. Conclusions and Recommendations

THERE IS A GENERAL CONSENSUS in the education research community on the need to increase the capacity of researchers to study educational problems scientifically. This report considers key issues involved in selecting research designs that allow investigators to draw valid causal inferences about treatment effects using large-scale observational datasets. It addresses why issues of establishing causal inference are of particular interest to education researchers, provides a brief explanation of how causality is commonly defined in the literature, and describes some of the tools that analysts use to approximate randomized experiments with observational data. The report also reviews four studies funded by NSF that illustrate the difficulties of and possibilities for making causal inferences when conducting studies focused on significant educational issues. These studies and other examples provided in this report are intended to help researchers and policymakers understand the strengths and weaknesses of various research designs and methods.⁶⁹

Government funding agencies in the United States and elsewhere are at a critical juncture as they seek to determine what types of research studies to fund in an era of declining

resources. At the same time, such agencies are faced with educational problems that have thus far proved intractable, such as closing the achievement gap between racial and ethnic groups with varying economic and social resources. As NSF, IES, and other government agencies review their portfolios and decide where they need to allocate scarce resources, we make the following suggestions.

Forming an Evidential Base With Observational Designs

National longitudinal datasets such as ECLS and the National Education Longitudinal Study of 1988–2000, designed and administered by NCES, are extremely useful sources of data for investigating educational problems and formulating policies. These datasets, constructed with stratified random samples based on population estimates, provide some of the most robust indicators of how students are performing academically and allow for exploratory analyses regarding why some children are more successful in school than others. The large samples on which these studies are based facilitate comparisons across various subgroups using measures such as age, gender, race/ethnicity, and social and economic resources. These datasets are widely accessible to researchers, enhancing capacity for replicating and extending findings to specific populations and settings. An additional benefit of these datasets is that they can be linked to other national datasets, including census information, facilitating the examination of neighborhood effects on achievement, school access, and resource inequities. Research based on these datasets has had a significant impact on our understanding of teacher effects on instruction, classroom resources that positively affect student learning, factors associated with dropout rates, high school graduation rates, postsecondary matriculation, and relationships between school organizations (sector effects, charter schools, and magnet schools) and student achievement. Secondary analyses of

educational datasets, particularly those that contain information from students, parents, and teachers within institutions over time, continue to serve as one of the richest sources for evidence-based educational policy evaluation.⁷⁰

Through statistical techniques, large-scale datasets can approximate some of the probable causes and effects that experiments can establish more conclusively. Analyses of large-scale datasets are particularly valuable when experiments are impossible or impractical, such as when examining the effect of corporal punishment on student learning. However, even with these data, which arguably are among the best we have, the findings have not consistently yielded information that could substantially improve our schools and change the educational opportunities of students, especially those who attend high-poverty schools and whose families have limited economic and social resources.

Certainly, these large-scale datasets could be more useful if the design and instruments were determined by some of the leading experts in the field. Often the design and instruments of longitudinal and other national large-scale studies are determined by precedent or produced within a short period so that careful review, discussion, and consideration of possible innovation are less likely to occur. In the instance of assessing effective accountability measures as identified in NCLB, a multipurpose longitudinal study could be conducted by embedding controlled field trials within a conventional stratified random sample of school districts that included an oversample of low-performing school districts.

In this report, we have highlighted how, with appropriate methods, observational datasets can be used to approximate randomized assignment to treatment and control conditions. Large-scale datasets have been somewhat underutilized for this purpose, and we encourage NSF, IES, and other funding agencies to promote studies that continue to explore and develop methodologies for approximating randomized experiments,

support work that is designed to undertake such analyses, and recognize the importance of these studies for testing hypotheses, designing subsequent experiments, and measuring contextual effects.

There are tradeoffs between experiments and analyses of observational data. Kish (1987) observed that what makes an experiment especially powerful is that the conditions are tightly controlled. Well-designed experiments maximize internal validity, whereas nationally representative observational datasets maximize external validity (Campbell & Stanley, 1963). Both are important. As Hedges recently commented, randomized controlled trials are particularly efficient in measuring main effects (Hedges, 2004). However, analyses of observational datasets may be beneficial for estimating contextual conditions such as classroom composition or school organizational practices that may be indirectly influencing the effect of a specific intervention.

