Advancing Educational Policy by Advancing Research on Instruction

Stephen W. Raudenbush

Am Educ Res J 2008 45: 206
DOI: 10.3102/0002831207312905

American Educational Research Journal
March 2008, Vol. 45, No. 1, pp. 206–230
DOI: 10.3102/0002831207312905
© 2008 AERA. http://aerj.aera.net

Advancing Educational Policy by Advancing Research on Instruction

Stephen W. Raudenbush
University of Chicago

Understanding the impact of "instructional regimes" on student learning is central to advancing educational policy. Research on instructional regimes has parallels with clinical trials in medicine yet poses unique challenges because of the social nature of instruction: A child's potential outcome under a given regime depends on peers and teachers, requiring the need for multilevel methods of causal inference. The author considers studies of the impact of intended versus experienced instructional regimes. Both are important; however, intended regimes are well measured and accessible to randomized trials, whereas experienced instruction is measured with error and not amenable to randomization. Multiyear sequences of experienced instruction are of central interest but pose special methodological challenges. A 2-year study of intensive mathematics instruction illustrates these ideas.

KEYWORDS: research on teaching, causal inference, classroom instruction, educational policy

Educational policy research has historically focused on the costs and consequences of providing resources such as facilities, teacher qualifications, class-size reduction, and instructional materials. These are things that policy makers can buy or regulate, and it is natural to ask about the likely payoff of making such investments. However, Cohen, Raudenbush, and Ball's (2003) review of 40 years of research on resources suggests that the yield for policy has been weak. To be sure, meaningful classroom learning without well-prepared teachers, usable facilities, comprehensible instructional materials, or

STEPHEN W. RAUDENBUSH is a professor of sociology and chair of the Committee on Education at the University of Chicago, Department of Sociology, 1126 East 59th Street, Chicago, Illinois 60637; e-mail: *sraudenb@uchicago.edu*. His areas of interest include improving statistical methods to study children's development within social settings such as classrooms, schools, and neighborhoods and developing better methods to measure the qualities of these social settings.

manageable class sizes is hard to imagine, yet an oft-repeated summary of the research is Hanushek's (1989, 1996) proposition that "money doesn't matter." Such a summary strikes many practitioners as nonsensical, generating a crisis in the credibility of educational research.

Hanushek's claim is, of course, open to dispute. Policy research has revealed evidence of some effects of resources on student learning (Hedges, Laine, & Greenwald, 1994) and particularly in regard to class-size reduction (Finn & Achilles, 1990; Krueger & Whitmore, 2001; Nye, Hedges, & Konstantopoulos, 2004). More recently, analysis of large-scale longitudinal data collected for school accountability has revealed some effects of teacher qualifications (Clotfelter, Ladd, & Vigdor, 2007). Yet most studies of teacher characteristics, teacher training, school facilities, and per pupil expenditures report remarkably small effects. Thus, despite its centrality in educational policy research, research on resources has shed few insights about how to improve student learning, and it is hard to argue that this research has had more than a marginal impact on the quality of classroom learning in the United States.

Educators use resources in such diverse ways for such varying purposes and with such varying effects that the potential benefits for student learning of spending money on schools are hard to discern. The implication is that we must look more deeply into how successful practitioners use resources in classroom instruction to understand how and when investments in resources come to have effects on student learning. The central focus of this article is therefore essentially methodological: How shall we study the impact of classroom teaching and the role that resources play in producing this impact?

First, I argue that improving educational policy research requires a new causal model. According to this model, the proximal cause of student progress is the "instructional regime." Resources play a prominent role in this model by facilitating the enactment of proven instructional regimes. Thus, knowledge about the impact of instruction supplies a scientific basis for policy concerning resources.

The study of classroom instruction therefore plays a role in educational policy that is similar to the study of clinical practice in health policy. Not surprisingly, the clinical trial immediately emerges as a model for testing instructional regimes. However, my second proposition is that despite important similarities, the *social structure of instruction*—the fact that it occurs within classrooms nested within schools—invalidates the canonical assumptions underlying the clinical trial in medicine. Implications for the design of research, experimental and nonexperimental, are fundamental.

Third, we will inevitably see differences between an *intended* instructional regime and the regime actually *enacted*. Understanding each of these is essential. A key methodological challenge is that inferences about the impact of instructional regimes students actually experience are not protected by randomization. Moreover, enacted regimes are inevitably measured with error, and errors of measurement threaten to bias estimated impacts of the enacted regime on student learning. It follows that studies of the measurement of classroom instruction is central to the proposed research agenda.

Fourth, discovering the impact of *multiyear sequences* of instruction is the fundamental aim of policy-relevant research on instruction. Most of the cognitive goals we seek for students, including, for example, reading with comprehension, writing an effective essay, or tackling a multistep math problem, require multiple years of instruction. To understand the impact of sequences of instructional regimes poses special challenges to valid causal inference, particularly in light of the fact that children's experiences in instruction occur in multiple social settings.

The Instructional Regime as the Proximal Cause of Classroom Learning

Past research on educational resources is implicitly based on a simple causal model in which investment in what Cohen et al. (2003) call a "conventional resource" (e.g., per pupil expenditures, teacher credentials, physical facilities, or class size) are the direct causes of a student outcome, typically an achievement test score (see Figure 1, Part a). The causal connection between such conventional resources and student outcomes has been found in many studies to be small (Clotfelter et al., 2007; Hanushek, 1989, 1996; Hedges et al., 1994). This sobering experience has evoked two kinds of conclusions. The first is that simply investing in conventional resources cannot generally be relied upon as a strategy for substantially improving the cognitive skill of the nation's youth. A corollary is that viewed as firms that produce cognitive skills, American schools are inefficient.

An alternative reaction to these findings is to criticize the simplicity of the causal model itself. Schooling serves multiple purposes. Increases in a school's budget might support new athletic facilities, new art or music programs, or new approaches to foster student health or social development. The impact of such uses of resources may be positive but are unlikely to be captured in measures of student cognitive skill. To understand the impact of resources, one must therefore consider how educators use those resources in pursuing specific aims.

Moreover, if schools are indeed inefficient, it makes sense to study resource use and how such use is linked to valued student outcomes. Only by identifying more and less effective ways of using resources will it be possible to increase school efficiency. This reasoning leads to an elaborated model (Figure 1, Part b). According to this model, any impact of investing in conventional resources is mediated by how those resources are used. If the model is well specified, there should be no direct connection between conventional resources and instruction. Rather, resources come to affect student outcomes only through specific uses that can be observed and statistically related to the outcome.

The more elaborate model that attempts to "explain" the link between resources and student outcomes has motivated substantial research on school climate and organization and classroom processes (Cohen et al., 2003). Ultimately, however, such a paradigm collapses under the weight of

