HAMMERS AND SAWS FOR THE IMPROVEMENT OF EDUCATIONAL RESEARCH

Margaret Eisenhart

School of Education
University of Colorado at Boulder

Quarrels over which method represents "the gold standard" make no more sense than arguing about whether hammers are superior to saws. The choice depends on whether you want to drive in a nail or cut a board.[1]

Determining causation is a fixation in U.S. society. Educational researchers are no exception. We are desperate to know what events and processes lead to what educational outcomes, so that we can promote the outcomes we want and eliminate the ones we do not want. I would venture to say that there is no educational researcher, parent, administrator, teacher, or government official who does not value the study of causation in this sense and, further, that the purpose of almost all educational research is to learn something about causation in general.

Disputes arise because not all educational researchers are doing the same thing when they search for causes. Metaphorically, some are hammering a nail while others are sawing a board. Quantitative researchers, for example, tend to investigate whether a change in one variable, *x*, causes a change in another variable, *y*. This approach has been called the variance, or regularity, theory of causation.[2] Variance theory addresses questions about causation by showing a correlation between an earlier event and a subsequent one. Causal processes — the means by which *x* is linked to *y* — are not directly investigated. Experimental and quasi-experimental research methods are appropriate for answering questions about cause framed in this way.

Qualitative researchers, in contrast, tend to investigate causal processes; that is, they are interested in *how x* influences *y*. This approach has been referred to as the process, or realist, theory of causation.[3] It relies on descriptions of a demonstrable sequence of events by which *x* flows into or leads to *y*. Ethnographies, case studies, discourse analyses, and narrative analyses are appropriate for answering questions about cause framed in this way.

These two theoretical approaches are not fundamentally different — both are means to investigate causation — but each is distinctive in approach and practice. They lead to different kinds of research questions and, in turn, to different research

1. Lisbeth B. Schorr and Daniel Yankelovich, "What Works to Better Society Can't Be Easily Measured," *Los Angeles Times*, B7, February 16, 2000.

2. Joseph A. Maxwell, *Qualitative Research Design: An Interactive Approach* (Thousand Oaks, California: Sage, 1996), 20–21; Joseph A. Maxwell, "Causal Explanation, Qualitative Research, and Scientific Inquiry in Education," *Educational Researcher* 33, no. 2 (2004): 3–5.

3. Ibid.

EDUCATIONAL THEORY | Volume 55 | Number 3 | 2005

designs and methods. The U.S. federal government currently privileges the first approach — variance theory and the experimental designs that so often go with it — as the "gold standard" for educational research. It is being encouraged and funded to the near-exclusion of the other approach.[4] In this article, I argue that just as both hammers and saws are needed to build a good house, both variance and process approaches are needed to build a good understanding of causation in education.

My position stands in contrast to the polarization being created by efforts to impose a gold standard on educational research, on the one hand, and to oppose such a standard, on the other. I do not think any thoughtful researcher today believes that experiments or randomized field trials are the "gold standard" for addressing all the important questions in educational research. Yet, because these designs are now required by the 2001 No Child Left Behind Act (NCLB) and are being strongly encouraged in other federal legislation and funding initiatives, scholars, practitioners, parents, and researchers must devote time and energy to fighting these designs when they are inappropriate or irrelevant, which is often the case. Despite long-standing objections from prominent methodologists and reservations expressed by national groups and committees, key policymakers in the federal government are encouraging the pursuit of experimental designs primarily or exclusively.[5] Educational scholars and researchers must counter this agenda, not by revisiting tired debates (such as the experimental versus qualitative debates, or the positivist versus interpretivist versus constructivist debates), but by underscoring the need for and the value of multiple and integrated approaches to the research questions that concern us the most. Surely questions about causation fit this criterion.

In this essay I discuss four research designs for pursuing questions about causation in education. Two of them take a variance approach; the other two take a process approach. The point of my discussion is to illustrate, first, their respective strengths and, second, their necessary interdependence.

4. For evidence of this, see the *No Child Left Behind Act of 2001*, Public Law 107-110 (HRl), http://www.ed.gov/nclb/landing.jhtml?src = pb; Institute of Education Sciences, "Predoctoral Interdisciplinary Research Training Program in the Education Sciences," (Request for Applications NCER-04-06, http://www.ed.gov/programs/edresearch/); and Joshua D. Angrist "American Education Research Changes Tack," *Oxford Review of Economic Policy* 20, no. 2 (2004): 201.

5. For examples of the methodologists' objections, see Donald T. Campbell, "Can We Be Scientific in Applied Social Science?" *Evaluation Studies Review Annual* 9 (1984): 26–48; Tom Cook, "A Discussion with Tom Cook" (roundtable presentation at the American Educational Research Association, Chicago, Illinois, November 1998); and Robert B. McCall and Beth L. Green, "Beyond the Methodological Gold Standards of Behavioral Research: Considerations for Practice and Policy," *Social Policy Report* 18, no. 2 (2004): 3–19. For examples of objections raised by national groups and committees, see American Educational Research Association, "AERA Advocates for Sound Science" (resolution passed January 26, 2003, http://aera.net/meeting/science.htm); and National Research Council, *Scientific Research in Education*, eds. Richard J. Shavelson and Lisa Towne (Washington, D.C.: National Academies Press, 2002).

MARGARET EISENHART is Distinguished Professor in the School of Education at the University of Colorado, Boulder, 249 UCB, Boulder, CO 80309-0249; e-mail <margaret.eisenhart@colorado.edu>. Her primary areas of scholarship are anthropology of education, ethnographic research methods, and women's studies.