Education research is facing new challenges and opportunities due to the confluence of high expectations and new methodologies and datasets. It is important to underscore that the types of research questions addressed by education research projects should be of first concern and that appropriate methods should be employed to answer these questions more definitively. Researchers should be encouraged to investigate questions that deliberately test theories of practice and to obtain empirical data to examine rival explanations for behavior. To do so requires developing a portfolio that tests specific hypotheses about educational practice, tailors research questions to address the effects of programs and practices on specific populations, and, most important, derives frameworks and theoretical approaches that address questions of causal effects and the multiple methods that can be used to examine such questions.

Assessing the Relative Strengths of Experimental and Quasi-Experimental Designs

In deciding which proposals best address the research questions of interest to a funding agency, it is important to develop decision rules for evaluating the quality of proposed research. Below, we identify several criteria for evaluating the appropriateness and strengths of various research designs for investigating the effects of particular interventions.

Are randomized controlled trials a feasible design for addressing the research question(s)? If not, can treatment and control groups be identified using existing large-scale observational datasets? If so,

- How reliable are the measures?
- Does the research design include identification of possible causal mechanisms?
- Does the design specify the investigation of treatment effects for different populations of students?
- Does the design allow investigators to take into account the nested quality of educational settings? For example, are treatment effects examined at various levels of the educational system (e.g., classroom, school, and school district)?
- Has the researcher proposed an appropriate quasi-experimental design, such as propensity score matching?

The What Works Clearinghouse has developed a set of decision rules that can be used to assess the strength of quasi-experimental designs. These include criteria for classifying experimental and quasi-experimental research designs and for determining the strength of various designs with respect to drawing valid causal inferences. Other important characteristics of studies that should be evaluated are also identified, including (a) intervention fidelity; (b) outcome measures;

(c) the extent to which relevant people, settings, and measures are included in the study; (d) the extent to which the study allows for testing of the intervention's effect within subgroups; (e) statistical analysis; and (f) statistical reporting (see What Works Clearinghouse, <http://www.whatworks.ed.gov>).

If a randomized experiment cannot be approximated using a rigorous method such as propensity score matching, can an alternative method, such as a regression discontinuity design, a fixed effects model, or an instrumental variable approach, be used? If so,

- For a regression discontinuity design, can the investigators indicate how they will establish whether individuals just above and below the cutoff point for program entry have similar characteristics and probabilities of being accepted into the program? For example, are appropriate matching procedures proposed?
- If a fixed effects approach is proposed, do the investigators provide a clear rationale for treating a variable or variables as fixed, or time-invariant?
- If an instrumental variable is used, do the investigators provide a clear rationale for its selection?
- Do the investigators propose appropriate statistical techniques for comparing treatment group outcomes, fixing effects, or correctly implementing the instrumental variable?
- If a regression discontinuity design, fixed effects model, or instrumental variable approach cannot be used, what methods are proposed for correcting for selection bias and controlling for potentially confounding variables? Have the investigators clearly indicated the strengths and limitations of these methods?

In some cases it may be possible to use both experimental and quasi-experimental designs to address a particular research question. In such cases, funding agencies need to assess the relative strengths and weaknesses of the proposed designs with respect to (a) their potential for producing unbiased estimates of treatment effects, (b) possible difficulties that might arise in implementing the designs, and (c) the cost of each type of study.

If the design is experimental,

- Can investigators recruit a sufficient number of participants (e.g., school districts, schools, teachers, and students) to conduct the study?
- Can the study be implemented with fidelity? Are steps proposed for monitoring implementation to identify problems that may arise in designing and fielding the study (e.g., unsuccessful randomization, insufficient sample sizes for detecting treatment effects, movement of students between treatment and control conditions when individuals are randomized to treatment and control conditions, and differential attrition)?

If the design is quasi-experimental and propensity score methods are used,

- Are available measures comprehensive enough to create an aggregate variable for purposes of computing propensity scores?
- Is there sufficient overlap in the pretreatment characteristics of the treatment and control groups to warrant further analyses?
- Can students in the treatment and control groups be matched with respect to pretreatment characteristics so as

to create equivalence in pretreatment characteristics within propensity score strata?

- Is the analytic sample large enough to detect treatment effects?

If both experimental and quasi-experimental designs are feasible and potential problems in implementation can be adequately addressed, then the decision regarding which design to implement may depend on the cost of each study. In the case of large-scale RCTs, these costs can be considerable. A well-designed quasi-experimental study using an existing large-scale national longitudinal dataset would generally be much less costly to implement. Using a quasi-experimental design does not preclude following it with a more tightly controlled experiment (i.e., it is not necessary to choose one or the other).