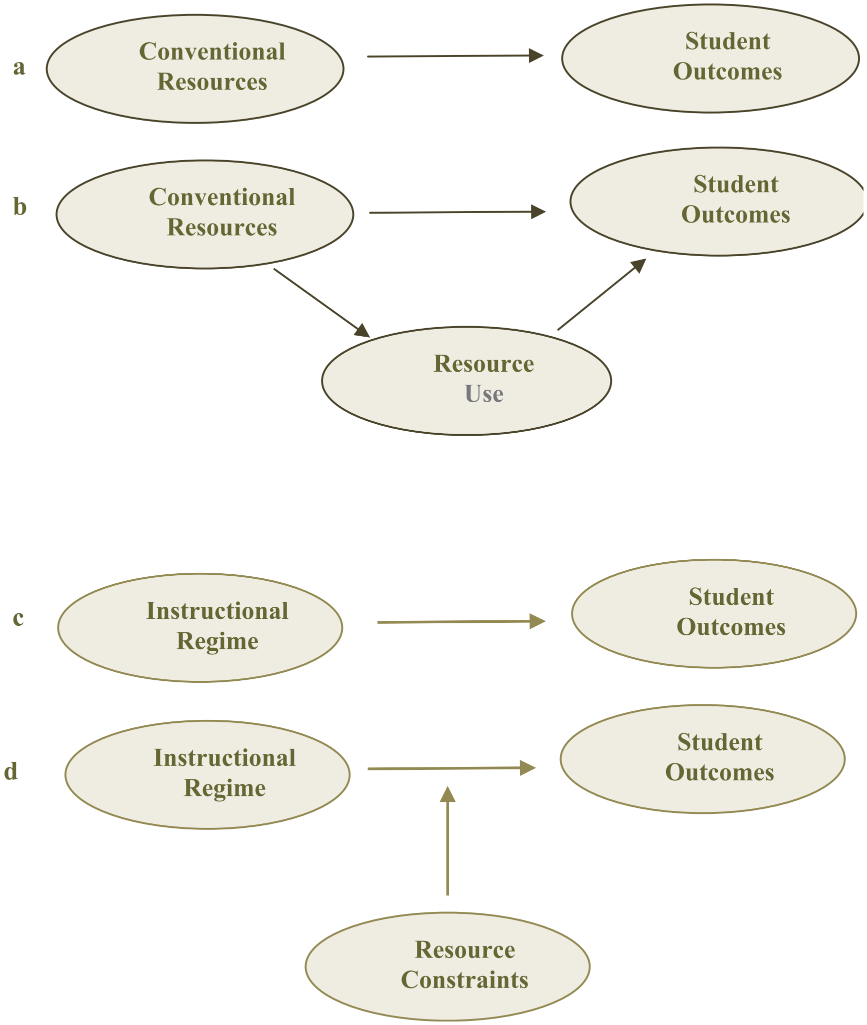


Figure 1. **Models for resources, instruction, and outcomes.**

its own aims. Resources are used for so many purposes, with such varied methods and such varied skill that the task of "explaining" how resources come to affect student outcomes is overwhelming. Moreover, studies of resource use have generally been fraught with selection bias. Practitioners who appear to use resources skillfully are likely to be skilled in many ways, and they may be assigned to comparatively advantaged students. Moreover, the impact of a particular aspect of resource use is difficult or impossible to pinpoint except in carefully controlled laboratory experiments. Such experiments could lead to promising new models of instruction. But studying those models takes us out of the “resource use” paradigm and encourages us to develop a new causal model.

Instruction as the Proximal Cause of Student Learning

A more straightforward approach is to formulate a model in which the clinical practice of education—classroom instruction—is the immediate cause of student learning (Figure 1, Part c). The idea is to test the impact of well-formulated systems of instruction that I shall label “instructional regimes” on student learning. In this setting, one can ask whether enhancing or constraining resources will modify the impact of the regime. Suppose that a well-conceived approach to promoting reading comprehension in second grade is found reliably effective across a range of school settings. Resource questions immediately follow: What kinds of preservice and in-service teacher preparation are required to enact the program? How small must class sizes be to make the program work? Are alternative regimes that require fewer resources equally effective?

An analogy with medical research is instructive. The effect of heart bypass surgery on angina patients would not have been discovered within a paradigm that sought to trace the connections between per patient expenditures on patient outcomes. Instead, clinical trials have produced evidence that for certain patients, heart bypass surgery produced more favorable effects than did conventional medical treatment. The question for resources is, then, how to increase the cost-effectiveness of the approach, for example, by providing such surgery in large centers that specialize in the procedure. In this case, heart bypass surgery is part of a treatment regime: One uses a clinical trial to identify its direct causal link to the outcomes of a well-defined set of patients; having found a positive effect, one considers resource constraints and cost-effectiveness.

Instructional Regimes

The notion of an instructional regime is founded on a picture of teaching as a continuous interplay of assessment and instruction. An effective teacher is ever attentive to the current level of student skill and knowledge. The interplay is dynamic: The appropriate next step in teaching a concept or skill is predicated on the observable success in having taught the prior step. Assessments can be formal and large scale (e.g., a districtwide test) or informal and small scale (e.g., evaluating a student response to a question during classroom discussion).

I therefore define an instructional regime as a more or less explicit set of rules for repeatedly assessing student skill and then tailoring instructional activities in response to the assessment in order to achieve more or less explicit aims for student learning. This process of assessment and tailored instruction produces, over time, considerable experience about a child’s progress and the success of attempts to improve that student’s learning. This cumulative knowledge, in principle, increases the effectiveness of subsequent efforts to engage the child in instruction.

This notion is closely linked to the "dynamic treatment regime" (Murphy, 2003), essentially a set of decision rules for allocating treatments to patients, given patient history and current status. In some cases, the classroom instructional regime is quite explicit. For example, in algorithm-guided individualized literacy instruction (Connor, Morrison, Fishman, Schatschneider, & Underwood, 2007), frequent student assessments of specific early literacy skills generate computer-produced recommendations for next instructional steps. In a similar vein, the Success for All program (Borman et al., 2005) prescribes periodic assessment of student literacy skills followed by a regrouping of students who are then targeted for quite specified modes of instruction. This process of assessing, regrouping, and instruction continues over time. The University of Chicago Charter Elementary Schools use a "balanced" literacy curriculum according to which every student is assessed three times per year (Kerbow, 2006). Each assessment generates a new plan of instruction for the next period. The record of assessments and instructional interventions provides evidence for staff discussions about how each student can reach explicit instructional aims.

In other cases, the interplay between assessment and instruction may not be formalized and may in fact be under the control of essentially autonomous classroom teachers. However, the assumption that effective teachers will engage in something like a regime remains intact even if the essential features of that regime are not readily open to public inspection. The decision to allow autonomous teachers to construct their own regimes might itself be regarded as a regime and evaluated as such. What appears to be a "nonregime" is in fact a hidden regime or a mix of hidden regimes.

The rationale for defining instruction at the level of "the regime" is in part a response to the unsuccessful attempts to define instruction at the more atomistic level of specific teacher actions or behaviors (Shulman, 1987). A thought experiment reveals the possibility that two teachers using apparently quite different instructional strategies might nonetheless be faithfully pursuing the same regime. For example, consider math instruction in two first-grade classrooms. Suppose that the first teacher is emphasizing word problems in single-digit addition whereas the second teacher is emphasizing practice in the single-digit addition computations. One might compare the outcomes of those teachers' students to compare the effectiveness of these approaches. Suppose, however, that the common regime prescribed (a) motivating the concept of addition by posing children with a simple word problem requiring addition, (b) practice in computation of a series of addition problems, and (c) posing a new series of word problems in addition. The regime involves assessing success in (a) before proceeding to (b) and assessing success in (b) before proceeding to (c). Our first teacher, who is teaching a word problem in addition, could have students at Stage (a) or (c) whereas the second teacher might have students at Stage (b). Thus, *within the regime*, the current status of the student and not any differences between the teachers is driving the instructional action.

This example illustrates a basic principle in the study of dynamic treatment regimes. If each of many teachers is faithfully enacting a regime, it is not possible to identify the causal effect of teacher behavior within that regime. Within such a regime, every instructional action is prescribed by the current level of student functioning. Therefore, given current student status, there is no variation in instructional behavior to study! However, it is entirely possible to compare different regimes. For example, one can readily imagine randomly assigning classrooms or even whole schools to pursue one of two different regimes as a way of evaluating their relative effectiveness (Spybrook, 2007).

Why Are Field Trials of Instruction Different From Clinical Trials in Medicine?

