Qual Quant (2013) 47:2225–2257
DOI 10.1007/s11135-011-9652-5

Merging the accountability and scientific research requirements of the No Child Left Behind Act: using cohort control groups

Jean Stockard

Published online: 11 December 2011

Abstract This article shows how assessment data such as that mandated by the No Child Left Behind Act can be used to examine the effectiveness of educational interventions and meet the Act's mandate for "scientifically based research." Based on the classic research design literature a cohort control group and a cohort control group with historical comparisons design are suggested as internally valid analyses. The logic of the "grounded theory of generalized causal inference" is used to develop externally valid results. The procedure is illustrated with published data regarding the *Reading Mastery* curriculum. Empirical results are comparable to those obtained in meta-analyses of the curriculum, with effect sizes surpassing the usual criterion for educational importance. Implications for school officials and policy makers are discussed.

Keywords Research designs · Cohort control groups · Assessment data · Evaluation research

Few would dispute that the No Child Left Behind (NCLB) Act has had an extraordinary influence on education in the United States. School administrators, teachers, and students have probably been most affected by the Act's focus on accountability and using standardized assessments to identify schools needing improvement. A key element of the Act was the requirement that states measure progress in reading and mathematics and compare students' scores to established benchmarks. Educational researchers have probably been more affected by the Act's requirement for "scientifically based research" and the call for "evidence-based practice."

J. Stockard (✉)
Department of Planning Public Policy and Management, University of Oregon, Eugene, OR 97403, USA
e-mail: jeans@uoregon.edu

J. Stockard
National Institute for Direct Instruction, Eugene, OR, USA
e-mail: jstockard@nifdi.org

To a large extent, the work resulting from these two aspects of the Act appears to have developed in parallel, and generally non-overlapping, avenues. The most powerful voices in the educational research community have defined "scientifically based research" in ways that emphasize small, randomized control trials within schools or artificial settings. Specialists in assessment have designed elaborate instruments to measure basic skills, gather data from schools throughout the country, and compare results to established benchmarks defining adequate progress. Yet, the two strands of work are largely unconnected.

This article suggests that the data schools obtain as part of their routine state and federally mandated assessment programs can be used to examine the effectiveness of educational curricula. More importantly, based on the logic developed in the classic research design literature, I suggest that appropriate analysis of these data can approximate the quality of results that could be obtained through randomized control trials of the same curricula. The analysis approach presented is simple, but, I assert, logically valid and "scientifically-based." It could easily be used by local school officials wanting to assess the impact of curricular changes. Thus, the approach directly addresses an issue of utmost concern to school authorities—examining the extent to which changes at their school produce desired results.

Section 1 of this article reviews the classic literature regarding how research designs assess causal relationships (Campbell and Stanley 1963; Cook and Campbell 1979; Shadish et al. 2002). Section 2 builds upon this discussion by showing how assessment data can be used to study the effectiveness of educational interventions in ways that provide high internal validity. Section 3 expands upon this analysis by reporting the results of several dozen analyses from eighteen different sites, following the Campbell et al. tradition of using multiple studies to develop generalizations about causal relationships and thus promote external validity. Section 4 summarizes the work and discusses potential implications for practitioners and researchers.

All of the examples involve applications of the *Reading Mastery* (RM) curriculum. This program is especially appropriate for this discussion because its publisher routinely disseminates reports of assessment scores of schools using the materials. These scores provided the basis of my analysis. The logic of the approach could, of course, be used to evaluate other types of interventions and curricula in institutional settings.

1 Research designs to assess effective programs

Lawrence Sherman (2003) suggested that the "political use" of "evidence-led policy" began in the U.K. in the 1990s. The term first appeared in medicine, apparently in reaction to studies finding that the vast majority of medical treatments in everyday use had never been scientifically tested (Millenson 1997, p. 4, cited by Sherman 2003, p. 7). The concern for such evidence quickly spread to other policy arenas. In education, the response has focused on calls for a greater reliance on experimental designs, particularly those involving randomized assignment (Cook 2002, 2003; Sherman 2003; Towne et al. 2005). In contrast, in medicine the response appears to have also included calls for systematic reviews of accumulated evidence (e.g., Noble 2006). The sections below discuss both of these foci utilizing the classical discussions regarding research design from the social science literature.

1.1 Experimental designs to establish causality

In recent years, building upon the NCLB's mandate for "scientifically based research," the most influential proponent of randomized control trials in education has no doubt been the

Institute of Education Sciences (IES), the "research arm" of the US Department of Education. It has defined randomized controlled trials as the "gold standard" for educational studies, stating that "randomized trials are the only sure method for determining the effectiveness of education program and practices" (Whitehurst 2003, p. 6). The use of these procedures has a strong influence on the ranking of curricula in literature reviews and in determining research funding (Julnes and Rog 2007; National Mathematics Advisory Panel (NMAP) 2008; Scriven 2005; St. Pierre 2006; U. S. Department of Education 2006; What Works Clearinghouse 2008).

The logical argument for randomized control trials is familiar to any undergraduate social science major who has studied experimental design. The literature builds on John Stuart Mill's delineation of three important criteria for inferring causation: co-variation between a cause and an effect, temporal precedence of the cause to the effect, and ruling out other possible causal variables. Expanding upon that logic, Campbell and Stanley (1963) and their successors (Cook and Campbell 1979, and Shadish et al. 2002, which are the later, expanded editions of the 1963 monograph) provided extensive discussions of the variety of designs that could be used in testing causal relationships. These writings have long been considered the standard reference for research design in the social sciences. In the pages that follow I will refer to the work as the Campbell, Cook, Shadish, Stanley (CCSS) approach.

The notion of validity is central to writings of CCSS. This term refers to the accuracy of inferences or conclusions that can be made from the results of a study. The most recent developments distinguish four types of validity:

1. Statistical conclusion validity, regarding the association (correlation) between receiving a treatment, such as exposure to a curriculum, and an outcome, such as achievement;
2. Internal validity, regarding whether the observed association between the treatment and outcome represents a causal relationship;
3. Construct validity, regarding the extent to which inferences from the study can be related to the theory on which it was based; and
4. External validity, the extent to which the results can be extended to other settings and conditions.

The arguments for using randomized control trials in educational research have focused on internal validity, because these designs logically deal with each of the classically defined "threats to internal validity." Because subjects are randomly assigned to either an experimental or control condition, they theoretically experience similar maturational processes, historical events, reactions to testing, etc. Thus, any differences accruing to the experimental group that are different than those accruing to the control group may be attributed to the experimental intervention. Like the general CCSS tradition, the argument stresses the vital importance of study design in producing results that should be used as the basis of social policy.

While accepting the logical argument of this approach, numerous authors have suggested that restricting analyses of educational interventions to randomized control trials may not be appropriate and could result in misleading conclusions. Using the language of experimental design, they suggest that randomized trials, when implemented in real-life settings, may have "many more potential threats to internal validity than would highly controlled quasi-experiments" (McMillan 2007, p. 1).[1] The concerns involve numerous issues ranging from how subjects are assigned to treatment when students must be grouped in classes or schools, to

[1] Concerns have also been voiced regarding the use of randomized control trials in medicine (e.g. Clay 2010; Williams 2010) and criminology (Sampson 2010).

how fidelity of treatment can be ensured in real-life settings where both the implementers (e.g. teachers) and the subjects (e.g. students) may routinely interact.

The classic "gold standard" experiment in a field such as medicine is "double blind," with neither the interventionists (e.g. doctors) nor the subjects (e.g. patients) knowing which group they are in. Yet, such a situation is generally impossible to attain in a field setting such as a school or even a school district with multiple schools. Teachers certainly know the situation and students no doubt also soon understand what is happening. Generations of research have documented the "Hawthorne" and "reverse Hawthorne" or "John Henry" effects, regarding how knowledge of experimental designs influences subject behavior (e.g. Zdep and Irvine 1970). Most importantly, such knowledge, and the potential of its effect, immediately adds another variable to the design. There is, logically, no way for an analyst to determine if differences between the control and experimental group are due to the intervention or to other types of actions by the subjects based on their knowledge of the situation (Cook et al. 2009; McMillan 2007; Raudenbush 2008; Scriven 2008, Scriven n.d.; Slavin 2008; see also Cartwright 2007). In summarizing these concerns, Michael Scriven concluded that "quasi-experimental designs…are alternative ways to establish conclusions, often better ways in particular circumstances" (Cook et al. 2009, pp. 4–5).[2]

A close reading of the classical CCSS writings on experimental designs, as well as more recent commentaries by Cook and Shadish, indicates that the essence of this conclusion also appears in that material, especially in the discussion of "recurrent institutional cycle" or "cohort control group" designs. In this discussion they propose a useful alternative to randomized control trials in organizational settings:

> Many institutions experience regular turnover as one group "graduates" to another level and their place is taken by another group. Schools are an obvious example of this, as most children are promoted from one grade to the next each year….The term cohort designates the successive groups that go through processes such as these. Cohorts are particularly useful as control groups *if* (1) one cohort experiences a given treatment and earlier or later cohorts do not; (2) cohorts differ in only minor ways from their contiguous cohorts; (3) organizations insist that a treatment be given to everybody, thus precluding simultaneous controls and making possible only historical controls; and (4) an organization's archival records can be used for constructing and then comparing cohorts (Shadish et al. 2002, pp. 148–149, emphasis in original; see also Cook and Campbell 1979, pp. 126–127 and Campbell and Stanley 1963, pp. 56–61).