EXPERIMENTAL STUDIES OF CAUSATION IN EDUCATION

The most familiar variance theory strategy for answering research questions about causation is experimental design. The most important point made by current proponents of this strategy is that experiments are powerful research tools that are too infrequently used in educational research. Although I do not think that experiments constitute a gold standard for educational research, I believe that they do have an important role to play. When it is possible and ethical to use experiments, they have the potential to yield useful results about relationships among variables. Consider the following scenario described in an article by Jeff Valentine and Harris Cooper:

> Imagine you are a parent reading a letter from your local PTA, suggesting that the PTA… recommend to the school board that it adopt a "school uniform" policy for your child's school.…[The] PTA letter asserts that research evidence suggests school uniforms would have a beneficial impact on student achievement and conduct. Being trained in [education] you decide to do some research on your own.…After examining the literature, you find that the studies vary in their quality (or, the confidence that you have in the validity of their conclusions).[6]

In this situation, the nature of the relation (if any) between school uniforms and school achievement is not obvious, and the research evidence is mixed, yet your school is planning to require uniforms. It is reasonable to ask whether requiring school uniforms is a good decision for a school wishing to improve achievement: How strong is the evidence of a connection between uniforms and achievement? Is this a cost-effective use of policy and resources for a school wishing to improve achievement? Given the goal of raising achievement and the reality of limited resources, is a uniform requirement likely to contribute positively toward reaching the goal?

This is the kind of public interest context in which the desire for more experimental research makes sense. Research questions such as Does the adoption of school uniforms contribute to improved school achievement? Is a proposed new reading program likely to improve reading scores? Does class-size reduction have a beneficial impact on student learning? call for studies that can test the relation between input and outcome variables. Well-designed experimental studies of such research questions can help teachers, parents, and policymakers separate programs based on opinion or fad from programs with demonstrated effects. They can also help consumers identify the conditions that are likely to be necessary to realize desired effects. This is worthwhile information to have about educational programs, and as numerous proponents of experimental research have pointed out, there have been relatively few well-designed experimental studies in education.[7]

6. Jeffrey C. Valentine and Harris Cooper, "Scaling the Quality of Causal Research in Education," in *Empirical Methods for Evaluating Educational Interventions*, eds Gary D. Phye, Daniel H. Robinson, and Joel Levin (San Diego: Academic Press, 2005), 2.

7. Herbert Turner, Robert Boruch, Anthony Petrosino, Julia Lavenberg, Dorothy de Moya, and Hannah Rothstein, "Populating an International Web-Based Randomized Trials Register in the Social, Behavioral, Criminological, and Education Sciences," *Annals of the American Academy of Political and Social Sciences* 589 (2003): 203–223.

But even those researchers who are staunch proponents of experimental designs know that they cannot do everything, nor are they perfect. For instance, in considering the school uniform scenario, Valentine and Cooper continue:

> [If] a principal asks, "If our school introduces a school uniform policy, can we expect achievement levels to rise?" — then experiments provide the best way of answering that question. [But not] all causes we are interested in can be manipulated, nor do all interesting questions pertain to causality. As a result, we are not arguing that experiments are the only way to acquire valid knowledge, nor do we mean to suggest that random assignment is synonymous with the scientific method. Rather, we believe there is a need to strike a balance between these indefensible positions and their equally indefensible opposites, specifically, denials that experimentalism in social science [or education] can result in any knowledge at all. Thus, we believe experimental research in social science should focus on a specific goal: to try to determine if some intervention causes changes in some outcome.[8]

Pursuing this focused goal is also the rationale for current initiatives such as the Campbell Collaboration and the What Works Clearinghouse, which are being supported by the government's new Institute for Education Sciences (IES).[9] Both initiatives are attempts to stimulate more high-quality experimental or quasi-experimental studies of educational interventions to test claims about the effects of specific programs and to encourage larger numbers of these studies so that systematic reviews and meta-analyses (summaries and syntheses of results across studies conducted in different places and with different participants) can provide information about the generalizability of results.

Researchers who promote more experimental designs in education are not blind to their problems. They are aware of how the characteristics of research sites, the limitations of measurement decisions, the requirements for ethical treatment of human subjects, and so forth can make these designs difficult or impossible in many situations.[10] They know these problems are compounded in meta-analyses that depend on additional, standardized criteria that can be applied across studies and contexts.[11] Despite these limitations, educational experiments and systematic reviews of their effects can distinguish educational programs that work (assuming a shared definition of success) in many situations from those that work in only a few or not at all.[12]

A recent argument about the role of resources in instruction made by David Cohen, Stephen Raudenbush, and Deborah Ball illustrates how experimental research can refine knowledge of educational effects (outcomes).[13] In the following

8. Valentine and Cooper, "Scaling the Quality of Causal Research in Education," 7–8.

9. For more information about the Campbell Collaboration, see http://campbellcollaboration.org/; for more information about the What Works Clearinghouse, see http://www.w-w-c.org.

10. McCall and Green, "Beyond the Methodological Gold Standard of Behavioral Research."

11. Derek C. Briggs, "Meta-analysis: A Case Study," *Evaluation Review* 29, no. 2 (2005): 87–127.

12. Anthony Petrosino, Carolyn Turpin-Petrosino, and John Buehler, "'Scared Straight' and Other Juvenile Awareness Programs for Preventing Juvenile Delinquency," in *The Campbell Collaboration Reviews of Intervention and Policy Evaluations* (C2-RIPE) (Philadelphia: Campbell Collaboration, 2003); and Valentine and Cooper, "Scaling the Quality of Causal Research in Education."