The argument here is that evaluating alternative instructional regimes ought to be at the core of the enterprise of educational policy research. How, then, shall we carry out such evaluations?

The clinical trial in medicine immediately emerges as a model for testing instructional regimes. Indeed, a number of educational randomized trials are now in the field (Spybrook, 2007), reflecting a dramatic shift in research policy at the U.S. Department of Education's Institute for Educational Sciences. That shift in policy itself reflects a kind of sea change in thinking about educational research (Mosteller & Boruch, 2002) favoring a much stronger emphasis on the use of randomized experiments in education.

However, despite important similarities, my second proposition is that the social structure of instruction—the fact that it occurs within classrooms nested within schools—negates the validity of canonical assumptions underlying the clinical trial in medicine. This distinction has fundamental implications for the design of research, experimental and nonexperimental, in education.

To clarify the differences between clinical trials in medicine and education, let us now consider the key statistical assumptions underlying a paradigm case: the randomized drug trial. It will become apparent that some of the key assumptions invoked in such a trial will generally not apply in education and that new assumptions are needed as the logical basis of experiments in education.

To motivate the argument, consider a drug trial in which each patient is assigned to receive to receive a "new drug" ($Z = 1$) or the "standard treatment" ($Z = 0$) with the aim of discerning the impact of the new drug relative to the standard treatment on some outcome of interest (e.g., blood pressure), denoted Y.

According to current thinking about causation in statistical science (Holland, 1986), we can then define for each patient two potential outcomes of treatment. If the patient is assigned to the new drug, we will observe the outcome $Y(1)$, literally, the value Y will take on if $Z = 1$. In contrast, if the patient is assigned to the standard treatment, we will instead observe the outcome $Y(0)$. This setup then defines for each patient the causal effect

$$\Delta = Y(1) - Y(0). \tag{1}$$

The idea of causal effects as person-specific differences between potential outcomes is at the heart of what statisticians now call the “Rubin causal model” because of Donald Rubin’s (1978) highly influential work and that of his colleagues (Holland, 1986; Rosenbaum & Rubin, 1983). However, the roots of this approach may be found in earlier statistical work (Neyman, 1935) and in economics (Haavelmo, 1943; Heckman, 1979).

The crucial point is that Δ is a patient-specific causal effect that, by definition, can never be observed: If the patient is assigned to $Z = 1$, we will observe $Y(1)$ but not $Y(0)$. For that patient, $Y(0)$ is known as the “counterfactual outcome.” In contrast, if the patient is assigned to $Z = 0$, we will observe $Y(0)$ but not the counterfactual outcome $Y(1)$. Holland (1986), who provides an exceptionally lucid presentation of this approach, has termed “the fundamental problem of causal inference” the fact that the key quantity of interest (the causal effect Δ) cannot be observed. However, under certain key assumptions, we can estimate the “population-average” causal effect, which I will label as

$$\delta = E(\Delta) = E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)]. \qquad (2)$$

The symbol E in Equation 2 denotes an “expected value,” that is, the population-average value of the term in brackets. The idea is that although we cannot observe the patient-specific causal effect Δ, every patient, in principle, possesses such an effect, and δ is the average of those causal effects in the population represented by our sample. Equation 2 says that the average causal effect is just the difference between the population average value of $Y(1)$ (the population mean if everyone in the population received the new drug) and $Y(0)$ (the population mean if everyone in the population received the standard treatment).

To estimate the population-average causal effect (Equation 2), one might compute

$$\hat{\delta} = \overline{Y}_1 - \overline{Y}_0. \qquad (3)$$

Equation 3 says that the sample estimate $\hat{\delta}$ of the true population-average causal effect δ is the simple difference between the sample mean $\overline{Y}_1$ of those actually assigned to the new drug and the sample mean $\overline{Y}_0$ of those actually assigned to the standard treatment. However, two key assumptions are required before we can infer that our sample estimate $\hat{\delta}$ will be an unbiased estimate of the true population-average causal effect δ.

First, the sample mean of those assigned to the treatment, that is, $\overline{Y}_1$, must be an unbiased estimate of the population mean $E[Y(1)]$. Recall that this population mean is the mean value of the potential outcome of $Y(1)$ for the entire population—including those assigned to receive the control treatment

$Z = 0$! In essence, the question is whether those actually assigned to receive the new treatment are representative of the entire population. This reasoning generates the rationale for a randomized experiment. Under randomization, those actually assigned to receive the treatment $Z = 1$ will represent the entire population. In a similar vein, under randomization, $\overline{Y}_0$ will be an unbiased estimate of the population mean $E[Y(0)]$: Those actually assigned to receive the standard treatment will represent the entire population. In essence, under randomization, one's potential responses to the two treatments have no bearing on whether one will be assigned to receive the new treatment or the standard treatment.

The Stable Unit Treatment Value Assumption (SUTVA)

An additional assumption is, however, required in order to motivate this entire discussion. We have thus far assumed that each patient possesses two and only two potential outcomes, one for each possible treatment assignment. Rubin (1986) called this the "stable unit treatment value assumption" (SUTVA). The basic idea is that one's potential outcome under a given treatment remains stable regardless of (a) the treatment assignment of other persons and (b) the mechanism used to assign the treatment.

a. In the context of our hypothetical drug trial, SUTVA implies that a patient's response $Y(1)$ to the new treatment $Z = 1$ does not depend on the treatment assignment of other patients. This is known as the assumption of "no interference between units." One can imagine instances in which such an assumption might not hold. For example, suppose that two married partners participated in the trial and that one of the drugs has negative side effects on one's ability to interact socially. Then we can suspect that a patient's outcome would depend on the treatment assignment of one's spouse. However, in most practical cases, the assumption of no interference between units seems reasonable in the context of a typical drug trial.
b. One entailment of (b) is that a patient's potential outcome will not depend on the physician who prescribed the drug. Again, one can imagine instances in which this assumption would not hold. For example, it may be that certain physicians are particularly skilled in convincing their patients to take their medicine. However, the assumption that the potential outcome of being assigned to a certain drug does not depend on the physician may be quite reasonable in many drug trials.

Without SUTVA, for an arbitrary patient whom we shall designate as Patient 1, we can denote the potential outcome of that patient as $Y_1(Z_1, Z_2, \ldots, Z_n; d)$. In words, the potential outcome for Patient 1 depends on that patient's own treatment assignment, Z_1, but also on the treatment assignment Z_2 of Patient 2 and that of all other patient assignments up to the nth patient's treatment assignment Z_n. Moreover, Patient 1's potential outcome also depends on the identity of that patient's doctor, d (for $d = 1, \ldots, D$), where D is the total number of possible doctors.

We can clearly see that without SUTVA, causal inference becomes extremely messy. A shift in one patient's treatment assignment would perturb the potential outcome of all other patients. How then can we even define, much less study, a causal question? Moreover, a patient's potential outcome, and therefore any causal effect of interest, will depend on the particular physician treating that patient, further complicating the framing of causal questions, their estimation, and their generalizability.

SUTVA thus makes causal inference tractable and can be defined mathematically in our context as $Y_1(Z_1, Z_2, \ldots, Z_n; d) = Y_1(Z_1)$, implying that the potential outcome of Patient 1 depends only upon that patient's own treatment assignment and is therefore independent of the treatment assignment of all other patients and also of the physician who prescribes the drug.