[2] While the psychological (and education) literature seems to generally use the notion of "reactivity," building on the CCSS tradition, the economics and sociological literature appears to also discuss violations of the "stable unit treatment value assumption" (SUTVA), using the logic of counterfactuals and the tradition of causal analysis developed by Donald Rubin (See Cook and Steiner 2010, Maxwell 2010, Rubin 2010, Shadish 2010, and West and Thoemmes 2010, for discussions of the differences of the two approaches). Writers using these terms note that the SUTVA assumption might be met with drug trials where the impact of a dose on one patient is independent of the dose that another patient receives. In schools and classrooms, however, administration of a treatment (such as instruction) can involve different forms of applications for each unit (students) given the impact of the multiple influences (other students and variability in teachers) in their environment (Raudenbush 2008). Put another way, Sloane (2008b), an educator and using experimental design terms, suggests that the nesting of data within schools results in variability in both fixed effects (differences in means of the dependent variable) and random effects (variability in the dependent variable). The traditional randomized control trial model only considers the former, potentially leading to misleading results and difficulties in generalizing to other contexts (See also Yin and Davis 2007). A large literature, largely within sociology and economics, has addressed methodological issues of causal inference when using observational data and the logic of the counterfactual approach (See Morgan and Winship 2007 and Winship and Morgan 1999 for especially readable discussions of this area).

Thus, the classic CCSS writings on research design explicitly note that there are valid alternatives to random assignment, especially in field settings, and that these designs can counter the "reactive effects" endemic to employing random assignment in institutional settings (Campbell and Stanley 1963, p. 57).[3]

As described more fully below, this cohort control group (CCG) design could be used by schools to examine the efficacy of new curricular programs for their own particular context. Routinely collected assessment data could be used to compare the achievement of students in cohorts that experienced the new curriculum with those that did not. In contrast to randomized control trials, which may reflect results on relatively few students in settings often quite dissimilar from a particular school district, this design could be explicitly applied to the actual context of a given school or district and use data that were already routinely collected (cf. Engelmann 2009). This concern relates to the notion of generalization.

1.2 The cumulative nature of science and generalizations

In a lengthy discussion of the role of experimental designs in establishing causality, Shadish and associates noted that

> the strength of experimentation is its ability to illuminate causal inference. The weakness of experimentation is doubt about the extent to which that causal relationship generalizes. … [E]xperiments yield hypothetical and fallible knowledge that is often dependent on context and imbued with many unstated theoretical assumptions (Shadish et al. 2002, pp. 18, 29).

The CCSS discussion of construct validity and external validity directly relates to their concerns with generalization. As noted above, construct validity refers to the relationship between a research design and the theory it is designed to test. For instance, a study of the impact of a particular mathematics curriculum on children's arithmetic skills might utilize assessment operations that involve "story problems," thus assuming certain reading skills. One could argue that the study might lack construct validity, as it applied to concerns with mathematics achievement, for it is not clear if math or reading skills were being assessed. External validity refers to the extent to which a causal relationship holds across a variety of settings and populations. As Shadish and colleagues put it, "external validity…[regards] the extent to which the effect holds over variations in persons, settings, treatments, or outcomes" (Shadish et al. 2002, p. 22).

CCSS stressed that sampling methods "are insufficient to solve either problem of generalization" (construct or external validity) (Shadish et al. 2002, p. 24). In an extended discussion they concluded that even relatively complex models such as random selection of units from a population followed by random assignment to treatments

> cannot be advocated as *the* model for generalized causal inference….Though we unambiguously advocate it when it is feasible, we obviously cannot rely on it as an all-purpose theory of generalized causal inference. So researchers must use other theories and tools to explore generalized causal inferences of this type (2002, p. 348, emphasis in original).[4]

[3] These criteria are very similar to those discussed in a recent paper by Shadish and Cook (2009).

[4] The call made by some authors for place-based randomization in educational contexts (e.g. Boruch 2005; Cook 2005; Fayer 2005) could be an example of the techniques to which Shadish and colleagues were referring.

They proposed a "grounded theory of generalized causal inference," which they suggested "is more practical than random sampling for daily scientific work," noting that it builds on numerous conceptualizations developed over the last half century (2002, p. 348, citing Brunswick 1956; Campbell 1986; Cook 1990, 1991; Cronbach 1982; Cronbach and Meehl 1955). The theory embodies five general principles that they suggested scientists routinely utilize as they make causal generalizations (See pp. 353–354 of Shadish et al. 2002, for a summary). The first is called the principle of "surface similarity," which involves "judging the apparent similarities between the things that they studied and the targets of generalization" (p. 353). For instance, an educational researcher reviewing numerous studies regarding a curricular approach might find that the same results obtain across applications in mathematics and reading. The second principle, "ruling out irrelevancies," involves determining what types of factors do not seem to make a difference, as for instance finding that conclusions hold in a variety of settings (rural vs. urban) or with populations of different racial-ethnic backgrounds. In contrast, the third principle, "making discriminations," involves finding discriminations that limit generalization, as in determining that an intervention appears to be more effective with younger children than with older children. The fourth principle, "interpolation and extrapolation," refers to the ways in which scientists may extend findings to groups that are within a sampled range but not yet studied (e.g. children in grade three when evidence is from those in grades two and four) or, more difficult to do, extending to those outside the sample range (e.g. children in grades three to four when the evidence relates to those in grades one to two). Finally, the fifth principle, termed "causal explanation," refers to the development and testing of explanatory theories about the target of generalization, whether they are persons, settings, treatments, or outcomes.

Shadish and associates then described how scientists can develop generalized causal inference, discussing techniques for single studies and multiple studies, or programs of research. For those engaging in a single study, they stressed the centrality of "purposive sampling." One approach is to purposively sample "typical instances," choosing samples of people, settings, times, treatments, and/or outcomes to which one wants to generalize. For instance, a school might be particularly interested in the extent to which a given curriculum affects the reading achievement of students in particular grades in their own localities using test instruments that they have previously found to be highly predictive of later success. Another alternative is to purposively sample "heterogeneous instances," defining the range of persons, settings, treatments or outcomes to which one wants to generalize and creating a sample that reflects that heterogeneity. For instance, a school district might wish to examine the extent to which an intervention impacts the achievement of children from different economic backgrounds and would thus ensure that the study sample included the full range of possibilities.

It is, of course, very difficult for one study to encompass the range of persons, settings, times, treatments, and outcome measures needed to develop causal generalizations, but multiple studies can provide this diversity. Sometimes multiple studies involve the gradual development of understandings and causal inferences, through a "phased model," typically moving from rather small and highly focused controlled experimental designs to tests with more varied settings, subjects, and outcomes. Through purposive sampling these series of studies can address the five principles of causal explanation described above. The Direct Instruction (DI) curriculum, which includes our targeted program of RM and was first developed in the 1960s, illustrates this approach. The curriculum developed through lengthy and detailed experiments with different populations of students and in different settings. Extensive work also validated the principles of learning and instruction that provide its theoretical base. Field testing with large and varied populations of students and teachers examined the extent to which it was

effective across heterogeneous populations (Collins and Carnine 1988; Engelmann 2007; Engelmann and Carnine 1982; Huitt et al. 2009).

The phased model of research is both time-consuming and expensive. Thus, systematic literature reviews are a much more common approach to using multiple studies to develop causal generalizations. In recent years, meta-analyses using quantitative techniques to summarize the average results of studies have become increasingly common. Typically these studies convert the outcomes from each study to a common "effect size" metric. Using the logic embedded in the five principles outlined above, meta-analysts can look for similarities and variations across studies, find gaps in the literature, and test for moderating effects. While meta-analyses are not immune from potential problems related to both internal and external validity (Shadish et al. 2002, pp. 446–454), it is clear that multiple examinations of a research problem are superior to single tests. As Cook and Campbell put it, "we stress the need for *many* tests to determine whether a causal proposition has or has not withstood falsification; such determinations cannot be made on one or two failures to achieve predicted results" (1979, p. 31, emphasis in original).

In the later volume Shadish, Cook, and Campbell noted that

> Among scientists, belief in the experiment as the *only* means to settle disputes about causation is gone, though it is still the preferred method in many circumstances. Gone, too, is the belief that the power experimental methods often displayed in the laboratory would transfer easily to applications in field settings (2002, p. 30, emphasis in original).

They described their work as

> about improving the yield from experiments that take place in complex field settings, both the quality of causal inferences they yield and our ability to generalize these inferences to constructs and over variations in persons, settings, treatments, and outcomes (Shadish et al. 2002, p. 32).

My analysis attempts to follow in this tradition by showing how the data routinely gathered by schools in response to the accountability and assessment movement can be used to test the effectiveness of curricular interventions. I use the concepts developed within the CCSS tradition to show how these analyses can provide valid tests of causal relationships. I first discuss two designs that could be used in single studies. Then, in the following major section, I summarize results from multiple analyses that used these designs, following the logic of CCSS regarding causal generalizations.

2 Using assessment data to examine the impact of curricular change

In this section I discuss two ways in which the assessment data often available to school personnel can be analyzed to examine the impact of implementing a new curriculum. Both of these designs use cohort groups. The difference is in the use of additional comparative data, with one using a simple CCG design and the other employing time series data over the same historical period as the cohort design as an additional control. In each of the sections I discuss how the designs deal with validity issues described in the CCSS tradition. The data for the examples came from an SRA/McGraw-Hill publication entitled *Results with Reading Mastery*, which reported results that schools around the country had with this program. For this section I use data from an elementary (K-5) school in rural Alaska, with slightly more

Table 1 At risk status based on end of year kindergarten DIBELS scores, 2002–2003 and 2006–2007, rural Alaskan elementary school

	2002–2003	2006–2007	z	*p* (one-tail)	Effect size	Min. CI
At risk	22	2	2.77	0.006	0.62	0.25
Some risk	32	12	2.17	0.002	0.48	0.12
Low risk	46	87	4.15	<.0001	0.87	0.52

Details on calculations are given in Appendix A. RM had not been implemented in 2003–2003, but had been implemented for 4 years by the end of 2006–2007. One-tail probabilities were used to match the question of theoretical interest

Table 2 CCG design

Cohort 1	O_1		
Cohort 2		X	O_2

than 300 students. Almost three-quarters of the students were Caucasian and one-quarter were of Native American or Alaskan Native descent.