13. David Cohen, Stephen Raudenbush, and Deborah Ball, "Resources, Instruction, and Research," *Educational Evaluation and Policy Analysis* 25, no. 2 (2003): 119–142. This work will be cited as *RIR* in the text for all subsequent references.

discussion, I draw heavily from their work in order to clarify the role of experiments and to provide a context for points I make in subsequent sections of this article.

Cohen et al.'s study focuses on the question of whether resources (money or the things that money buys) are systematically related to instructional outcomes. This is, of course, one form of a very important research question in education: Does money matter? Findings of previous studies were inconsistent.

The authors began by reviewing the findings and limitations of previous studies and, on that basis, developed a theoretical model of how the use of resources might intervene to affect the relation between resources and instructional outcomes. They described their model as follows:

> [We] distinguished among types of resources, and offered a view of causality. Conventional resources include teachers' formal qualifications, books, facilities, class size, and time. Personal resources include practitioners' will, skill, and knowledge. Environmental and social resources include state guidance for instruction, academic norms, professional leadership, and family support. Each type counts.…Yet…resources only count as they enter instruction, and that happens only as they are noticed and used…and…some sorts of resources are easier to use than others (*RIR*, 127–128).

Next Cohen et al. argued that, given the complex ways teachers and students might use various resources, or compensate for their absence, and given the interactivity of influences in classrooms, descriptive (or naturalistic) research alone was not likely to yield valid causal inferences about resource use. Descriptive studies of resource use were needed to generate hypotheses about the effects of resource use on instruction, but the hypotheses would have to be tested systematically — that is, experimentally — in order to identify a causal explanation for the effects of resources and resource use on instructional outcomes (*RIR*, 128).

With existing evidence (including data from a randomized field experiment[14]) about the effects of one resource, class-size reduction, Cohen et al. applied their model to develop a series of hypotheses about the influence of class-size reduction on instructional effects through the intervening variable, resource use. The researchers identified four hypotheses that helped them to distinguish among resource uses leading to different instructional effects (positive effects, negative effects, or no effects):

> When added conventional resources appear to directly affect learning, it is because they are usable, because teachers and students know how to use them, and because environments enabled or did not impede their use. If these ideas are correct, then when added resources lie outside the range of teachers' and students' knowledge, norms, and incentives, they will have no discernible effect. A hypothetical legislature might mandate that teachers use innovative content standards to engage learners in more creative and demanding work. The legislature might even provide money to write and disseminate the standards and support discussion of them. Yet research on the effects of such policy would probably show that the new resources had no average positive effects on students' learning, for the policy would have required most teachers and students to work well beyond their skills, knowledge, and will, without providing opportunities or incentives for them [to do so] (*RIR*, 132).

14. Jeremy D. Finn and Charles M. Achilles, "Tennessee's Class Size Study: Findings, Implications, Misconceptions," *Educational Evaluation and Policy Analysis* 21, no. 2 (1999): 97–109.

Cohen et al. explicitly note that this explanation is not news to classroom teachers or to those who regularly observe them; their point is that the interdependence of these various factors poses a serious challenge to descriptive research on instructional effects (*RIR*, 132):

> If teachers adjust the tasks they assign and the material they use, correct estimates of resource allocation and effects would depend on valid evidence of use. That would depend on teachers knowing and articulating what they did, and having the time and inclination to do so, or on researchers' valid observations of teachers' reports and practices, or both. Such evidence would not be easy to define or collect, especially since teachers often adjust their own knowledge, skill, and will as they apply them.…Nonexperimental studies of resource effects on student outcomes that fail to take account of how teachers adjust instruction in light of their judgments about students will likely misestimate and confound the resources used, those merely present, and their effects.…[If] teachers calibrate instruction…, one could make accurate causal inference about instructional effects only by reconceiving and redesigning instruction as a…system, and comparing it with different systems…in which the desired outcomes are specified and observed, and in which the intended outcomes are rationally related to consistent methods of producing those outcomes (*RIR*, 133).

On this basis, Cohen et al. conclude that the important research question to ask about the relationship between resources and instructional effects is not the conventional "Does money matter?" or even "Do resources effect instructional outcomes?" Rather, the key questions are "What instructional approach [cause], aimed at what instructional goals [effects], is sufficient to insure that students achieve those goals? [and] What resources are required to implement this instructional approach?" (*RIR*, 134–135):

> A first line of work should probe the effects [of the instructional system] for students on its central outcomes, when resources are plentiful. A second line could test the effects…under various resource constraints, which also could allow various modifications…under different conditions. Pursuing the two lines of research for [systems] that share outcomes, wholly or in part, would yield evidence about their robustness, generalizability, and cost effectiveness. As each was tested and modified, the research program would reveal the resources needed, as well as the ways in which they must be coordinated to produce effects. This active research agenda does more than passively discern the effects of extant resource configurations; it seeks valid causal inferences about specific instructional designs (*RIR*, 136).