SUTVA in Research on Instruction

Suppose now that we uncritically import the paradigm of the randomized drug trial into research on instruction. That is, we randomly assign students to receive a novel instructional regime, say, $Z = 1$, or to the standard approach to instruction, that is, the control condition $Z = 0$. Although some might regard such a study as meeting a gold standard for causal inference, we would immediately confront the implausibility of SUTVA, which might (for the case of an arbitrary student labeled Student 1) be written as $Y_1(Z_1, Z_2, \ldots, Z_n; t) = Y_1(Z_1)$, implying that Student 1's response to receiving treatment Z_1 will not depend on the treatment assignment of any other students or on t, the teacher to whom that student is assigned. To most educators, such an assumption would seem on its face implausible. First, a student's response to any approach to instruction may well depend on the composition of the class, that is, the background, experiences, and motivation of that student's classmates. Second, the teacher to whom that student is assigned is likely to be influential.

We can conclude that SUTVA as invoked in the clinical trial in medicine is not generally plausible in the context of research on instruction. This conclusion has profound implications for how we define causal effects and for how we design research, both experimental and nonexperimental. Hong (2004) and Hong and Raudenbush (2006) consider these issues in detail, and the following discussion draws heavily on their work.

Implications for Experimental Design

The first implication of the failure of SUTVA as invoked in medical trials is that we will often prefer to assign entire classes or even whole schools to alternative instructional regimes rather than assign individual students. We can then invoke SUTVA, but two other assumptions are required.

Let us consider the case in which classrooms within schools are assigned at random to one of two instructional regimes. Let j denote the identification number of a classroom. The potential outcome of the first student in classroom j may be written as $Y_{1j}(Z_{1j}, Z_{2j}, \ldots, Z_{n_j j}; t_j)$. In words, Student 1 in classroom j will display this outcome if that student is assigned to treatment

Z_{1j}, the second student in that class is assigned to treatment Z_{2j}, and so on, right up to the last student, that is, student n_j in classroom j, who is assigned to treatment $Z_{n_j j}$. Here n_j is the number of students in class j. Moreover, the outcome will depend on which teacher t_j is assigned to teach that classroom. Defining the potential outcome this way implies two key assumptions identified by Hong and Raudenbush (2006).

First, we are assuming "no interference between classrooms." Note that the potential outcome $Y_{1j}(Z_{1j}, Z_{2j}, \ldots, Z_{n_j j}; t_j)$ does not depend on students or teachers in classrooms other than j. This means, for example, that teachers assigned to different instructional regimes do not share information with each other in a way that changes their behavior so as to modify the outcomes of their students. If interference ("spillover") between teachers and students in different classrooms is likely, it will tend to make more sense to assign entire schools (or even entire districts) to instructional regimes.

Second, we are assuming "intact classrooms." That means that students and the teacher will not move from one classroom to another after that classroom has been assigned to an instructional regime. This assumption becomes problematic in studies of long duration. In particular, if the aim is to evaluate the impact of instruction occurring over more than 1 year, the intact-classrooms assumption will become implausible in most settings, an issue we discuss in detail later on.

Under these two additional assumptions—no interference between classrooms and intact classrooms—we can now define the potential outcome of our Student 1 in class j if that class is assigned to the new instructional regime, that is, if $Z_j = 1$. We then have the potential outcome $Y_{1j}(1, 1, \ldots, 1; t_j)$. Note that because the entire class is assigned to the common treatment $Z_j = 1$, every member of that class is assigned to that treatment. In the same vein, if classroom j is assigned to the traditional instructional regime, Student 1 will display the outcome $Y_{1j}(0, 0, \ldots, 0; t_j)$. We therefore can define the child-specific causal effect

$$\Delta_{1j} = Y_{1j}(1, 1, \ldots, 1; t_j) - Y_{1j}(0, 0, \ldots, 0; t_j). \qquad (4)$$

Under our additional assumptions of no interference between classrooms and intact classrooms, Equation 4 shows that SUTVA now applies: Every student has one and only one potential outcome under each possible treatment. We can therefore define a person-specific causal effect (Δ_{1j}) as the difference between these two potential outcomes. Causal inference becomes tractable, and the random assignment of classrooms to instructional regimes will supply a basis for unbiased inference about the causal effect of one instructional regime relative to another. Moreover, this formulation allows the causal effect Δ_{1j} for classroom j to vary randomly over schools or districts.

Cost and Statistical Power

The forgoing discussion suggests that group-randomized experiments (where groups are classrooms are schools) will typically be the gold standard for assessing the impact of instructional regimes on cognitive outcomes. This finding is consistent with current practice in that the vast majority of experiments supported by the U.S. Department of Education have in fact used randomization by group (Spybrook, 2007).

Questions that immediately arises is how many classrooms or schools are required to achieve adequate statistical power in such studies and how much it will cost to run each study. The answer is disconcerting—unless great care is taken in the planning of these studies (Raudenbush, Martinez, & Spybrook, 2007).

Consider a fairly typical scenario in which schools will be assigned to a new instructional regime or a standard regime. We want to know how many schools must be studied in order to achieve an adequate statistical power of .80, meaning that our study will have a probability of .80 of detecting the true positive effect of the intervention. Assume that the expected effect size is 0.30 in standard deviation units, somewhat large but within a realistic range, and that about 20% of the variation in the outcome lies between schools within treatments with 10% between classrooms within schools, a fairly typical finding (Bloom, 2007; Hedges, 2007). Suppose further that we shall sample three classrooms per school with 20 children in each classroom. Under these assumptions, we would need 86 schools—43 in each treatment condition, a daunting result!

Suppose, however, that a pretreatment covariate, the prior mean achievement in a school, has a correlation with increases in the outcome of .80, a plausible correlation when the outcome is also a measure of achievement (Bloom, 2007). By using this covariate in the analysis, the required number of schools is sharply reduced to 42, that is, 21 per treatment—still a fairly large study but not implausibly so (see Borman et al., 2005).

An alternative way to use prior information is to match pairs of schools on one or more such pretreatment covariates. This approach will produce power that typically approximates that in the case of analysis of covariance (Raudenbush et al., 2007). An advantage is that matching can insure that the two groups are closely balanced on salient characteristics such as ethnic or social class composition.

A key conclusion from this line of work is that randomized studies of alternative instructional regimes are likely to be fairly large. Even then, careful exploitation of prior information is needed to ensure that those studies will have adequate power. A fleet of such studies, as recommended here, will require a fairly substantial investment. Given that many nonexperimental studies will also be needed to support this research agenda (see Raudenbush, 2005, and also the next section of this article), we can anticipate that a focus on comparing instructional regimes as a strategy for improving educational policy will require sustained financial support at a reasonably high level.

Intended and Enacted Regimes and the Contributions of Nonexperimental Research

The foregoing section suggests the need for a fleet of randomized studies evaluating the impacts and costs of alternative instructional regimes.

However, a successful agenda for research on the impact of instructional regimes actually requires a well-orchestrated interplay between experimental and nonexperimental research.

We will inevitably see differences, often substantial, between an *intended* regime and the regime actually *enacted*. If so, two new research questions immediately emerge. First, how does the planned intervention (that is, the intended regime) affect the regime children actually experience (the enacted regime)? Second, how does the enacted regime affect student learning? Randomization helps us with the first question but not the second. Let us consider why each of these questions is important before taking up the methodological challenges entailed in answering them.

Why Study the Impact of the Intended Regime on the Enacted Regime?

The claim here is that knowing how an intervention affects instruction is essential to interpreting impacts on student learning. This is most clearly the case when experiments show no effects but is also true in the presence of a positive effect.

Interpreting null results. Consider a study in which the assignment of schools or classrooms to an intended regime is found to have no significant impact on student learning. Assume that the study design was unbiased and provided adequate statistical power to detect a non-negligible effect. Two explanations immediately arise.

First, it may be that adoption of the intended regime changed classroom instruction in the ways intended but that those classroom changes made no difference in student learning. Program evaluators use the term *theory failure* to describe this scenario, because the theory that links intended changes in instruction to intended student outcomes will have proven incorrect.