2.1 CCG design

One set of data from this school was changes in Dynamic Indicators of Basic Early Literacy Skills (DIBELS) scores of kindergarten students from the 2002–2003 year, before RM was implemented, to the 2006–2007 year, after four years of implementation. The percentage of students classified as at "low risk," "some risk," and "at risk" for poor language and reading outcomes in both years based on the established DIBELS benchmarks were given (See Table 1). Based on their overall scores, 22% of the students in 2002–2003 were classified as at risk, but only 2% had this classification in 2006–2007. At the other end of the spectrum, less than half of the kindergarten students were at low risk of future difficulties before RM was implemented, but 87% were in this category four years later.

To analyze the changes in DIBELS scores over time we can employ a CCG design (Cook and Campbell 1979, pp. 126–133; Shadish et al. 2002, p. 137 and called "recurrent institutional cycle design" by Campbell and Stanley 1963, pp. 57–60), the logic of which is diagrammed in Table 2.[5] Following the custom used by Campbell and Stanley (1963) and successors, I use X to refer to the intervention (the curriculum that the school used—RM) and O_i to refer to the observation or assessment (DIBELS). Each line represents a cohort in the school. For this analysis Cohort 1 attended kindergarten before RM was implemented (the 2002–2003 school year), while Cohort 2 attended kindergarten after the program was established (2006–2007). Thus, data are available for a cohort that had not been exposed to the experimental intervention (O_1) and for a cohort with such exposure (O_2), and the data from O_1 and O_2 can be compared to determine the impact of the intervention (X).

A comparison of proportions test can be used to examine the statistical significance of the differences between these two years. The results for each measure are in the fourth and fifth columns in Table 1 and indicate that all the differences were statistically significant. Also included is an effect size for each comparison (Cohen's d), and each of these surpassed the usual criterion of educational importance, ranging from 0.48 to 0.87. Finally, the lower limit of a one-tailed confidence interval is reported. This may be interpreted as the

[5] Miron (2005) briefly described a similar design that he called "successive cohorts," but did not cite the CCSS tradition or use the analysis approach that is described below.

lower limit of an expected effect size at 95% confidence (e.g. for the "at risk" comparison, $P[D > 0.25] = 0.95$, where D is the effect size associated with the difference in at risk status between the two cohorts in a hypothetical larger population of cohorts). The mathematics involved in these comparisons is simple and can be completed with an excel spreadsheet or an on-line calculator. Details on computations are in Appendix A.

To what extent is this design valid? As noted above, the CCSS literature distinguishes four types of validity with which researchers should be concerned: statistical conclusion, construct, external, and internal validity. As described above, statistical conclusion validity refers "to the appropriate use of statistics to infer whether the presumed independent and dependent variables covary" (Shadish et al. 2002, p. 37). Most researchers now agree that the best way to handle these issues is to report both inferential and descriptive statistics (such as effect sizes) and to report complete details of the inferential results, as was done above with both the z value and the associated exact probability. Construct validity regards the extent to which inferences from the study can relate to the underlying theory. For this analysis the key question is no doubt the extent to which the DIBELS measure accurately reflects and measures the school's concern with increasing reading skills. Numerous studies indicate the predictive validity of DIBELS measures, in terms of both later reading achievement and comprehension (Fuchs et al. 2001; Good et al. 2001). External validity refers to the extent the results can be applied to other settings and conditions. While it is unclear to what extent the findings from this small community in Alaska could apply to other settings, the issue for the district itself is probably more limited. For school officials, the question of external validity probably involves the extent to which the findings can apply to later cohorts of students. This issue of comparability of cohorts is central to determining the final type of validity—internal validity.

No doubt the trickiest element of the CCG design revolves around internal validity, inferences regarding whether the observed relationship represents a causal relationship. In response to misunderstandings regarding his original use of this term, Campbell (1957) suggested relabeling the concept as "local moral causal validity." Building on this suggestion, Shadish et al. stressed that "internal validity is about whether a complex and inevitably multivariate treatment package caused a difference in some variable-as-it-was-measured within the particular setting, time frames, and kinds of units that were sampled in a study" (2002, p. 54). In other words, the important question is, within the particular setting of this Alaskan community, as RM was implemented with these two cohorts, did this implementation result in changing reading skills?

The list of potential threats to internal validity should be familiar to all social scientists: (1) ambiguous temporal precedence, (2) selection, (3) history, (4) maturation, (5) regression, (6) attrition, (7) testing, (8) instrumentation, and (9) additive and interactive effects of the preceding threats (Shadish et al. 2002, p. 55). Several of these threats would appear to be unlikely with the CCG design. For instance, temporal precedence issues would appear quite unlikely, as long as the new curriculum was implemented after Cohort 1 finished kindergarten. Following the standard testing protocols would help to minimize testing and instrumentation effects. Each of the other threats, however, relates directly to the extent to which the cohorts involved are comparable. In other words, with reference to this example, did kindergarten students in 2006–2007 differ systematically from kindergarten students in 2002–2003?

As noted above, CCSS provided explicit guidance on this issue. I suggest that the design easily meets three of their four criteria: (1) the earlier cohort did not have the treatment (RM), while the later cohort did; (3) the treatment given to the later cohort applied to all students, and (4) archival records provided comparable data (the DIBELS scores). The most difficult potential issue is their second criterion, ensuring that the "cohorts differ in only minor ways

from their contiguous cohorts" (Shadish et al. 2002, p. 149). For instance, changes in reading skills could reflect changes in the demographic characteristics of cohorts if the community experienced dramatic changes in the composition of its residents or if the school had a dramatically altered student body over time. School officials would, of course, be in the best position to judge such changes (and see the discussion in the next section for examples). In addition, however, empirical data could be used to address this question. First, schools routinely gather data on the percentage of enrolled students who receive free and reduced lunch and are members of racial-ethnic minorities, common measures of "at-risk" status, and these data could be examined to determine any differences in demographic characteristics. Second, most schools now have excellent data on the mobility and turn-over of student cohorts and the extent to which members of cohorts have had a "full dosage" of a given intervention. Third, and applicable to this example, the DIBELS system provides measures of skills at the start of schooling, and comparisons of these scores between cohorts would help to establish any differences. If school officials could be convinced that the cohorts were essentially equivalent, then one could logically argue that threats from selection, maturation, regression, and attrition would be minimized.

A more difficult question involves threats to internal validity from history, the impact of events that occur naturally or concurrently with the treatment. Might there be circumstances or events within the broader social-historical context that contributed to these changes other than changes in the school curriculum? Given the strong legislative and political impetus for school improvement around the nation it would be important for a school to know how they were faring relative to other districts. In other words, even if achievement in this Alaskan school improved after implementing RM, is this improvement greater than what was observed in other schools over the same historical period? This question is addressed by our second design: the cohort control group with comparative historical data (CCG-H) design.

2.2 CCG-H design

The logic of the CCG-H is diagrammed in Table 3. Like the CCG design, this design compares outcomes between cohorts with differential exposure to the treatment. Two cohorts from the same setting are compared, with one cohort not exposed to the treatment and the other receiving the treatment. In addition, however, to provide a control for historical change, comparisons are made with cohorts from a larger group that provides an appropriate reference (This logic is analogous to the modification of the CCG design described by Cook and Campbell 1979, p. 130 and Shadish et al. 2002, pp. 149–150). For schools, this larger group could logically be students within the same state or district, who would presumably be affected by similar opportunities and restrictions related to political and funding issues. This additional comparison group is shown in the bottom lines of Table 3, with observations paralleling those obtained from both of the cohorts. The logic involved in determining the equivalence of school cohorts for the CCG design would apply here. The addition of the larger comparison group, however, allows a control for history effects. To be considered significant relative to changes in the larger historical context, differences between O_2 and O_1

Table 3 The CCG-H design

Cohort 1	O_1		
Cohort 2		X_2	O_2
Comparison group cohort 1-C	O_3		
Comparison group cohort 2-C			O_4

would have to be greater than differences between O_4 and O_3, the comparable time points for these observations for the comparison group.

As an example I again use data from the Alaskan school, looking at the percentage of third graders who scored at the proficient or advanced level in reading and writing in the state's annual assessment test. I compare the change in third graders' scores from before the school began to use RM (spring, 2003) to 2007, when third graders continuing in the school would have been exposed to the curriculum since kindergarten, to changes over this time period within the state as a whole. I test the hypothesis that the changes over time in the community were greater than the changes over time in the state or, equivalently, that the difference between scores of students in the community and those in the state of Alaska was greater in 2007 than in 2003. In other words, I examine the extent to which third graders' reading achievement increased in the target school after the implementation of the curriculum *relative to increases for all third graders within the state of Alaska*.