Cohen et al. make especially clear why experiments are valuable for this type of research question. Yet they also make clear that their position does not eliminate the need for important contributions from other research designs, such as ethnographies and surveys. These nonexperimental designs are needed to reveal how teachers and students think and act within a setting, to identify and describe the range of resource uses and instructional systems, to clarify the internal dynamics of instructional systems, and to estimate the resources available. Ideally, according to the authors,

> Active [experimental] and passive [nonexperimental] research would be interdependent: without ethnographic research on instructional dynamics, it would be difficult or impossible to grasp the role or importance of various influences on instruction, and so to interpret the experimental evidence.…Active programs of research, in which deliberate interventions vary resources in relation to well-articulated [systems], are at the heart of our proposal, but so are well-designed programs of passive research (*RIR*, 137).

In summary, experimental research on relations among variables is extremely important for educational research. Few well-designed studies of this kind have been done recently, and more could be done. But such studies cannot be done well

without contributions from other kinds of research. To say that causal effects or causal explanations as defined by Cohen et al. can be determined by experiments alone is absurd.

SCIENTIFICALLY BASED RESEARCH ON CAUSATION

A second strategy for investigating causation derives from variance theory, includes experimentation, and further clarifies the importance of contributions from other research designs. This strategy is the one endorsed in the 2002 report of the National Research Council (NRC), *Scientific Research in Education* (*SRE*).[15]

The NRC Committee that produced this report was formed in 2000, in response to then-pending federal legislation that aimed to make educational research more scientifically based. This proposed legislation later became the Education Science Reform Act of 2002 (ESRA).[16] During the period of the Committee's work but independent of it, Congress passed NCLB into law. The language in the original version of ESRA and in the final version of NCLB limited "scientifically based research" to experimental designs and recommended randomized field trials.[17] *SRE* and, to a lesser extent, the final version of ESRA took a broader view of what counts as scientifically based research in education.

For the present discussion, the most important point made in *SRE* is that there are three kinds of research questions that scientifically based educational research tries to answer. These question types can be addressed separately (as in different research studies), but they are interdependent in the sense that answers to each type of question contribute to studies focused on the other types of questions and to the validity of overall conclusions about educational phenomena. Different research designs and methods are needed for each question type. Only one of these question types focuses on causal *effects*; thus, only one is potentially answerable with experimental designs.

The three kinds of research questions are (1) What is happening? (referred to in *SRE* as "descriptive questions"); (2) Is there a systematic effect? ("causal questions"); and (3) Why or how is it happening? ("explanatory questions," that is, questions about mechanisms or processes).[18] A number of commentators have criticized *SRE* for privileging the second class of questions — the so-called "causal questions" — and one means of addressing them — experimental methods,

15. National Research Council, *Scientific Research in Education*.

16. I was a member of the NRC committee that prepared *Scientific Research in Education.* Statements made here represent my view only; they do not represent the views of other Committee members or of the NRC generally.

17. Margaret Eisenhart and Lisa Towne, "Contestation and Change in National Policy on 'Scientifically Based' Education Research," *Educational Researcher* 32, no. 7 (2003): 31–38. See also *No Child Left Behind Act of 2001*; and *Education Science Reform Act of 2002*, HR 3801, Public Law 107–279, http://www.ed.gov/about/offices/list/ies/index.html (note that this is the final version of ESRA).

18. National Research Council, *Scientific Research in Education*, 99–123.

especially randomized field trials.[19] While I have to acknowledge that a small fraction of the report can be read in this way, my reading of the whole report indicates that answers to all three question types are crucial to scientifically based research in education.[20]

DESCRIPTIVE RESEARCH QUESTIONS

> The first question — What is happening? — invites description of various kinds, so as to properly characterize a population of students, understand the scope and severity of a problem, develop a theory or conjecture, or identify changes over time among different educational indicators.[21]

SRE discusses the value of this kind of information but tends to present it as only preliminary to the real work of scientific research. This representation is misleading. I agree with Fred Erickson and Kris Gutierrez, who have argued that good descriptive knowledge is essential if causal analysis is to succeed.[22] In order to understand whether or how *x* causes *y*, it is first necessary to know what, exactly, *x* and *y* are, and how in actual practice *x* can exert an influence on *y*. This is one of the impressive contributions of the Cohen et al. article discussed previously. Without good descriptions of how teachers and students use the resources available to them or compensate for the lack of resources, and without good descriptions of the internal dynamics by which use occurs or does not occur, experimental studies to identify causal relationships will be partial if not useless, and attempts to explain why or how causal agents work as they do are likely to fail. On this point, Erickson and Gutierrez note that "Unless considerable proportions of a research budget, even in a large-scale formal experiment, are devoted to documenting the treatment as delivered on the ground, the causal inferences drawn from the inspection will remain unwarranted."[23]

In order for descriptions to serve this fundamental role in research, they must be factually accurate, that is, they must have "descriptive validity." Joe Maxwell defines "factually accurate" descriptions as accounts about which both researchers and participants agree, specifically, when they agree that the account is physically, concretely, or behaviorally accurate.[24] Factually accurate descriptions can be obtained, evaluated, and justified only by means of qualitative data: field notes, audiotapes, or videotapes that can be used as illustrations and reviewed by others for consensus. In the case of descriptions of constructivist-oriented mathematics

19. Frederick Erickson and Kris Gutierrez, "Culture, Rigor, and Science in Educational Research," *Educational Researcher* 31, no. 8 (2002): 21–24; Kenneth Howe, "A Critique of Experimentalism," *Qualitative Inquiry* 10, no. 1 (2004): 42–61; and Maxwell, "Causal Explanation, Qualitative Research, and Scientific Inquiry in Education."

20. See National Research Council, *Scientific Research in Education*, 108–112, for the discussion that most explicitly seems to privilege causal questions and randomized trials.