Second, the instructional regime may never have been effectively implemented in classrooms. Perhaps the innovators lacked skill in working with teachers, or perhaps the teachers lacked the skill, knowledge, or motivation to put the innovative ideas to work in their teaching. In any case, program theory about the relationship between instruction and student outcomes was never tested, leading to what evaluators often call *implementation failure*.

Without valid assessments of the enacted regime, it would be impossible to distinguish between these two explanations, severely limiting the study's contribution to knowledge. One would never know whether the theory underlying the program had in fact been tested—a big problem given the likely cost of such an experiment.

Interpreting non-null results. Suppose instead that assignment to the intended regime did produce gains in student learning. One might then assume that the enacted regime "worked" by improving instruction according to its theory. But without valid measurements of the instruction children actually experience, that is, without measurement of the enacted regime, this conclusion would be unwarranted. Perhaps the innovation worked in other ways, an assertion that could not be probed without studying the impact of the innovation on instruction. Once again, a failure to measure key aspects of classroom life yields major ambiguities in the findings, leaving open key questions about how to use the findings to improve teaching and learning.

Why Study the Impact of the Enacted Regime on Student Learning?

Suppose now that a new regime, if enacted faithfully, effectively boosts student learning. If all teachers faithfully followed an experimental instructional regime, the intended and enacted regimes would be identical and, of course, their impacts on student learning would be identical as well. Suppose, however, that some teachers assigned to a new regime fail to enact it. Then the experimental estimate of the impact of the intended regime would underestimate the impact of the instruction children actually experience, that is, the enacted regime. The effect of the intended regime would be attenuated further if some of the control teachers had, contrary to expectation, adopted and used the new experimental methods in their classroom.

Such a scenario is not hypothetical. The Success for All (SFA) regime specifies an approach to reading comprehension that Rowan, Camburn, and Correnti (2004) refer to as "skill based." These authors also studied an alternative regime, America's Choice (AC), which uses a "literature-based" approach to reading comprehension using considerable writing about what children read. A third set of teachers in comparison schools had adopted neither SFA nor AC. Rowan et al. collected detailed information on classroom practice for all three kinds of schools using teacher logs of daily activities. The results showed that about 80% of the SFA teachers adhered quite faithfully to the skill-based approach, whereas another 20% did not. Similarly, about 70% of the AC teachers relied heavily on the literature-based approach. Interestingly, quite a large minority of comparison teachers, who ostensibly had no relation to either SFA or AC, appeared to adhere quite faithfully to an approach that closely resembled either the skill-based or the literature-based approaches. It is clear in this setting that the effects of the intended regimes would underestimate the effect of the enacted regimes.

One might argue that only the impact of the intended regime is important to the policy maker: Policy makers and administrators can encourage adoption of an intended regime, but they have no direct control over teacher behavior and therefore cannot ensure that the intended regime will be enacted. If the intended regime is taken to scale, one would expect slippage in implementation, and the resulting attenuation should be reflected in any calculation of the expected benefit of adopting the intended regime.

Yet understanding the impact of the enacted regime is also important for policy. Once the enacted regime's full benefit becomes visible, incentives shift: Teachers may become more inclined to give a new approach a try if evidence shows it to be especially effective, and policy may directly or indirectly reward them for doing so.

Methodological Challenges

Once we accept the rationale for studying the causes and consequences of the enacted regime, important methodological entailments follow. First, we have to develop technologies for measuring instruction as it is enacted in classroom settings. Inevitably, we will measure classroom instruction with error. Quantifying the degree of measurement error then becomes essential in research design and statistical analysis. Second, studies of the impact of enacted regimes on student learning will not benefit from random assignment. Sound methods of causal inference in nonrandomized studies therefore become essential. Let us consider how these methodological issues play out as we pursue the two new questions of interest.

Studying the Effect of the Intended Regime on the Enacted Regime

We first consider the case in which the aim is to study the impact of assigning a classroom to a new regime on the enacted instructional approach. A key problem is that the enacted instruction, whether measured through direct classroom observation, teacher logs, interviews, or questionnaires, will generally be measured with some degree of error. These errors of measurement do not cause bias, but they do reduce precision. To obtain adequate power requires sampling a larger number of classrooms than would be required if the enacted instruction were measured with perfect reliability.

Raudenbush, Martinez, Bloom, Zhu, and Lin (2007) have shown that the impact of reliability on statistical power can be substantial. In the case where classroom observations assess the enacted instruction, reliability can be increased by better training of observers, use of more observers per classrooms, or observing each classroom on more occasions. A trade-off arises between investing resources in increasing the reliability of measurement and investing resources in recruiting and sustaining the involvement of more classrooms and schools. Raudenbush et al. (2007) consider these trade-offs. This work demonstrates the importance of careful studies to discern the reliability and validity of measures of classroom instruction within the framework of a research agenda focusing on instructional regimes.

Studying the Effect of the Enacted Regime on Student Learning

Next we consider the case in which the aim is to assess the impact of the enacted regime on student learning. Once again, the enacted regime is measured with error. The key result in this case is that measurement error creates bias. Specifically, if the effect of the enacted regime on student outcomes is positive, the estimate based on a fallible measure will be negatively biased. One of the benefits of conducting a study of the reliability of the classroom measurement is that information from such a study can be used to correct the bias that arises from measurement error as explained in detail by Raudenbush (2007).

The Challenge of Studying Multiyear Instructional Sequences

Discovering the impact of multiyear sequences of instruction might well be the fundamental aim of policy-relevant research on instruction. Whether children can read or reason mathematically is the cumulative result of sequences of instructional experiences over several years. To understand the impact of sequences of instructional regimes, however, poses a special challenge to valid causal inference, particularly in light of the fact that children's experiences in instruction occur in multiple social settings.

Background

Causal comparative studies of instruction tend to be bounded by a single academic year. While understanding instructional effects during a single year is necessary, it is essential to understand how sequences of instruction extending over 2 or more years cumulatively affect learning. The effect of a multiyear sequence of instructional experiences cannot logically be equated to the sum of the effects of instruction occurring each year. Early reading instruction provides a useful example.

We now have considerable confidence in our methods to teach a broad range of children how to decode familiar text during Grade 1. Suppose we had equal confidence that certain strategies for building comprehension during Grade 2 are also effective. To maximize comprehension by the end of Grade 2, however, does not necessarily imply simply maximizing the additive effects of these Grade 1 and Grade 2 approaches. The problem is that many children come to school with comparatively weak vocabulary, general knowledge, and fluency in standard English (National Institute of Child Health and Human Development, 2000). These limitations plausibly constrain learning in Grades 3 to 5 when comprehension ("reading to learn") replaces the goal decoding of familiar text ("learning to read"). It may well be that during Grade 1, an immersion in language, ideas, oral comprehension, and text are necessary along with a strong emphasis on explicit instruction in decoding in order to maximize the child's gains in comprehension observable in later grades. Such an effect cannot be discerned from even the best studies of teaching and learning during a single academic year.

The Problem of Time-Varying Confounding

If we could randomly assign students to sequences of instructional experiences across years and if students' treatment assignment remained intact, we could obtain unbiased inferences about the impact of a sequence of instructional experiences using standard and quite simple methods of analysis. The essential problem is that students will tend to migrate across classrooms and even across schools from one year to the next. In many urban areas, such school mobility is remarkably high. And in a study where instructional sequences differ within schools, parents and teachers may find reasons to opt into treatment conditions not assigned. In these cases, the randomization will tend to break down. If so, this scenario generates the problem of *time-varying confounding*.