The first panel of Table 4 reports data provided by the SRA report and the comparable state assessment data, which was obtained from the web site for the Alaska State Department of Education. In 2003 the percentage of the school's third graders scoring proficient or advanced was very similar to the percentage for the state in both reading (74–75%) and writing (60%). However, by 2007, the percentages in the school were higher than those in the state (86% vs. 80% for reading and 89% vs. 77% in writing). The final column of the first panel gives z-scores (standard scores) comparing the proportions in the school to those for all third graders in the state. As would be expected given the greater increase in scores in the school, these values were near zero for 2003, but positive for 2007.

The second panel of Table 4 reports effect sizes describing the change in scores from 2003 to 2007 in the school relative to the change in scores in the state. While both effect sizes were positive, only the one associated with writing reached the usual level used to denote educational importance. Also included are t-values testing the null hypothesis that the changes from 2003 to 2007 in the school, relative to changes in the state, equaled zero.

Table 4 Changes in third graders reading and writing scores on state assessments, 2003–2007, an Alaskan elementary school and state of Alaska

Proportion of students reaching proficient or advanced benchmarks			
	School	Alaska	*Z*-score
Reading			
2003	0.75	0.74	0.01
2007	0.86	0.80	0.15
Writing			
2003	0.60	0.60	0.00
2007	0.89	0.77	0.29
	Reading	Writing	
Change 2003 to 2007			
Effect size	0.13	0.28	
t-value	0.73	1.55	
df	102	102	
Prob.	0.24	0.06	
Min. CI	−0.17	−0.02	

There were 41 third graders in 2003 and 63 in 2007. In 2003 the third graders had no exposure to RM; in 2007 third graders in the school had been exposed to RM since kindergarten. Probabilities are one-tail, reflecting the hypothesis that exposure to RM will increase achievement

The probability level associated with the change in writing scores approached significance, while that associated with the change in reading scores did not. As would be expected given these t-values, the lower bounds of the uni-dimensional 95% confidence intervals were less than zero (Details on computations are in Appendix A).

Note that the design used in this analysis explicitly controls for the "history" threat to internal validity. By comparing changes in the community to those that occurred within the state as a whole it adjusts for the changes that were occurring within the state. As described more fully in Appendix A, because achievement scores were increasing in both the community and the state, the estimates of change in the community would have been too high if this control were not included. If scores in the state had been declining during this time period omitting the control would have under-estimated the magnitude of change in the school.

2.3 Summary

This section demonstrated analyses with two research designs that schools can use to test the impact of curricular interventions on student achievement with assessment data that they routinely gather. Both designs compare the achievement of student cohorts, an approach that the classical literature suggests is appropriate in institutional settings such as schools. Most important, in these settings these designs could provide more internal validity than randomized control trials because they may minimize the possibility of reactivity, a serious concern when conducting experiments within intact organizations. I also showed how, with some types of data, this design can be expanded to include comparative data from similar cohorts in a larger entity representative of the sample groups, such as their state or perhaps other schools in their district. This CCG-H design thus helps promote internal validity by helping to control for historical changes that might be affecting the cohorts. In short, I suggest that these designs provide valid ways to assess the impact of curricular innovations and avoid the possible reactive effects that can occur by attempting to randomly assign students within intact groups.

However, as described previously, the CCSS research design literature stresses that the results of one study are far from sufficient to determine a causal relationship. While individual schools and districts may be convinced by the results of their own intervention and analysis, other schools and districts, as well as researchers, would want additional evidence. They would be concerned with "external validity," the extent to which results of one study may generalize to other students, settings, and outcomes. I suggest that the logic involved in Shadish et al.'s (2002) "grounded theory of generalized causal inference" can be used to develop this understanding. This can occur by examining the results of multiple observations, numerous tests of the efficacy of a curriculum or intervention. In the next section I use this logic to examine several dozen analyses of the implementation of RM from around the country using the CCG and CCG-H designs.

3 Generalizing from replications

My analyses come from data in the report disseminated by SRA/McGraw Hill, the publisher of RM. I use the logic outlined by Shadish et al. (2002) to compare results obtained across different settings and assessment techniques, examining the extent to which they were similar across schools with different student characteristics and modes of assessment. Further details on the results for each site are given in Appendix B.

Table 5 Characteristics of SRA/McGraw Hill reports on RM interventions

Report #	Design	# of comparisons	State	Grades in school	# of students	Assessment	Grades for analysis
1	CCG & CCG-H	5	Alaska	K-5	300	CBM and SA	K, 3
2	CCG-H	4	Alabama	K-6	470	SA	3–6
3	CCG	1	Arizona	PreK-3	813	NR	3
5	CCGH	6	Colorado	K-5	135	SA	3–5
6	CCGH	1	Delaware	PreK-7	160	SA	3
7	CCGH	11	Florida	Pre-K-5	723	SA	3–5
8	CCG	5	Florida	K-6	593	NR	3–6
10	CCGH	1	Florida	K-5	1577	SA	4
11	CCGH	1	Florida	PreK-5	554	SA	4
13	CCGH	1	Kentucky	PreK-5	250	SA	4
14	CCG	5	Kent. (5 schools)	Pre-k-6	318 (average)	NR	3
15	CCG	4	Minnesota	K-12	216	CBM	K-3
17	CCGH	6	N. Carolina (2 schools)	gr 3–5	675	SA	3–5
18	CCGH	2	Ohio	K-6	208	SA	4
19	CCGH	3	Oregon	K-12	2059 (district)	SA	3
21	CCG	1	Tennessee	K-12	9918 (district)	SA	3–8, SPED only
23	CCG & CCGH	3	Washington	K-6	652	NR	3–4
24	CCGH	1	Wisconsin	PreK-5	374	SA	4

Reports numbered 7 and 19 provided information that allowed comparison of students with disadvantaged status. Report numbered 17 also gave data for a third school, but provided no information for the period before implementation of RM, so those data were not used. Numbers are not contiguous because some of the 26 reports that were reviewed did not have data appropriate for the CCG or CCG-H analysis. CBM refers to curriculum-based measures, SA refers to state assessments, and NR refers to norm-referenced tests

Of the 26 reports that were reviewed, eight were deemed to fall outside the parameters of the CCG or CCG-H design because they did not meet the criteria outlined above.[6] The analysis below focuses on the remaining 18 reports. The second column of Table 5 reports the study design that was determined appropriate for the data. Five of the reports had data appropriate for the CCG design, eleven had data appropriate for analysis with the CCG-H design, and two provided data appropriate for both designs.

Table 5 also summarizes the characteristics of the 18 reports that were reviewed, including geographic location, the grades served by the school or district, the number of students, and the type of assessment used. It can be seen that the reports came from all regions of the country and involved data from individual schools as well as several schools within one district.

[6] Three reports examined the growth over one year of small numbers of special education students, rather than achievement of larger cohorts over multiple years. One report reported data for one cohort over only one year, which precluded comparisons between cohorts and years. Two sites implemented the programs with only a sub-set of students, but reported data for the entire school, and thus it was not possible to match the intervention with the available data. Finally, two districts had been using the programs for an extended period of time and did not provide data from before the implementation began.

The schools varied in size from just over 100 to over 1,000. Assessments described included curriculum-based measures (DIBELS in all instances), norm-referenced tests such as the ITBS and CTBS, and state assessments. Data were given for students in grades K-6 as well as special education students (grades K-8), but most often involved third and fourth graders, perhaps because that is often the age at which state assessments are first given. All of the schools implemented RM, and reading achievement was the focus of the assessments. Seven of the sites also used various other Direct Instruction programs for areas such as language, spelling, and/or interventions for older children.[7]

Most of the reports provided information about several grades and/or schools and thus yielded a number of possible comparisons (60 in all). However, comments within two of the reports indicated that some comparisons would be less valid than others, specifically addressing the second criterion listed by Shadish and associates—that the "cohorts differ in only minor ways from their contiguous cohorts" (Shadish et al. 2002, p. 149). In these situations, the school officials noted that one of the cohorts for which data were presented differed from other cohorts due to an unexpected influx of new students. To account for this possible non-equivalence, I completed analyses with and without these potentially incomparable cohorts. The sample that omits the incomparable cohorts is referred to below as the "reduced sample."

Table 6 provides summary descriptive statistics on the characteristics of the schools (or, in some cases, districts) that the students attended and the effect sizes associated with the analyses. The top part of the table reports data when individual comparisons were used as the unit of analysis, while the bottom part reports data using the report (site) as the unit of analysis. Thus the information in the top part of the table can represent multiple analyses from a school or district, while the information in the bottom reflects one summary number for each site. The data indicate that the analyses involved schools and districts with a wide range of characteristics, thus providing the large number and variety of comparisons required by the logic of the grounded theory of generalized causal inference. For instance, they ranged from schools where less than a quarter of students qualified for free or reduced lunch to those where all students qualified. Similarly, they varied in racial-ethnic composition, from schools that were predominantly minority—Hispanic, African American, or Native American—to those that were virtually all Caucasian in composition. They also varied in the percentage of students who were English Language learners. On average, the sites had implemented RM for four to five years, but ranged from as little as two years to nine. The comparisons involved students who had begun their instruction in RM in kindergarten through ones involving students who began the program only in the upper elementary grades. Some involved only a few students (a minimum of 34), while others involved well more than a thousand. The average was slightly less than 200.

The last two lines of each panel of Table 6 give summary statistics regarding the effect sizes for both the total sample and the reduced sample, which omitted the cohorts reported to be non-comparable. The effect sizes ranged from −0.26 to 1.66 for the single analyses and 0.04 to 1.14 for the analyses aggregated to the site/report level. As would be expected, the minimum values were slightly higher in the reduced samples. However, the means changed only slightly and ranged from 0.47 to 0.57 across the four sets of analyses, well above the level usually considered educationally important.