21. Ibid., 99.

22. Erickson and Gutierrez, "Culture, Rigor, and Science in Educational Research."

23. Ibid., 21.

24. Joseph A. Maxwell, "Understanding and Validity in Qualitative Research," *The Qualitative Researcher's Companion*, eds. Michael Huberman and Matthew B. Miles (Thousand Oaks, California: Sage, 2002), 45.

instruction, for example, relevant questions about factual accuracy might include Do participants and researchers agree that the account portrays a teacher conducting a constructivist mathematics lesson? Do they agree that the students followed the directions given by the teacher? Do they agree that the students in small groups completed the work they were assigned? Do they agree that all but five of the students displayed stronger math skills after the lesson? Maxwell makes clear that descriptive accuracy or validity is not independent of theory or values. Rather, "descriptive validity is free from disagreement about the theory [and values] in question. This assertion does not mean that there cannot be disagreement about the descriptive validity of an account, only that such disagreement could in principle be resolved by the appropriate data."[25]

Factually accurate descriptive accounts of the social and cultural facts of educational events and interventions are the basis for answering all three kinds of research questions in *SRE*, and they are the basis for establishing the validity of these answers. Without precise, detailed, consensual accounts of *what* is going on — that is, without "descriptive validity" — researchers will not be able (1) to inform others about the characteristics of an event or intervention, (2) to identify important causal variables, (3) to generate reasonable hypotheses, (4) to determine the causes of success or failure, and (5) to identify, with confidence, the mechanisms that contribute to success and failure. In other words, without descriptive validity a study cannot make a contribution to scientific research on causation in education.

Several recent publications that make the case for more experiments in educational research have claimed that experiments are a powerful design that is used too infrequently; one implication is that there is currently too much descriptive work and too little experimentation.[26] The paucity of experimental research is demonstrated, in part, by the small number of experimental or quasi-experimental studies that are available for systematic reviews. However, Susan Flinspach, one of the few qualitative researchers to enter this debate, found that there were hardly any *descriptive* studies that include enough information to give a clear picture of what goes on in relations among school board members and superintendents.[27] I would argue that this lack of descriptive data is characteristic of many if not most other areas of educational research.

So, although one can make a case for the need for more good experimentation, more good descriptions are needed as well, and good descriptions of what is going on are crucial to the development of good experiments. Good descriptions are not merely helpful in suggesting hypotheses for future experimental testing, or in bringing contexts to life; they are fundamental to the validity of any causal claims that are made from research.

25. Ibid., 46.

26. See, for example, Robert Slavin, "Evidence-Based Educational Policies: Transforming Educational Practice and Research," *Educational Researcher* 31, no. 7 (2002): 15–21; and Turner et al., "Populating an International Web-Based Randomized Trials Register."

27. Susan Flinspach, *Interpretive Synthesis: A Methodology for Reviewing Qualitative Case-Study Research* (PhD diss., University of Chicago, 2001), 103–115.

CAUSAL AND EXPLANATORY RESEARCH QUESTIONS

According to *SRE*'s definition, "causal" research questions focus on establishing causal effects — that is, on answering the general question, Is there a systematic effect?[28] As in its discussion of experimental research, the report emphasizes the value of randomized experiments to make claims about causal effects. In *SRE*, "explanatory" research questions focus on Why or how does *x* cause *y*? Paul Holland has offered a helpful clarification of the commonalities and distinctions between *SRE*'s "causal," "explanatory," and "descriptive" types of research questions.[29]

Holland argues that all three question types are in fact causal.[30] The first type (labeled "descriptive" in *SRE*) asks directly about causes, as in What is the *x* (cause) that led to this *y*? or What is the antecedent that led to improved reading scores? Holland's second type (identified as "causal" in *SRE*) asks directly about effects, as in What *y* (effect) resulted from *x*? or What are the effects on reading scores of using, for example, the Open Court reading program? Holland's third type of question (called "explanatory" in *SRE*) asks about causal mechanisms, as in How does *x* work to cause *y*? or How does the Open Court program work to improve reading scores? Different information is needed to answer each causal question. To answer questions of the first type, antecedents must be identified. For the second type, relations among variables must be determined. For the third type, a mechanism by which antecedents produce the relations must be identified and justified.

Holland's typology helps to clarify the respective contributions of experimental and nonexperimental research designs to scientifically based studies of causation in education. The second causal question type — about effects — is relatively easy to answer because relationships among variables can be directly addressed in series of experiments or quasi-experiments (assuming confidence in the validity of the variables).[31] Using experimental research designs in studies of this question type gives estimates of causal effects. Estimates can be wrong (again assuming validity of the variables) but mainly by being biased, and biased estimates can be minimized through conducting subsequent experiments.

In contrast, the first causal question type can be quite hard to answer because there are often multiple causes of an effect. Confidence in identifying the "right" cause or causes requires more than repeated or refined experimentation, as the Cohen et al. argument about the effects of class-size reduction (discussed previously) illustrates. In the class-size example, the causal effects were already known

28. National Research Council, *Scientific Research in Education*, 99.

29. See Paul Holland, "Statistics and Causal Inference," *Journal of the American Statistical Association* 81, no. 396 (1986): 945–960; and Paul Holland, "Evidence for *Causal* Influence in Education Research" (paper presented at the annual conference of the American Educational Research Association, San Diego, California, April 2004). Holland, who also served on the NRC Committee that produced *SRE*, locates his view of causation in the tradition of John Stuart Mill.