Experience From Epidemiology

Once again, we can learn important lessons from public health research about causal inference in this case (Robins, 2000). A classic example involves exposure to toxic chemicals in a workplace. An obvious hypothesis is that the more exposure a worker has to toxic chemicals, the worse that worker's health will be over time. Nevertheless, we can easily imagine a scenario in which the association between years of exposure to toxic chemicals and health is actually *positive*. Suppose that some workers respond immediately and negatively to such exposure and therefore quit their jobs in the factory that exposes them to toxic chemicals. However, suppose that other workers are actually quite resistant to the effects of such chemicals. These hardy workers stay on the job with no negative effects. Then an analysis of the association between years of exposure to toxic chemicals and health would find a positive association even though the causal effect is, on average, negative.

This idea is formalized in Figure 2. Define Y_0 as the baseline health of a worker. Define $Z_1 = 1$ if that worker is exposed to toxic chemicals in Year 1 and $Z_1 = 0$ if not. The outcome at the end of Year 1 is Y_1. Next, $Z_2 = 1$ if that worker is exposed to toxic chemicals in Year 2; $Z_2 = 0$ if not. The outcome at the end of Year 2 is Y_2. We are interested in the effects of Z_1 on Y_1 and the joint effects of the entire sequence—that is, the Z_1, Z_2 sequence—on Y_2. We note, however, that Y_1 is a "time-varying confounder," that is, both an outcome of Z_1 and a predictor of Z_2. Suppose that the impact of Z_1 on Y_1 is negative: Exposure to toxic chemicals hurts health in the short term. Suppose also that Y_1 negatively predicts Z_2: Bad health in the short term predicts avoiding toxic chemicals next year. Suppose that Z_2 has little effect on Y_2 for those hardy workers, having suffered no ill consequences of Z_1, who have stayed on the job. Then the correlation between $Z_1 + Z_2$ and Y_2 will be positive: The greater the exposure to toxic chemicals, the better the health!

Now the question arises: How do we solve this problem? One option is to control for Y_1 when we examine the impact of Z_2 on Y_2. That makes perfect sense—if our only aim is to study what happens in Year 2! But if our original aim is to study the impact of a sequence of treatments, in this case, the cumulative effects of toxic exposure over 2 years, we will not answer our question by controlling Y_1 in evaluating the impact of Z_2 on Y_2. The problem is that Y_1 is "in the causal pathway" between Z_1 and Y_2. Controlling for Y_1 will "control away" the impact of Z_1 on Y_2.

So we are between a rock and a hard place: If we fail to control Y_1, we bias our estimate of the effect of Z_2 on Y_2; if we do control Y_1, we bias our estimate of the effect of Z_1 on Y_2. In neither case will we achieve our goal of understanding the impact of the sequence of treatments.

Robins (2000) has developed an ingenious solution to this problem. In sample survey research, groups of respondents who are underrepresented

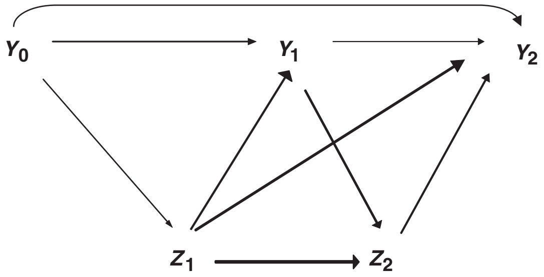


Figure 2. **A simple model for a sequence of two treatments.**

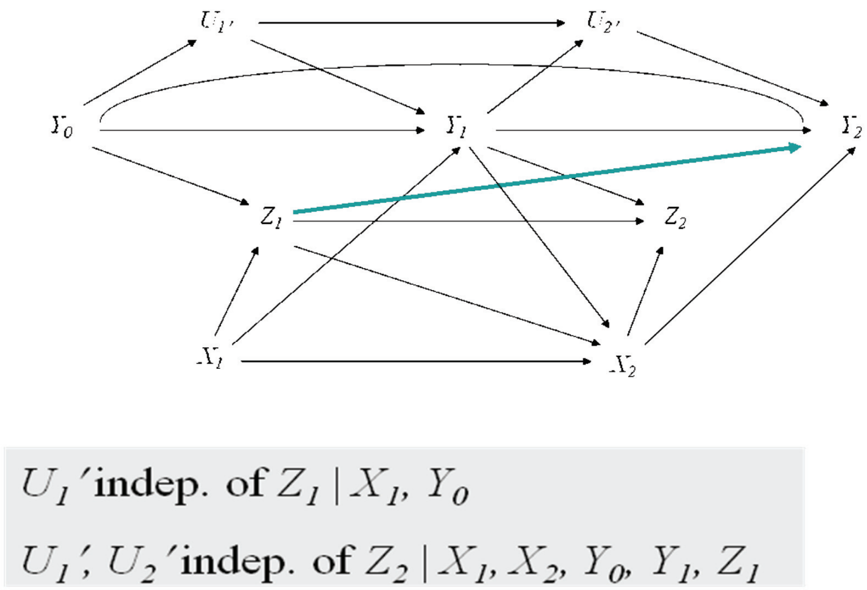
in the sample are "weighted up," while persons who are overrepresented are "weighted down." Robins showed that the same idea can be used in causal inference in the presence of time-varying confounding. Persons who are very likely to be selected into a treatment (because their Y_1 predicts Z_2) should be weighted down, and those who are underrepresented (who receive Z_2 despite their Y_1) should be weighted up. This weighting scheme, known as "inverse probability of treatment weighting" (IPTW), will remove the bias potentially caused by Y_1. The key is that Y_1 is *not* used as a covariate to make adjustments. Rather, Y_1 is used to construct a weight. The analytic model looks just like the simple model needed if we had a randomized experiment. Weighting removes the bias.

This idea is generalized in Figure 3. Now we have not only outcomes (the *Y*s) and treatments (the *Z*s) but also covariates of two types: The *X*s are observed covariates and the *U*s are unobserved. Of particular interest are covariates of type X_2. These are time varying. They may be outcomes of earlier treatments and predictors of later treatments.

The key assumptions are two. First, treatment assignment Z_1 at Time 1 does not depend on unobserved covariates U_1 given what we have observed—the baseline outcome Y_0 and other pretreatment covariates X_1. Second, treatment assignment Z_2 at Time 2 is independent of unobserved covariates U_1 and U_2 given the past observables, that is, given the pretreatment covariates X_1, time varying covariates X_2, prior outcomes Y_0 and Y_1, and prior treatment Z_1. In that case, we can construct weights for each time point based on the past observables and obtain unbiased estimates of the impact of the sequence of treatments Z_1 and Z_2 on Y_2 (along with the effect of Z_1 on Y_1).

Incorporating the Social Settings in Which Instruction Occurs

Hong and Raudenbush (in press) showed how to adapt Robins's (2000) methods to the social setting in which instruction occurs. We know that children will migrate across teachers within a school and even between schools as they pass through the elementary grades. A child will have one set of



U_1' indep. of $Z_1 \mid X_1, Y_0$

U_1', U_2' indep. of $Z_2 \mid X_1, X_2, Y_0, Y_1, Z_1$

Figure 3. **A simple model for a sequence of two treatments, with covariates.**

classmates within Grade 1 and then will have a somewhat different set of classmates in Year 2. How can we account for the shifting social memberships characteristic of schooling in order to use Robins's results?

Hong and Raudenbush (in press) employed a special four-level hierarchical model in which children are cross-classified by teachers who themselves are nested within schools. They proved that Robins's (2000) weighting approach can apply in this setting. However, great care is needed in defining and applying the weights in this multilevel setting. They applied the method to study the impact of "intensive math instruction" experienced over 2 years (Grades 4 and 5) based on the Longitudinal Evaluation of School Policy and Change (LESCP). They defined intensive math instruction as characterized by a high level of conceptual demand combined with a large amount of time devoted to math.