[7] Sites 8 and 21 also used *Corrective Reading,* an intervention program for older students. Site 17 used *Corrective Reading* for older students and *Language for Learning* with younger students. Site 18 used *Language for Learning* , Site 24 used *Language for Learning* and *Language for Thinking*, Sites 6 and 23 used Spelling programs, and Site 6 also used *Reasoning and Writing.* The analysis reported below examined variations in effects between these sites and others and found minimal differences.

Table 6 Descriptive statistics, case study reports

	Min.	Max.	Mean	SD
Comparisons as unit of analysis (n = 60)				
School/district size	135	9918	713.6	1282.9
FRL %	21	100	68.3	21.0
African American %	0	94	17.1	22.3
Caucasian %	0	97	54.5	30.0
Hispanic %	0	90	17.4	20.9
Asian %	0	7	1.7	2.7
Native American %	0	100	9.2	25.4
ELL %	0	58	14.8	20.1
Years RM implemented	2	9	3.7	1.7
Grade last cohort began RM	0	5	1.1	1.5
N in comparison	34	1800	192	251
Effect size (d)	−0.26	1.66	0.47	0.38
Effect size (d), reduced sample ($n = 58$)	0.00	1.66	0.48	0.37
Site as unit of analysis (n = 18)				
School/district size	135	9918	1108.3	2253.9
FRL %	21	100	67.8	25.4
African American %	0.0	94.0	23.1	31.8
Caucasian %	0.0	97.0	51.4	34.0
Hispanic %	0.0	90.0	16.4	24.4
Asian %	0.0	7.0	1.1	2.0
Native American %	0.0	100.0	7.6	23.8
ELL %	0.0	58.0	11.1	20.1
Years RM implemented	2.0	9.0	4.40	2.35
Grade last cohort began RM	0.0	3.5	0.99	1.31
N in comparison	36	3236	649	824
Effect size (d)	0.04	1.14	0.56	0.34
Effect size (d), reduced sample	0.14	1.14	0.57	0.34

The data with site as the unit of analysis represent the average value across the individual comparisons within a site. The reduced sample omits the cohorts noted in the SRA report as being non-comparable to other cohorts. $N = 58$ for the comparisons for the reduced sample, and n remains at 18 for the site statistics

The grounded theory of generalized causal inference described by Shadish et al. (2002) and summarized above calls for multiple comparisons of results, looking for similarities and differences across numerous studies, searching for factors that might suggest moderating influences or conditions that limit the generalizability of findings. The wide range of settings in the SRA report allows the comparison of findings across schools of different size, with different socio-demographic student characteristics and varying experience with the curriculum, including the years of implementation and utilization of other programs. They also allow us to examine the relation of design characteristics to the results. Table 7 provides a summary of these comparisons. It reports the average effect size calculated for different sets of comparisons, using the four sample groups used in Table 6. The last rows of the table report the maximum and minimum average values across the subgroups.

Table 7 Effect sizes by sample and sub-groups

	Comparisons as unit of analysis—all cases		Comparisons as unit of analysis—reduced sample		Sites as unit of analysis—all cases		Sites as unit of analysis—reduced sample	
	Mean	*N*	Mean	*N*	Mean	*N*	Mean	*N*
School/site size								
135–469	0.64	27	0.65	26	0.74	8	0.76	8
469–9918	0.32	33	0.34	32	0.41	10	0.42	10
Free and reduced lunch %								
21–71	0.47	26	0.47	26	0.50	8	0.50	8
72–100	0.47	34	0.48	32	0.61	10	0.63	10
% African American								
0–29	0.48	51	0.48	51	0.55	13	0.55	13
30–94	0.42	9	0.49	7	0.57	5	0.63	5
% Caucasian								
0–49	0.47	34	0.48	32	0.61	10	0.63	10
50–97	0.47	26	0.47	26	0.50	8	0.50	8
% Hispanic								
0–29	0.50	39	0.52	37	0.55	14	0.57	14
30–90	0.40	21	0.4	21	0.58	4	0.58	4
% ELL								
0–29	0.52	37	0.54	35	0.55	14	0.57	14
30–58	0.40	21	0.40	21	0.58	4	0.58	4
Comparison cohort began RM in kindergarten								
No	0.39	27	0.41	26	0.53	7	0.54	7
Yes	0.51	32	0.51	31	0.48	9	0.50	9
School implemented RM for 4 or more years								
No	0.48	31	0.50	30	0.64	9	0.65	9
Yes	0.45	29	0.45	28	0.48	9	0.50	9
Number of students in comparisons								
22-100	0.76	18	0.77	17	0.92	2	0.92	2
101-1800	0.34	42	0.35	41	0.51	16	0.53	16
Design type								
CCG	0.68	20	0.68	20	–	–	–	–
CCG = H	0.36	40	0.37	38	–	–	–	–
Type of assessment								
CBM	0.79	7	0.79	7	–	–	–	–
Norm refer.	0.58	12	0.58	12	–	–	–	–
State assess.	0.38	41	0.39	39	–	–	–	–
Used other DI programs								
Only RM	0.49	42	0.51	41	0.55	11	0.56	11
Other DI	0.41	18	0.40	17	0.57	7	0.59	37

Table 7 continued

	Comparisons as unit of analysis—all cases		Comparisons as unit of analysis—reduced sample		Sites as unit of analysis—all cases		Sites as unit of analysis—reduced sample	
	Mean	*N*	Mean	*N*	Mean	*N*	Mean	*N*
Summary								
Minimum	0.32		0.34		0.41		0.42	
Maximum	0.79		0.79		0.92		0.92	

Averages for categories of design type and type of assessment could not be computed when the site was the unit of analysis because some sites had multiple designs and assessment types

The results in Table 7 indicate a good deal of consistency across the various analyses. In all of the sub-groups the average effect sizes surpassed the usual criterion for educationally important effects. At the same time there were some variations. The largest differences appear to be those related to the design and assessments used. As would be expected, effects with the CCG-H design, which explicitly controls for historical changes, were smaller than those with the CCG design. The analyses using state assessments, virtually all of which used the CCG-H design, also yielded slightly smaller effect sizes. In addition, effect sizes tended to be higher in analyses of smaller schools and sites. There were very small differences in effect sizes of schools with different socio-demographic characteristics or with those with different levels of experience with the curriculum. In other words, using terms from the CCSS research design literature, these multiple comparisons suggested that the results regarding the efficacy of the RM curriculum appear to generalize to schools with different compositions and settings and with different types of assessments.

4 Summary and discussion

This article used the classic research design literature to demonstrate how assessment data that are routinely collected by schools can be used to analyze the impact of curricular interventions on student achievement. This work shows how two areas of work fostered by the NCLB Act—the requirements of regular assessment of student achievement and the call for "scientifically based research"—can be merged in a way that provides useful information to school officials and researchers in a cost-effective manner. Following the suggestions of Campbell and Stanley (1963), Cook and Campbell (1979), and Shadish et al. (2002), the classic textbooks on experimental design, I showed how comparisons of intact cohorts within schools can result in designs that help counter the problems of reactivity that are likely to appear when randomized trials are implemented within institutional settings and thus promote internal validity. In addition, I used the logic of Shadish, Cook, and Campbell's grounded theory of causal inference to systematically compare several dozen tests of the efficacy of one curriculum across settings with widely differing characteristics, promoting external validity or generalization of results. Below, I briefly discuss how the work relates to discussions of methodologies that should be employed by educational researchers, potential limitations of this approach and how these could be addressed, and implications of this work for practitioners and policy makers.

4.1 Methodology in education research

As noted above, the educational establishment has indicated a strong preference for randomized control trials as a way to establish "scientific validity" of research results. Several authors, including those in the CCSS tradition (Cook 2002, 2003) have chastised educational researchers for appearing to avoid randomized assignment (see also Sherman 2003; Towne et al. 2005). This preference for randomization as a "gold standard" is no doubt well-meaning and based on the general outlines of the CCSS tradition. However, as noted above, when viewed in its entirety, the CCSS literature is quite clear in noting the limits of randomized assignment in field settings, such as schools. More important, as discussed above, it offers alternative designs that promote internal validity, can utilize the type of data schools routinely collect, and can answer the types of questions with which schools may be most often concerned—how the implementation of a new curriculum in their setting, with their teachers and their students, has affected students' achievement.

Thus, this article supplements the growing body of scholarship that suggests that there has been an over-emphasis on randomized control trials in educational research, as well as other field settings. One element of this literature is empirical in nature and has compared the results obtained with randomized trials and quasi-experimental designs. Some have found systematic differences between these types of studies (e.g. Agaodino and Dynarski 2004; Glazerman et al. 2003; Weisburd et al. 2001), while others suggest that the results are minimal when appropriate statistical controls are employed (Heinsman and Shadish 1996; Shadish et al. 2008; Slavin 2008).[8] Some scholars have also focused on theoretical issues regarding research designs and, especially, those that are appropriate given the "social structure of instruction" and ways in which instruction is nested within classrooms and schools, resulting in multiple influences on student outcomes (Raudenbush 2008; Sloane 2008a,b).

An alternative approach that has appeared within the critical literature, echoing Shadish et al.'s (2002) call for a "phased model" of work, is programmatic research, series of studies that address an issue, such as the impact of a curriculum. Scholars urging this approach stress that programs of research should use a variety of methodological approaches to develop a fuller and complete picture related to a research question (Odom et al. 2005; Robinson 2004). Sloane (2008a,b) argued that a programmatic approach more closely parallels the model used in drug efficacy research, which begins with basic research and small feasibility studies and proceeds to small-scale efficacy tests often using randomized trials. These are followed by larger scale experimental studies to test effectiveness in broader populations and implementation, scaling, and sustainability research with much larger groups, often using quasi-experimental designs. All of these researchers stress that randomized control trials are more appropriate at some stages of the research program than others.