30. Holland, "Evidence for *Causal* Influence in Education Research."

31. Ibid.

(improved achievement outcomes in small classes overall, but large within-class variations in achievement regardless of class size), but the cause itself was unknown. Careful reading of the previous literature, including experimental and nonexperimental studies, as well as their own experiences as teachers enabled Cohen et al. to develop a theoretical model that may account for achievement variations. Additional research will be necessary to determine the power of these causes individually and together and to eliminate other causal contenders. Some of this additional research will need to be experimental, but much of it will not. Research to identify additional kinds of resource use, and the nuanced conditions and contexts of the use of these resources, will require nonexperimental designs such as focused observations, case studies, and surveys.

In Holland's view, the third causal question type is hard to answer because causes can never be directly observed. Holland argues that it is impossible to observe simultaneously the action of *x* on *y* and the action of *not-x* on *y*, and thus it is impossible to observe directly the effect of *x* on *y*.[32] Thus, causal mechanisms must always be inferred from regularities in effects (outcomes), and specifying a causal mechanism itself depends upon interpretation and insight, not experimentation. As I will illustrate in subsequent sections, researchers with process views of causation argue that causal mechanisms can be more directly investigated through the careful use of qualitative research methods.

According to the authors and proponents of scientifically based research as well as the proponents of more experimental research in education, experimental research is a powerful tool when research questions ask about causal effects. But qualitative and other research designs and methods are needed to answer the harder causal questions about the antecedents that lead to effects and about the processes that make antecedents work. Scientifically based research claims about causation must be based on using a range of research designs and methods to seek sound answers to all three of *SRE*'s and Holland's question types.

Process Views of Causation: Realist and Intentional

Some scholars of education have challenged the view of causation that informs *SRE*, experimental designs, and Holland's work.[33] The important point made by these critics is that educational research predicated on a regularity or variance conception of causation can be blind to the role of human intentions, social interaction, and context in producing effects. Yet, human behavior and social interactions are often the causes of educational effects. These researchers do not reject the regularity view for certain purposes but emphasize that it leaves out much that should be included in research that focuses on human activity, such as educational research.

32. Holland, "Statistics and Causal Inference," 947.

33. See, for example, Howe, "A Critique of Experimentalism"; and Maxwell, "Causal Explanation, Qualitative Research, and Scientific Inquiry in Education."

An alternative to the regularity view is what Maxwell calls the process or realist approach to causality. On this view, the actual mechanism of causality — that is, the means by which some persons, events, or actions influence others — is the primary focus: "A mechanism…is not a variable but an account of the makeup, behaviour and interrelationship of those processes which are responsible for the regularity."[34] An example comes from my own study, *Educated in Romance*, a longitudinal descriptive study of college women's career aspirations and choices.[35] Based on ethnographic and survey data obtained over eight years, from the women's first year in college through four years past their college graduation, we discovered a "culture of romance" — a taken-for-granted system of beliefs about the way college male and female relationships work (a causal mechanism) — that could account for and explain the women's career-related choices. Maxwell is correct when he complains that neither experimentalists nor *SRE* elaborates on the value of this kind of research on causal mechanisms.[36]

Maxwell's view highlights the role that context and meaning play in causal mechanisms. Not only can context or meaning, or a combination of the two, generate a causal mechanism, but context and the people involved in the setting cannot be controlled or manipulated (in the experimental sense) without distorting the mechanism of interest.[37] Taking this position seriously requires that we recognize causal mechanisms as specific to context and intentions; thus, it is likely that they will have to be identified descriptively and cannot be isolated into components or manipulated to produce consistent outcomes without losing their causal power. If Maxwell's view of causal mechanisms in human activity is correct, then experimentation is certainly not an appropriate methodology for investigating it.

Ken Howe makes a similar distinction between forms of causation but frames it in a different way. Based on an interpretivist perspective, Howe distinguishes between "regularity" and "intentional" conceptions of causation:

> The regularity conception construes causation in terms of relationships among descriptive variables grounded in the outsider's perspective. The intentional conception construes causation in terms of relationships among intentional states and actions grounded in the insider's perspective.[38]

Howe's point is that causal explanations (mechanisms) for many human actions depend upon the beliefs, values, and norms of the persons engaged in those actions. Such mechanisms "really exist" for the people involved and compel their actions. While a regularity view of causation may identify causal explanations for actions that are outside human awareness, an intentional view is necessary to identify the mechanisms by which people decide to act in one way or another.

34. Maxwell, "Causal Explanation, Qualitative Research, and Scientific Inquiry in Education," 5.

35. Dorothy C. Holland and Margaret Eisenhart, *Educated in Romance: Women, Achievement and College Culture* (Chicago: University of Chicago Press, 1990).

36. Maxwell, "Causal Explanation, Qualitative Research, and Scientific Inquiry in Education," 6–7.

37. Ibid.

38. Howe, "A critique of experimentalism," 53.

Clearly, for exploring the kind of realist or intentional causation Maxwell and Howe describe, qualitative research designs and methods are preferred and superior. Designs such as ethnography, case study, narrative research, life history, and discourse analysis permit the direct examination of human behavior, beliefs, and intentions. Research methods of participant observation, interviewing, videography, document analysis, and reflection are the tools needed to do this work.