If existing observational datasets do not contain sufficient information for conducting a well-designed quasi-experiment, then funders should consider developing a study that builds on the strengths of both designs. For example, it may be possible to embed a multi-site randomized controlled trial within a large-scale longitudinal study based on a nationally representative sample of students, teachers, and schools, with an oversampling of low-income and minority students, or other groups most likely to benefit from a particular intervention.

Sustaining a Program of Evidential Research

In the past, NSF and other governmental agencies and private foundations have funded few randomized controlled trials in education. The enactment of the No Child Left Behind Act, in conjunction with other evidence-based movements internationally, raised awareness of the importance of conducting RCTs, particularly in education (Schneider, Kertcher, &

Offer, 2006). The importance of RCTs is clear, and it seems important that NSF and government agencies that fund education research develop and support a coherent and sustained program of experimentation to complement qualitative data on best practices such as interviews and classroom observations and descriptive and quantitative data on teacher quality, instructional practices, and student and teacher characteristics obtained from large-scale observational studies such as ECLS.

As the NRC's Committee on Scientific Research on Education makes clear in *Scientific Research in Education* (2002), the question of causal effects is but one of three general questions that drive research. This report has focused on how to establish that there is an effect (i.e., "Is there a systematic effect?"). What has been less emphasized are the two other questions identified by the NRC: (1) "What is happening?" (i.e., what is occurring in a particular context, usually documented through thick description); and (2) "Why or how is it happening?" (i.e., What mechanisms are producing the effect that is observed?). These two questions are central to the design of experiments and their usefulness. They are also important for developing theories of cognition, learning, and social and emotional development. A program of evaluation built on a solid foundation of closely linked research using a variety of methods is needed to establish the basis for reliable and enduring knowledge about the effects of educational innovations.

The recent review of NSF's portfolio of mathematics projects provides a window into the research priorities of a specific program within a federal funding agency. This report concluded that NSF-funded projects in this area tend to focus on designing and implementing new interventions, tools, and methods, but are much less likely to address basic problems of teaching and learning or to synthesize results and identify new questions (NSF, 2004). Although NSF projects that

focus on the design and implementation of new interventions or methods often include an evaluation component, project quality or effectiveness seldom has been evaluated using rigorous experimental and quasi-experimental designs. The challenge for funding agencies such as NSF is to develop a culture of both development and evaluation—one that attends to all points of the cycle of discovery, innovation, and application. In this report we recommend that researchers be required to discuss more directly their hypotheses and models of educational practice. Proposed research programs should answer questions about what mechanisms are important and how practitioners can apply the results of research or evaluation.

Considerable resources are currently available to help funding agencies and researchers evaluate the strengths of different study designs and to develop better-designed experiments and quasi-experiments. We have attempted to add to these resources by providing decision rules specific to the evaluation of studies based on large-scale, nationally representative datasets. Although embedding RCTs within future national longitudinal studies would strengthen the design of such studies, existing large-scale datasets remain a rich resource for descriptive statistics on nationally representative samples of students and subgroups (e.g., low-income and minority students); for identifying potential causal effects and mechanisms; and for providing valid evidence of causal effects through the use of rigorously designed quasi-experiments. These datasets have been underutilized for purposes of study replication. Properly analyzed, they present cost-effective alternatives for addressing causal questions about the effectiveness of educational interventions. The methods described for approximating randomized controlled experiments underscore the value of these datasets for generating and informing educational policy and practice.

Chapter 5 Notes

- 69** This report is not intended to be a “how to” manual for designing research studies or analyzing experimental or observational data. A number of resources are currently available for helping education researchers develop and implement well-designed randomized controlled experiments or quasi-experiments. We would recommend that Shadish, Cook, and Campbell (2002) be one of the first sources consulted.
- 70** Although most of the national datasets are unusually broad in scope, analytic limitations exist even among datasets that have hundreds of variables, many of which can be triangulated across different respondents. Common problems that researchers encounter with these datasets are missing data. Fortunately, researchers now have sophisticated techniques for imputing missing data (see, e.g., King, Honaker, Joseph, and Scheve, 2001, and Little and Rubin, 2002, on multiple imputation techniques). Similarly, Institutional Review Boards and government agencies are finding ways to secure confidentiality so that researchers can now link different datasets or create equating assessment protocols that allow them to identify and use similar variables across local, state, and national datasets.