Interestingly, research studies of Direct Instruction curriculum, including RM, illustrate this progression. In a recent review of the literature Couglin (forthcoming) has found that randomized control trials were most common in the 1970s through the early 1990s. After that time, as would be suggested by Sloane's description, the literature moved to larger implementation trials, studies of how the curriculum could be used in whole systems (scaling studies) and how it could be sustained over time (sustainability research, the last element in Sloane's description). I suggest that the analytic model described in this article could be seen as reflecting the later stages in Sloane's description of research phases, especially

[8] The reviews suggesting differences between the techniques examined studies on criminal justice, welfare and job training programs, and dropout preventions. Those suggesting more minimal effects examined issues related to student achievement and learning, arguably the substantive area that would be of more concern to educational researchers.

implementation trials, large-scale quasi-experiments "to determine the effectiveness of the … intervention under real world conditions" (Sloane 2008a, p. 628).

4.2 Limitations and potential modifications

There are of course, several potential limitations to the approach described in this document. First, assessment data are often reported as the percentage of students meeting certain benchmarks. Compared to analyses with scale scores, these percentages have less variability and are more apt to reflect ceiling and floor effects. To avoid this problem, schools could use, whenever possible, the actual scale scores. This could also counteract the issue of states setting unusually high or low benchmarks, such as a "Lake Woebegone effect," with cut-offs that result in the vast majority of students being "above average."

Second, the designs could be criticized for not controlling for the threat to internal validity that comes from regression toward the mean. This threat is most likely to occur when a sample represents groups that are at extremes of the distribution (higher or lower than the overall mean) and chance variation produces scores closer to the overall average. The use of the historical control group in the CCG-H design and the use of several years of data for baseline and follow-up measures can help control for this possibility. In addition, analysts could directly address this concern by selecting a sub-sample of comparable schools from a larger set of comparative data. For instance, using data sets often made available by state departments of education, they could choose other schools with similar levels of baseline achievement as a comparison group.

Third, these designs ignore the issue of treatment fidelity and the extent to which the programs have been implemented as designed. The analyses outlined in this article describe the effect of programs as they were implemented within a given setting. Variation in fidelity from the optimum would logically affect both mean levels of achievement as well as variations across classrooms and/or schools. Thus, one way to examine variations in fidelity could be to directly study variations in outcomes across classrooms and schools and potential factors that might be contributing to these differences.

Fourth, even though the approach described in this article helps to control for reactivity that might be related to random assignment within a school, it does not remove the possibility of other sources of this phenomenon. It is not unreasonable to expect that the ways in which teachers taught and students learned in one year would influence the ways that they taught and learned in subsequent years, even if a new curriculum were implemented. This is one reason that experts in implementation suggest that a substantial amount of time may be needed before new programs can be fully established, or stabilized, within a school (Engelmann and Engelmann 2004). One way in which researchers might handle this issue with the current designs is to wait to examine results until a program has been in place for several years.

Fifth, the use of state-level data as a control in the CCG-H design, while cheap and efficient, is, of course, far from precise, for it is possible that other schools in a state were also using the curriculum under study. Fortunately, however, this possibility would lead to smaller differences between the two groups, thus providing a conservative test of the curriculum's impact.

Finally, it is possible that the set of data analyzed in this document could be highly selective, for the publisher might have chosen only to report data from successful sites. As a result, the reported effect sizes might over-state the impact of the curriculum. Interestingly, however, the effect sizes found in my analysis are similar to those reported in meta-analyses of other studies of Direct Instruction programs. For instance, Hattie's recent meta-analysis

of meta-analyses incorporated results from 304 studies of DI programs, with 597 effects and over 42,000 students and concluded that the average effect size associated with DI was 0.59 (2009, pp. 206–207). This value is slightly higher than the values reported above. Importantly, as with the other studies, our analysis found no difference in the effects across a variety of settings.

4.3 Implications for practitioners and policy makers

The approach described in this article can be seen as a way to empower school practitioners. They could easily employ the techniques described above. The calculations are easy to complete, and on-line calculators can usually be used to compute the statistics. In addition, the approach is cheap and efficient. The calculations use readily available data, which are, for the most part, in the public domain and/or already gathered in response to government regulations. Perhaps most important, because the analyses involve locally gathered data, school officials can be confident that they apply to their setting. In short, the analyses described in this report can empower local school officials—giving them the ability to determine, with relative accuracy and confidence, the extent to which curricular interventions produced desired results within their local setting.

The discussion above could also be used to support calls for those who control educational research agendas, both research funding and the reviews of research findings, to move away from a single-minded focus on randomized control trials. While this focus is justified as a way to guarantee "scientific" research, the CCSS literature is clear in its suggestion that other types of designs are often more appropriate in field settings such as schools and that it is much more important to examine results from a series of investigations in a range of settings rather than to seek out only a very few studies that meet limited and strict criteria.[9]

It is perhaps reasonable to expect that at least some educational researchers will resist the suggestions of this report, for conducting randomized control trials has become big business within that sector, employing dozens of researchers and their assistants, providing support for large research institutes, and elevating scientific reputations of the "experts" vis-à-vis practitioners. In the long run, however, one could suggest that students—those whom we most want to benefit—would be best served by using assessment and analysis procedures that are cost-effective and also provide the greatest internal and external validity. To do anything less could be termed educational, as well as scientific, malpractice.

Acknowledgements The author thanks Douglas Carnine, Kurt Engelmann, Zig Engelmann, Dan Johnston, and Robert O'Brien for helpful comments on earlier drafts. Any errors are the sole responsibility of the author.

Appendix

A.1 Statistical calculations

This appendix describes the statistical calculations used to assess data in the CCG and the CCG-H designs. The data presented are from the Alaskan school discussed in the second major section of this article.

[9] The What Works Clearinghouse is typical of this approach, often examining dozens of studies of curricula but finding that only a handful, almost always fewer than five, are worthy of review (Slavin 2008).

A.2 Analyzing the CCG design

As described in the text, the question of interest with the data in Table 1 is the extent to which spring DIBELS scores of the 2007 cohort differed from those of the 2003 cohort. The cohorts can be considered statistically independent and thus a simple difference of proportions test can be used (If mean scores were available then a difference of means test would be appropriate). The hypotheses to be tested are:

$$H_O: p_{2003} = p_{2007} \text{ versus}$$
$$H_1: p_{2003} < p_{2007}.$$

That is, we wish to test the hypothesis that the proportion of students meeting the benchmarks (in the case of being at low risk) is greater in 2007 than in 2003.

Such tests can be easily accomplished with on-line calculators (Traditionalists could, of course, use simple pencil and paper, calculator, or excel-aided computations; however, the on-line calculators provide important safeguards against computational errors). For the results in this article I used the calculator available at Dimension Research, Inc. (http://www.dimensionresearch.com/resources/resources_overview.html, accessed 24 Oct 2011) and all results obtained were equivalent, within rounding error, to those calculated by hand. To provide more exact estimates of the associated probabilities I used a different on-line calculator (http://faculty.vassar.edu/lowry/tabs.html#t, accessed 8 Dec 2011). To adjust for variations in sample size I used the usual formula for comparisons [$df = (n_1 + n_2) - 2$].

I used Cohen's d as a measure of effect size, defined as the difference between the proportions or means divided by the common standard deviation. In contrast, the z or t scores used to test hypotheses about the differences between these values are calculated by dividing the difference by an estimate of the standard error (The standard error is inversely related to the sample size, and is equal to the standard deviation divided by a function of n). To minimize computational error, I again used an on-line calculator, choosing to employ one that converts the z-scores to effect sizes (http://www.uccs.edu/~faculty/lbecker/, accessed 8 Dec 2011). Again, the results obtained were equivalent, within sampling error, to those obtained with pencil and paper calculations.

Finally, I used these data to compute uni-dimensional confidence intervals, given in the last column of Table 1. These confidence intervals tell the minimum value of the effect size, at a given level of confidence, which I set at 0.95, following conventions in the field. Using the logic of confidence intervals, if the sample value of the effect size is used as an estimate of the population value,

$$P[D > d - (1.645)\,(\text{s.e.})] = 0.95 \tag{1}$$

where d is the sample value of the effect size, D is the population value of the effect size, and s.e. refers to the standard error. The standard error can easily be calculated from the values of d and z as in

$$\text{s.e.} = d/z \tag{2}$$

Thus, for the proportions of students at risk,

$$\text{s.e.} = d/z = 0.62/2.77 = 0.22, \text{ and} \tag{3}$$
$$P[D > (0.62 - (1.645 \times 0.22)] = P[D > 0.25] = 0.95. \tag{4}$$

That is, we can be 95% confident that the effect size in a hypothetical larger population is greater than 0.25. Note also that to maintain comparability, all effect sizes and minimal values

of the confidence intervals were coded so that positive values indicate that the students with the RM program had higher achievement and negative values indicate lower achievement.