In my view, neither the authors of *SRE* nor experimentally-oriented researchers such as Cohen et al. or Holland are likely to disagree with using these designs or methods for these purposes. They might disagree, however, about what to do once such mechanisms have been identified. For instance, consider Howe's use of Shirley Brice Heath's *Ways With Words* to exemplify research based on an intentional conception of causation.[39] Heath's finding of a causal mechanism by which differences in linguistic practices led to distorted communication in school is enough for him. Having identified (with evidence and argument, of course) this causal mechanism, Heath's *research* is done. I do not think Cohen et al. or Holland (or the spirit of *SRE)* would want to stop the research there. I expect they would try to translate Heath's mechanism into a regularity model so as to investigate through experimentation various conditions and manifestations of the mechanism, its predictive power, and its limitations. Howe, Maxwell, and perhaps Heath would not want to do that. They would likely argue that such experimental manipulation would destroy the power of the causal mechanism. They are likely to move forward by introducing the causal mechanism and its evidence into public deliberations about school practices and goals. For them, research itself would stop at this point, while democratic deliberations based (in part) on it would proceed.

The disagreement between these two research camps is not about causation in itself but about what to do once causal effects and causal mechanisms have been identified and supported with evidence. One camp tends to argue for more fine-grained precision in specifying effects and mechanisms that sometimes comes from experimental research; the other tends to argue for placing the research evidence (such as it is) in the public domain for widespread discussion along with other non-research-based considerations, such as values and ethics. In my view, these two strategies are not mutually exclusive nor should they be. There is no reason that I can see why both strategies cannot be pursued simultaneously.

In summary, to the extent that human actions and volition cause educational outcomes — and surely they must — then realist or intentional causation cannot be ignored in research on education. Correspondingly, nonexperimental research designs and methods must be used and perfected to explore this type of causation. Once research evidence and arguments for causation in education are available, they should be pursued in various ways, including research-oriented efforts to improve understanding and public deliberations to widen the scope of their influence.

39. Ibid. See also Shirley Brice Heath, *Ways With Words: Language, Life, and Work in Communities and Classrooms* (Cambridge: Cambridge University Press, 1983).

A Practical Approach to Science and Causation

The final approach to causation that I take up here can be called "practical." In a recent book, Bent Flyvbjerg, taking his cue from Aristotle's three intellectual virtues, (re)makes the case for three types of science: epistemic science (familiar empirical inquiry that aims to predict the future by discovering universal or general laws, that is, experimental science in the sense that I have been using it here); technical science (empirical inquiry that aims to produce things that work); and phronesis or practical science (empirical inquiry that aims to understand how knowledge comes to be important or consequential in practice).[40] Flyvbjerg argues that the search for epistemic and technical knowledge is not a mistake in the social sciences (or, for my purposes, in education[41]). But such knowledge should not be privileged over practical science when human activity is at issue. The value of practical science lies in its potential to illuminate how rationality (knowledge) is constructed and acted upon in public deliberations and decision making (*MSS*, 142–143).

In some ways, Flyvbjerg's argument for practical science research is similar to the arguments Maxwell and Howe presented for a realist or intentional view of causation. All three call for more attention to the causal mechanisms generated by insiders as they attempt to make sense of their own actions in context. However, the kind of causation that interests Flyvbjerg does not come directly from individual or group intentions but from what actually happens in a given case and how it happens. Relying on the work of Pierre Bourdieu and Michel Foucault, Flyvbjerg finds answers to his questions through such means as historical case study methods and analyses of the power dynamics that lead to observed outcomes.

Flyvbjerg provides an illustration from his work on urban planning in Aalborg, Denmark (*MSS*, especially chap. 10). In this example, the research question is Why were certain changes taking place in the Aalborg urban landscape — changes that could not be rationally connected to the intentions of the key actors or the public? The context was a decision on the part of the Aalborg City Council, backed by the public, to improve the quality of the downtown urban core. To this end, the City Council approved a proposal prepared by urban planners to prohibit cars and to increase public transportation, bicycle paths, and walking corridors. However, as time passed and changes were actually implemented, cars were not prohibited, the number of cars increased, and the quality of the downtown core (in terms of air pollution, traffic, and health concerns) deteriorated even further. Flyvbjerg wanted to find out how and why this happened.

His research design was to trace historically the interests and actions of key persons and groups, their deliberations and decisions, the policies that were implemented, the compromises that ensued, representations of events in the media, and

40. Bent Flyvbjerg, *Making Social Science Matter: Why Social Inquiry Fails and How It Can Succeed Again* (Cambridge: Cambridge University Press, 2001). This work will be cited as *MSS* in the text for all subsequent references.

41. It should be noted that Flyvbjerg does not discuss education or educational research.

the outcomes that were produced. Using data from documents, interviews, surveys, and analyses of business trends and traffic patterns, he found that the power dynamics between two key groups — the City Council and the Chamber of Commerce — privileged the Chamber's position that the vitality of the downtown core depended upon the shoppers who drove there. Although survey data later showed otherwise, the Chamber held firmly to its position and used its power to influence decisions and to ensure that media accounts supported its view. Over time, this led to changes to the original plan and to outcomes (lower quality of life) that neither the City Council nor the public desired:

> The fate of the Aalborg Project would be decided by these two rationalities [the knowledge/position of the City Council and that of the Chamber of Commerce] fighting it out, and the group who could place the most power behind their interpretation of what was rational and what was not would win.…Distorted relations of power produced a distorted project. Power thus defined a reality in which the real Aalborg Project, that which has become a reality, deviates from and on principle objectives directly counteracts the formal [original] Aalborg Project, which was ratified by the City Council with a vote of 25–1, but which exists only on paper (*MSS*, 147–148, 154).