Their results are graphed in Figure 4. It turned out that experience in Year 2 was decisive. Although there was some evidence of a positive effect of Year 1 intensive math ($Z_1 = 1$), this estimated effect did not significantly differ from zero. However, the effect of Year 2 instruction ($Z_2 = 1$) was significant and positive. The data were not fully adequate for estimating the joint effect of receiving 2 years of intensive math instruction, because the sample size of students who switched from $Z_0 = 0$ to $Z_1 = 1$ or from $Z_1 = 1$ to $Z_2 = 0$ was not large enough to support a sufficiently precise statistical inference, a result that has implications for research design on sequences: Every sequence must be adequately represented in the sample to uncover "amplifying effects."

SUTVA Once More

Our earlier discussion warned of the uncritical importation of paradigms from medical research into education. Our key concern involved SUTVA—the assumption that each participant possesses one and only one potential outcome under each treatment. One key problem was the problem of no interference between units. We reasoned that the treatment assignment of other children would likely be important in determining the outcome of any particular child. We also reasoned that the teacher who delivers the treatment would likely have an effect on how that child responded to the treatment.

The problem of SUTVA once again emerges as important in the case of time-varying instructional regimes. The essential problem is that a teacher in Year 2 will confront a set of children whose treatment histories vary. This will tend to affect how the teacher enacts the regime. As a result, the treatment history of one's classmates becomes a potentially important contributor to a child's potential response. For example, the effect of being assigned to intensive math instruction in Year 2 may depend not only on one's own treatment assignment in Year 1 but also on the treatment histories of one's classmates.

The comparison with a case in medicine when SUTVA applies straightforwardly is again instructive. Suppose again that we have a new drug for treating blood pressure. A patient will be assigned to receive the new drug in Year 1, so that $Z_1 = 1$. The potential outcome of that patient at the end of Year 1 is then $Y_1(1)$. As before, $Z_1 = 0$ if the patient is assigned to the traditional treatment, in which case that patient will display the Year 1 outcome $Y(0)$. In general, we can say that the patient will display the outcome $Y(Z_1)$ for either $Z_1 = 1$ or $Z_1 = 0$.

In Year 2, a patient once again may or may not receive the new drug, in which case the outcome in Year 2 will be $Y_2(Z_1, Z_2)$. In particular, a patient may be assigned to receive the new drug in Year 2, so that $Z_2 = 1$. The potential outcome of that patient is then $Y_2(0, 1)$ if that patient had the traditional drug in Year 1 and $Y_2(1, 1)$ if that patient had the new drug in Year 1. Similarly, a patient assigned to the traditional treatment in Year 2 will display either $Y_2(1, 0)$ or $Y_2(0, 0)$, depending on treatment assignment in Year 1.

The causal effects of interest flow from these four potential outcomes in Year 2. We have the average causal effect of receiving the new drug in Year 1 only, that is, $\delta_1 = E[Y(1, 0) - Y(0, 0)]$; the average causal effect of receiving the new drug in Year 2 only, that is, $\delta_2 = E[Y(0, 1) - Y(0, 0)]$; and the amplifying effect of receiving the drug in both years, that is, $\delta^* = E[Y(1, 1) - Y(0, 0)] - \delta_1 - \delta_2$. These causal effects are estimated without bias using IPTW if unobserved covariates are sequentially independent of treatment group assignment given past observables, as discussed in the previous section.

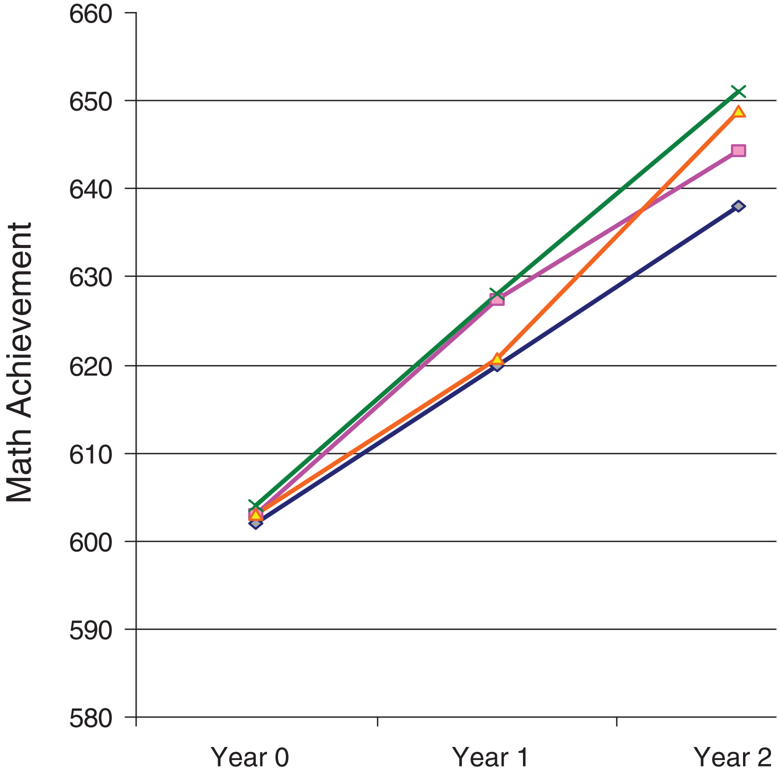


Figure 4. **Estimated trajectories of math outcomes as a function of regime sequence.**

To import this paradigm into research on instruction, we have to assume that one's teacher and other children's treatment assignments do not affect one's potential response to receiving a certain approach to instruction. We saw earlier how this problem was nicely solved by the random assignment of the entire classroom to one of two treatments. However, recall that two additional assumptions were needed: the assumption of no interference between classrooms and also the assumption of intact classrooms. The latter requires that the social membership of the class remains stable, because children do not migrate in response to the treatment.

When we now move to a 2-year sequence of instructional regimes, the assumptions that we made earlier may be reasonable in Year 1. However, by definition, children will migrate to new classrooms in Year 2, and we reason that this will tend to create complications even in a randomized study.

Following Hong and Raudenbush (2006), we might therefore modify our notation as follows. Define the Year 2 outcome as $Y_2(Z_1,\ Z_2,\ f)$, where f

is the fraction of one's Year 2 classmates who received intensive math instruction in Year 1. One's potential outcomes depend not only on the sequence of instructional regimes to which one has personally been assigned but also on the prior treatment history of one's classmates.

This elaborated definition defines a series of causal questions, some of which are important for policy and practice but *that cannot even be defined* if SUTVA is assumed as in the world of clinical trials! First, we have questions of the type

$$\delta_1(f) = E[Y(Z_1, Z_2, f) - Y(Z'_1, Z'_2, f)]. \tag{5}$$

These questions parallel those listed above in the case of the drug trial about the impact of an assignment to a particular sequence Z_1, Z_2 as compared to the sequence Z'_1, Z'_2. What is different is that now we are holding constant the treatment histories of one's classmates in Year 2. We are also interested in causal questions of the form

$$\delta_1(f, f') = E[Y(Z_1, Z_2, f) - Y(Z_1, Z_2, f')]. \tag{6}$$

Here we are asking about the impact of one's classmates' treatment history on one's Year 2 outcome—holding constant one's own treatment history.