A.3 Analyzing the CCG-H design

The basic logic of the CCG-H is summarized in Table 3 in the text. In this design the differences between cohorts within a school are compared with differences between cohorts in a larger group to which the cohorts belong. In the examples used in this report, because the dependent measure was third graders' scores on the state assessment tests, the larger comparison group was third grade students in the state. Table 4 summarizes the data used in the analysis, showing the percentage of third graders in the school and in the state who had scores on the reading and writing state assessments that reached or surpassed the "proficient" level in 2003 and in 2007. The school implemented RM after the 2003 testing and thus the change from 2003 to 2007 demonstrates the difference that occurred after that implementation. During that period schools throughout the state were trying to improve achievement, and the increase in the state scores illustrates that change. To get an accurate picture of the change in the school that might be attributed to RM it is important to control for these more general changes in the state. The analysis involves three steps: (1) Comparing the school values to those for the state using simple standard deviation or z-scores; (2) Using these scores to calculate effect sizes that describe the change over time in the school, controlling for change within the state; and (3) Calculating t-ratios that can be used to compute the statistical significance of the change and confidence intervals regarding its magnitude. Each of these steps is described below.

A.3.1 Comparing school and state values using standard scores

Standard scores, or z-scores, are used to compare the scores of a sample with a population in the first step of the analysis. The formula for a z-score is

$$Z = (M_i - \mu) / \sigma \tag{5}$$

where M_i is the sample mean, μ is the population mean and σ is the population standard deviation. The resulting value tells the magnitude of the difference between a sample and a population in standard deviation terms. Thus, a z value of 1.0 indicates a difference of one standard deviation; a value of 0.50 indicates a difference of one-half of a standard deviation, etc.

The fourth column of Table 4 reports z-scores comparing the results for the school to those for the state as a whole for each year. Because we wish to compare the values for the school (a sample) to the state as a whole (the population), we need to compute the standard deviation for the state. For proportions (or percentages) this can be done with the binomial distribution. Translating the percentages to proportions (by simply dividing by 100),

$$\sigma = \sqrt{p_u \times q_u} \tag{6}$$

where $p_u =$ the proportion in the population and $q_u = (1 - p_u)$.

As an example, consider the computations for 2003. For this year, $p_u = 0.739$, $q_u = 0.261$.

$$\text{Thus } \sigma = \text{ sq root } [0.739 \times 0.261] = \text{sqrt } [0.193] = 0.439 \tag{7}$$

Substituting in the formula (5) for standard (z) scores, where 0.745 is the sample value M (or p_s) (for the district) and 0.739 is μ (or p_u), the value for the population,

$$z = (0.745 - 0.739)/0.439 = 0.006/0.439 = 0.014. \quad (8)$$

This value (rounded to two significant digits) is given in the first line of Table 4 of the text. It indicates that in 2003 the percentage of third graders in the school who scored at the proficient level or above on the reading test was 0.014 of a standard deviation greater than in the state.

Similar calculations were completed for 2007 and for scores on the writing test, and they are also reported in Table 4. They indicate that by 2007 the percentage of the school's third graders scoring at or above the proficient level in reading was 0.15 of a standard deviation greater than in the state as a whole. For writing, there were no differences between the school and the state in 2003, but by 2007 the percentage for the school was 0.29 of a standard deviation higher than in the state.

A.3.2 Effect sizes for changes over time in a district controlling for changes in the state

Educational researchers often use Cohen's d, a measure of effect size, to describe the magnitude of an effect. Cohen's d is calculated as

$$(M_1 - M_2)/\text{s.d.} \quad (9)$$

where M_i = mean of a group and s.d. = the common standard deviation.

The resulting value tells the magnitude of the difference between two groups in standard deviation terms. A value of 1.0 indicates that the means differ by an entire standard deviation; a value of 0.50 indicates that they differ by one-half of a standard deviation. Note that this interpretation of an effect size is precisely the same as interpretations of z-scores—a difference in standard deviation terms. Thus, one can see z-scores as effect sizes; they are equivalent to Cohen's use of standard deviation units as an effect size for differences between means, but involve comparing a sample mean to a population mean.

While the results given in Table 4 provide a snapshot of achievement in each year relative to the state, the question of greater interest is the extent to which changes occurred over time in the school. To be most accurate we also need to control for changes that occurred within the state as a whole. Specifically, we want to know the extent to which a school's or district's performance changed over time relative to the performance of the state. A simple way to describe these changes is to compare the z-scores from one year to another. In other words, we can simply calculate the change in the standard deviation scores. Again, we build on the standard formula for Cohen's d, where

$$d = (M_1 - M_2)/\text{s.d} \quad (10)$$

Because the standard deviation for z-scores is, by definition, 1.0,

$$d_z = (Z_1 - Z_2) \quad (11)$$

Thus, the effect size that describes the change in a district relative to changes in the state can be calculated simply by comparing the z-scores using Eq. 11.

Again the data in Table 4 can illustrate. For reading scores in 2003, the z-score comparing the proportion of third graders scoring at the proficient level or higher in the school with the proportion in the state was 0.014, indicating that the proportion for the school was 0.014 standard deviations above the proportion for the state as a whole. In 2007 the z score for

this comparison was +.147, indicating that the proportion was 0.147 of a standard deviation higher than the score for the state.

The difference of these scores can be easily calculated

$$d_z = 0.147 - (0.014) = 0.134 \tag{12}$$

From 2003 to 2007, the proportion of third graders scoring at the proficient level or higher increased by 0.13 of a standard deviation relative to changes in the scores of third graders throughout the state. The comparable result for writing scores was

$$d_z = 0.289 - 0.004 = 0.285 \tag{13}$$

One could, of course, simply look at the change in scores over time and compute an effect size with this information (using the formula $d = (M_{t1} - M_{t2})/\text{s.d.}$). This is the type of calculation used with the CCG design and described above. However, as described in the text, one potential problem with this design is that it fails to control for possible historical changes. Comparing the changes in the district with those in the state as a whole, as occurs with formula (11), provides a control for such historical changes.

The data in Table A-1 illustrate the importance of including such a comparison. The data report the results of a simulation involving changes in achievement over time. We assume that the proportion of students meeting criteria in an imaginary district changed from 0.30 in year 1 to 0.70 in year 10. If only these data were considered, the effect size reflecting the change would be 0.87[$= (0.70 - 0.30)/0.46 = 0.40/0.46$]. This indicates a large change, one that would be considered educationally significant.

The data within the body of Table A-1 provide information on different possible changes within the state over the ten year time period. For instance, the first line reports the situation where 0.50 of the students in the state met criteria in Year 1 and the same proportion met criteria in Year 10. The third and fourth columns report the population standard deviation for years 1 and 10 (using Eq. 6 above); and the fifth and sixth columns give the z-scores comparing district values with the state values for each year, using the formula in Eq. 8 above. For Year 1, the district proportion of 0.3 compared to a state proportion of 0.5 results in a z-value of −0.40; for year 10, the district proportion of 0.7 compared to a state proportion still at 0.5 results in a z-value of +0.40. In other words, in year 1 the district proportion was 0.4 of a standard deviation below the state value, but in year 10 the value was 0.4 of a standard deviation above the state value. Using formula (11) above, the effect size for the

Table A-1 Example of calculating Cohen's effect size, d, with and without controlling for changes in the state

State year 1	State year 10	SD year 1	SD year 10	z year 1	z year 10	d_z	Bias
0.5	0.5	0.50	0.50	−0.40	0.40	0.80	0.07
0.4	0.6	0.49	0.49	−0.20	0.20	0.41	0.46
0.3	0.7	0.46	0.46	0.00	0.00	0.00	0.87
0.2	0.8	0.40	0.40	0.25	−0.25	−0.50	1.37
0.6	0.4	0.49	0.49	−0.61	0.61	1.22	−0.35
0.7	0.3	0.46	0.46	−0.87	0.87	1.75	−0.88
0.8	0.2	0.40	0.40	−1.25	1.25	2.50	−1.63

It is assumed that the values for the district were $p = 0.30$ for year 1 and $p = 0.70$ for year 10. Thus the effect size, d, without controlling for any changes within the state, would be 0.87

district change, while controlling for the state values $d_z = 0.40 - (-0.40) = 0.80$. This value is quite close to the unadjusted value of 0.87, which would be expected given the lack of change in the state (See the last column of Table A-1, which reports the difference of d_z and the value of d without controlling for changes in the state; that is, the bias from not controlling for changes in the state).

The second, third and fourth lines of data depict a situation where there were positive changes within the state. The second line shows a change from 0.40 to 0.60 meeting criteria, slightly less than in the district; the third line has a change that exactly matches that within the district, from 0.30 to 0.70; and the fourth line depicts a situation with a larger change, 0.20 to 0.80. In other words, these are situations where the district improved over time, but so did all students within the state. In one situation the change was slightly greater in the district than in the state (line 2); in the next situation (line 3) the change was exactly the same as in the state; and in the third situation (line 4) the change in the district was actually less than in the state. Clearly the effect size of 0.87 does not accurately portray what really happened in the district *relative to what was happening in the state as a whole*. However, the effect size calculated with Eq. 11 (in the next to last column of the table) accurately reflects these changes. The d_z in line 2, with district changes that are slightly greater than in the state, is 0.41. This is still positive, but smaller than with the unadjusted value. The d_z in line three with changes equal to the state is, as one would expect, equal to zero. Finally, the d_z in line four, where the state had greater positive changes than the district, is negative. This is appropriate because the changes in the district, although positive, were less than in the state as a whole. Relative to the state as a whole the students in our hypothetical district lost ground over time.

The final three lines in Table A-1 depict a situation where the students in the state became less likely to meet criteria over time. In these situations the effect size that only considers district data ($d = 0.87$) underestimates the actual magnitude of change. For instance, the fifth line of data simulates a change in the state from 0.60 meeting criteria in Year 1 to 0.40 in year 10. The effect size when these state level changes are considered is 1.22, substantially larger than the value of 0.87 calculated without this control.