Having reached this conclusion, Flyvbjerg moves forward by publicizing his findings:

> I reasoned that if the arrangements and outcomes…in Aalborg were not publicly justifiable, as my studies showed they were not, then, perhaps, I could help change things for the better [that is, toward a more publicly justifiable process]…by calling public attention to my results (*MSS*, 156).

Over time, his strategy worked: public indignation and protest began to grow, and changes were eventually made to bring the project more in line with public views and research evidence.

The kind of research questions Flyvbjerg addressed and his approach to conducting research are germane to many research questions in education: How and why does a decision to require more standardized testing lead to desirable outcomes in some cases but undesirable ones in others? How and why does the adoption of a constructivist math program lead to desirable (or undesirable) outcomes? How and why does a school choice policy lead to desirable (or undesirable) outcomes?[42] Studies of this kind of "practical causation" can reveal how outcomes are produced by decisions made and actions taken in the minutiae of everyday practice, by who is involved and who is left out, and by whether the outcomes are desirable to those involved.

Flyvbjerg's methodological guidelines for producing high-quality case studies (or other accounts) that can contribute to an understanding of practical causation include (1) getting close to the people involved, (2) emphasizing concrete detail and context, (3) relying on practice (what people do) rather than on what they report, and (4) contributing to dialogue by narrating how things are, how they came to be that way, what their implications are likely to be, and what alternatives there

42. For an example of a study with implications for the school choice question, see Kenneth Howe, Margaret Eisenhart, and Damian Betebenner, "A Crucible of School Choice," *Phi Delta Kappan* 83, no. 2 (2001): 137–146.

might be (*MSS*, 132–139). With a different perspective and for different reasons, Flyvbjerg comes to a conclusion about social science research that is similar to the one reached by many qualitative researchers regarding educational research: accounts of regularities must include interpretations of intentions and analyses of power if they are ever to apply to practical circumstances.

In this, Flyvbjerg's perspective is similar to the position taken by many critical educational researchers. Critical researchers argue that educational research focused on either causal effects or intentions is blind to history (how values have developed over time and whose interests they serve, the effects of social and institutional contexts over time, and the impact of historical memory) and power (who has the power — and who does not — to define the questions asked, to establish the meaning of "doing science," to determine the designs and methods funded, and to decide what findings will be widely publicized). Critical researchers stress that there are many different ways to conceive of the issues to be studied within educational research, many different standpoints (based on race, gender, socioeconomic status, and so on) from which the results are viewed, and different material effects that the specific uses of the findings will have. I do not have space to explore these issues further here, but the work of such researchers as Patti Lather, Phil Carspecken, and Elizabeth St. Pierre provides useful insights.[43]

CONCLUSION

All four views of causation discussed here are relevant to educational phenomena. Hammers, saws, and more are needed in the toolbox of educational research if we are to study causation effectively. Having laid this out, it is shortsighted if not ridiculous to encourage attention to only one tool.

Certainly, educational researchers can be divided into camps based on their promotion of different designs and methods. This is not surprising since no single researcher can be trained equally well in all the relevant designs and methods or can hope to become expert in all of them. Debates among the camps are heated, sometimes fierce. These debates should not be discouraged. They are the hallmark of scholarship in a democratic society. What is not productive (for research or scholarship) is when debate stalls. What is not productive (for democracy) is for government to regulate the debate. Both of these things seem to be occurring. Debate is stalled if educational researchers continue to rehash old debates about qualitative versus quantitative methods, conventional versus critical perspectives, and so on. Debate is regulated if the federal government is allowed to mandate the kind of research designs that will be funded with public dollars.

One way to oppose these forces is to build on shared commitments. We all want to improve student learning, especially for students who are struggling in

43. See, for example, Patti Lather, "Research as Praxis," *Harvard Educational Review* 56, no. 3 (1986): 257–277; Patti Lather, "This *IS* Your Father's Paradigm: Government Intrusion and the Case of Qualitative Research in Education," *Qualitative Inquiry* 10, no. 1 (2004): 15–34; Phil Carspecken, *Critical Ethnography in Educational Research: A Theoretical and Practical Guide* (New York: Routledge, 1996); and Elizabeth St. Pierre, "'Science' Rejects Postmodernism," *Educational Researcher* 31, no. 8 (2002): 25–27.

school and society. We all want to support programs that succeed at this goal. Our differences are about how best to approach these issues, what specific questions to ask, what research priorities follow, and what outcomes are most desirable. But why not pool our diverse resources (perspectives, interests, and expertise) in all the ways suggested by the authors whose work I have reviewed here so as to learn as much as we can about important educational issues, from a variety of perspectives, all of which are relevant to educational phenomena? In saying this, I do not mean that we should all learn to dabble in various research designs, nor that we should all try to become expert in every design, but that we should work together in groups, collaboratively, and from all the angles and with all the tools we have available, to pursue important questions about education. On the basis of this collaborative work, we should then come together to exert political pressure on public officials to take our work and our findings seriously and to provide the money necessary to learn more. We should come together to influence the public to care more about what goes on in schools and more about the evidence for (or against) the programs and priorities of schools. These are worthwhile goals. Let's pick up the hammers and saws and go for them.

THANKS TO Hilda Borko, Susan Flinspach, Joe Harding, Ken Howe, and the *Educational Theory* reviewers for their comments on earlier drafts of this paper.