Finally of interest are questions about interaction effects: Does the impact of a sequence of instructional regimes depend upon the treatment history of one's classmates? These questions have the form

$$\begin{aligned}\delta_1(f) - \delta_1(f') = {} & E[Y(Z_1, Z_2, f) - Y(Z'_1, Z'_2, f)] \\ & - E[Y(Z_1, Z_2, f') - Y(Z'_1, Z'_2, f')].\end{aligned} \tag{7}$$

Knowing the quantity in Equation 7 would tell us if the impact of a sequence of instructional regimes depends on the treatment histories of one's classmates. It may be, for example, that 2 years of intensive math instruction (compared to having intensive instruction in Year 2) may be most beneficial when all of one's classmates have also had 2 years of such instruction.

This discussion reflects a key feature of instruction that makes teaching distinctly different from—and more complex than—medical practice. The physician takes the history of patients one by one and, in each case, prescribes a treatment. The teacher must assess the instructional history of the entire class and then adopt an instructional approach that is tailored to the collection of histories. This same feature of instruction makes causal inference about instructional regimes more complex than causal inference about clinical practice in medicine.

Conclusions

I began with the assertion that a research agenda focusing on the impacts of alternative instructional regimes has great potential for improving policy. The argument, built on the work of Cohen et al. (2003) entails four propositions.

1. The approach recommended here asks how instructional regimes influence student learning. Questions for policy immediately follow: What resources, incentives, and governance structures cost-effectively support the best regimes? Methodological questions loom large: How does one design research to define and answer questions about the impacts of instructional regimes on student learning, and how does one study the moderating effect of resources, incentives, and governance?
2. Research on instruction has similarities but also fundamental differences from research on clinical practice in medicine. These differences arise from the central role of the teacher as the enactor of any instructional regime and from the intrinsically social nature of classroom instruction. These differences have fundamental implications for research design and analysis.
3. Research on instructional regimes requires a deliberate interplay between experimental and nonexperimental research. Experimental research can reveal the impact of intended instructional regimes on student learning. However, sound nonexperimental methods are needed to understand the impact on children of the instruction they actually experience. Moreover, to study such enacted regimes requires good measures of classroom instruction. When the aim is to study the effect of the intended regime on the enacted regime, low reliability of measurement causes no bias but will weaken statistical power to detect effects. However, when we turn to studies of how enacted instructional regimes affect learning, errors of measurement of the enacted instruction will lead to bias. This bias can be corrected, however, if we have good knowledge of the properties of our measures of instruction.
4. Finally, research on sequences of instruction is central to this agenda. Most of the skills and knowledge we hope children will develop emerge from the cumulative effects of sequences of instruction. Again, there are important parallels with clinical research in medicine, but there is also a fundamental difference. Unlike the physician, who takes a patient's history and develops a plan of treatment tailored to that history, the teacher must consider the instructional histories of the entire class and then develop instructional approaches tailored to the whole class or to subgroups of children. The child's response will, in principle, depend not only on that child's own instructional history but also on the instructional history of that child's classmates. The social nature of instruction makes the teachers' task complicated and also complicates the researcher's attempt to draw causal inferences about the impact of instructional sequences.

If these arguments make sense, it is essential that educational researchers develop distinctive models and methods for studying causal effects on student development as it unfolds in the social settings of classrooms and schools. This effort will require an infusion of well-trained, creative social scientists in education who have strong methodological skills. Such experts must understand how to integrate their statistical expertise with the best available thinking about classroom instruction and student learning.

Notes

The research reported here was supported by a grant from the Spencer Foundation titled "Improving Research on Instruction: Models, Designs, and Analytic Methods." The paper was originally presented as an invited address at the annual meeting of the American Educational Research Association in Chicago. I wish to thank Guanglei Hong, Anthony Bryk, and the editors for comments and suggestions. Earlier work with David Cohen, Deborah Ball, and Brian Rowan laid the foundation for the ideas presented here.

[1] One could set up an experiment to examine two alternative ways of responding to a given level of student knowledge within a regime. Such a study would likely be a small-scale laboratory study, one that would influence the architecture of future regimes. However, this more micro study would be more remote from informing policy than the "between-regime" comparisons emphasized here.

References

Bloom, H. S. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, *29*(1), 30–59.

Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A., Madden, N., & Chambers, B. (2005). Success for all: First-year results from the National Randomized Field Trial. *Evaluation and Policy Analysis, 27*, 1–22.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). *How and why do teacher credentials matter for student achievement* (NBER Working Paper 12828). Washington, DC: National Bureau of Economic Research.

Cohen, D. K., Raudenbush, S. W., & Ball, D. B. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, *25*, 119–142.

Connor, C. M., Morrison, F. J., Fishman, B. J., Schatschneider, C., & Underwood, P. (2007). The early years: Algorithm-guided individualized reading instruction. *Science*, *26*(315), 464–465.

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal*, *27*(3), 557–577.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrika*, *11*, 1–12.

Hanushek, E. A. (1989). The impact of differential expenditures on school performance. *Educational Researcher*, *18*(4), 45–65.

Hanushek, E. A. (1996). School resources and student performance. In G. Burtless (Ed.), *Does money matter? The effect of school resources on student achievement and adult success* (pp. 43–73). Washington, DC: Brookings Institution.

Hedges, L. V. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87.

Hedges, L. V., Laine, R. D., & Greenwald, R. (1994). Does money matter? A meta-analysis of studies of the effects of differential school inputs on student achievement. *Educational Researcher*, *23*(3), 5–14.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrika*, *47*, 153–161.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–960.

Hong, G. (2004). *Causal inference for multilevel observational data with application to kindergarten retention*. Doctoral dissertation, University of Michigan, Ann Arbor.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for observational data. *Journal of the American Statistical Association*, *88*, 221–228.

Hong, G., & Raudenbush, S. W. (in press). Causal inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics*.

Kerbow, D. (2006). *Strategic teaching and evaluation of progress*. Chicago: University of Chicago, Center for Urban School Improvement.

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 results: Evidence from Project STAR. *Economic Journal*, III, 1–28.

Mosteller, F., & Boruch, R. (2002). *Evidence matters: Randomized trials in education research*. Washington, DC: Brookings Institution.

Murphy, S. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society*, *Series B*, *65*(2), 331–366.

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

Neyman, J. (with Iwaszkiewicz, K., & Kolodziejczyk, S.). (1935). Statistical problems in agricultural experimentation (with discussion). *Supplement to the Journal of the Royal Statistical Society*, *2*, 107–108.

Nye, B., Hedges, L.V., & Konstantopoulos, S. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*, 237–257.

Raudenbush, S. W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher*, *34*, 25–31.

Raudenbush, S. W. (2007). *Statistical inference when classroom quality is measured with error*. Chicago: University of Chicago, Department of Sociology.

Raudenbush, S.W., Martinez, A., Bloom, H., Zhu, P., & Lin, F. (2007). *The reliability of group-level measures and the power of group-randomized studies*. Working paper, University of Chicago.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, *9*, 5–29.

Robins, J. (2000). Marginal structural models versus structural nested models as tools for causal inference. In M. Elizabeth Halloran & Donald Berry (Eds.), *Statistical models in epidemiology, the environment, and clinical trials* (pp. 95–134). New York: Springer.

Rosenbaum, P., & Rubin, D. B. (1983). The central role of the propensity score in observational studies of causal effects. *Biometrika*, *70*, 41–55.

Rowan, B. E., Camburn, E., & Correnti, R. (2004). Using teacher logs to measure the enacted curriculum in large-scale surveys: Insights from the Study of Instructional Improvement. *Elementary School Journal*, *105*, 75–102.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, *6*, 34–58.

Rubin, D. B. (1986). Which Iffs have causal answers? *Journal of the American Statistical Association*, *81*, 961–962.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1–22.

Spybrook, J. (2007). *Examining the experimental designs and statistical power of group randomized trials funded by the Institute of Education Sciences*. Unpublished doctoral dissertation, School of Education, University of Michigan, Ann Arbor.