Similar results occur with other patterns of change. The essential point is that, if one wants to examine the amount of change relative to some type of larger comparison group, data for the comparison group need to be considered. Examining changes in the z-scores relative to this larger population provides a simple, accurate way to describe these changes.

A.3.3 Inferential tests to examine changes over time in a district controlling for changes in the state

While the effect size computed above (d_z) provides a descriptive measure of the magnitude of change, practitioners, policy makers, and researchers are often interested in whether or not the changes might have occurred by chance. To answer this question researchers use simple hypothesis tests. For the data presented in this article one can use t-tests. As mentioned above, the data used in the CCG designs, comparing cohorts of students from one year to another, involve samples that may be seen as independent. The vast majority of children in a grade in one year would not be in that grade in a subsequent year.

The null hypothesis tested is that there was no change in a district's results, relative to the state, over time. In other words, the null hypothesis is that the z-scores in year a equal the z-scores in year b. If a curriculum had no effect we would expect that the z-scores relative to the state would be the same in both years. This is our null hypothesis: the difference between the two z-scores is zero.

$$H_o : z_b - z_a = 0 \tag{14}$$

Alternatively, if there were a positive effect, we would expect there to be fewer students at risk or with a deficit and more students at low risk in the later years (This is our alternative hypothesis).

$$H_1 : Z_b - Z_a > 0 \tag{15}$$

To test this hypothesis we can do a simple comparison of means test using the t-distribution, treating the z-scores for each year as the means and using the standard formula for a t-test

$$t = (M_2 - M_1) / s.e._{2-1} \tag{16}$$

The standard error is simply a function of the standard deviation and the sample size.

$$s.e. = \sqrt{[(s.d.1/n1) - (s.d.2/n2)].} \tag{17}$$

By definition, the z-scores have standard deviations of 1.0.

$$\text{Thus, } s.e. = \sqrt{[(1/n1) + (1/n2)].} \tag{18}$$

To illustrate these calculations, consider the data for 2003 and 2007 from the Alaskan school reported in Table 4 in the text. For 2003, n = 41 and the z-score comparing the value to the state = .014; for 2007, n = 63, and the z-score comparing the value to the state = +0.147.

$$s.e. = \sqrt{(1/41) + (1/63)} = 0.184. \tag{19}$$

$$t = (Z_2 - Z_1) / s.e. = (0.147 - 0.014) / 0.184 = 0.726. \tag{20}$$

The degrees of freedom associated with this test are

$$df = n_1 + n_2 - 2 = 41 + 63 - 2 = 102. \tag{21}$$

Using a standard t-table (see an on-line calculator at http://faculty.vassar.edu/lowry/tabs.html#t), it can be found that the probability of getting a t-value of 0.726 by chance with samples of this size is 0.2348 (one-tail).

The corresponding values for writing are

$$t = (Z_2 - Z_1) / s.e. = d/s.e = 0.285/0.184 = 1.55 \tag{22}$$

and the associated probability is 0.0621.

These values can also be used to place confidence intervals around the effect size. Because schools would logically be interested only in interventions that enhanced student achievement, one tail hypothesis tests, as well as unidirectional confidence intervals, would be most appropriate. For such a confidence interval one would assume that the mean of the sampling distribution equals a sample value (the effect sizes calculated above) and then calculate the value above which 95% of the sample values would fall. In other words, based on information from this analysis, what is the minimal effect size that we could expect with 95% confidence. In a typical normal curve this would correspond to values at 1.645 standard errors below the mean. This corresponds to the following formula:

$$P[D > (d - (1.645)(s.e.))] = 0.95 \tag{23}$$

where D is the estimate of the effect size within the population.

For the Alaska site, the application of these formulas leads to

$$\begin{aligned} P[D > (0.134 - (1.645)(0.184))] &= P[D > (0.134 - 0.303)] \\ &= P[D > -0.169] = 0.95 \text{ for reading and} \\ P[D > (0.254 - (1.645)(0.184))] &= P[D > (0.285 - 0.303)] \\ &= P[D > -0.018] = 0.95 \text{ for writing.} \end{aligned}$$

Other values could, of course be used for more narrow or wide confidence intervals.

Appendix B

B.1 Statistical analyses of case studies

This appendix summarizes the results of the separate analyses used in the third section of this article, which illustrates the grounded theory of generalized causal inference. The first section gives the results for analyses that used the CCG design, and the second gives results for analyses that used the CCG-H design.

B.2 Analyses using the CCG design

Table 5 in the text describes the sites included in the SRA report. Data for seven sites were deemed appropriate for analysis with the CCG design. They represent all parts of the country, schools with varying grade levels and focus on students at different grades. Table B-1 summarizes the socio-demographic characteristics of the schools, and this also indicates substantial variation. While five of the seven sites had a majority of Caucasian students, Hispanic students comprised 90% of one school and Native American students comprised 100% of another. Data from these sites were available for 21 separate analyses.

Table B-2 summarizes the information obtained in the individual analyses (Details on each separate analysis are available on request from the author). The first panel of the table gives statistics for each of the 21 analyses, and the bottom panel aggregates the information across each of the seven sites. As would be expected, the range of effect sizes was larger with the group of individual comparisons (0.08 to 1.66) than with the values averaged to the site

Table B-1 Socio-demographic characteristics of schools in CCG comparisons

Report #	Free or reduced lunch	African American	Caucasian	Hispanic	Asian-American	Native Amer./Al. Native	ELL
1	64	0	75	0	0	25	0
3	95	4	3	90	1	2	52
8	47	1	97	1	0	1	1
14	42	1	97	1	1	0	0
15	100	0	0	0	0	100	0
21	47	2	91	6	1	0	5
23	30	4	89	2	5	0	0

All numbers are percentage of students listed in the report as having a given characteristic

Table B-2 Summary of results with analyses using CCG designs, 7 sites and 21 comparisons

Site #	Grade	Year began RM	Years DI at school	Effect size	z	Prob.	Min. for CI
1	K	K	4	0.62	2.77	0.006	0.25
1	K	K	4	0.48	2.17	0.002	0.12
1	K	K	4	0.87	4.15	<.0001	0.53
3	3	1	3	0.85	7.64	<.0001	0.67
8	2	K	3	0.58	3.8	0.0001	0.33
8	3	1	3	0.08	0.5	0.31	−0.18
8	4	2	3	0.08	0.49	0.31	−0.19
8	5	3	3	0.23	1.49	0.07	−0.02
8	6	4	3	0.18	1.15	0.13	−0.08
14	3	K	5	0.71	2.94	0.0023	0.31
14	3	K	5	0.81	3.91	<.0001	0.47
14	3	K	5	1.66	7.24	<.0001	1.28
14	3	K	5	0.66	4.23	<.0001	0.40
14	3	K	5	0.71	3.19	0.001	0.34
15	K	K	2	1.22	3.21	0.002	0.59
15	1	K	2	1.43	3.86	0.003	0.82
15	2	1	2	0.30	0.53	0.3	−0.63
15	3	2	2	0.64	1.52	0.07	−0.05
21	3–8—SPED		2	1.14	20.83	<.0001	1.05
23	3	K	7	0.16	1.04	0.15	−0.09
23	3	K	6	0.35	2.40	0.01	0.11
Min.				0.08			−0.63
Max.				1.66			1.28
Aver.				0.66			0.29

Site #	Average effect size	Aver. Min. for CI
Site-Level Summaries		
1	0.66	0.30
3	0.85	0.67
8	0.23	−0.03
14	0.91	0.56
15	0.90	0.18
21	1.14	1.05
23	0.25	0.01
Min.	0.23	−0.03
Max.	1.14	1.05
Aver.	0.71	0.39

The three analyses for site 1 are those reported in the text regarding the three different DIBELS measures. The five analyses for site 14 reflect analyses for different schools. Other cases with multiple analyses for a site reflect the different grades examined as listed in the second column of the table

Table B-3 Socio-demographic characteristics of schools in CCG-H comparisons

Report #	Free or reduced lunch	African American	Caucasian	Hispanic	Asian-American	Native Amer./Al. Native	ELL
1	64	0	75	0	0	25	0
2	72	55	42	3	0	0	0
5	92	4	47	49	0	0	44
6	85	92	2	6	0	0	0
7	76	26	32	34	7	1	32
10	21	3	73	17	1	6	4
11	91	94	1	3	1	1	3
13	92	61	38	1	0	0	0
17	51	23	70	6	0	0	0
18	100	38	47	12	0	0	0
19	72	0	49	50	0	0	58
23	30	4	89	2	5	0	0
24	44	8	73	15	3	1	0
min	21	0	1	0	0	0	0
max	100	94	89	50	7	25	58
mean	68.5	31.4	49.1	15.2	1.3	2.6	10.8

All numbers are percentage of students listed in the report as having a given characteristic

Table B-4 Summary of results with analyses using CCG-H designs, 13 sites and 40 comparisons

Site #	Grade and comparison	Year began RM	Years DI at school	Effect size	z	Prob.	Min. for CI
1	3—reading	K	4	0.13	0.73	0.24	−0.17
1	3—writing	K	4	0.28	1.55	0.06	−0.02
2	3	2	2	0.11	0.65	0.26	−0.17
2 (a)	4	3	2	−0.26	−1.35	0.91	−0.58
2	5	4	2	0.21	1.26	0.10	−0.06
2	6	5	2	0.09	0.52	0.30	−0.2
5	3—proficient +	2	2	0.63	2.08	0.02	0.13
5	4—proficient +	3	2	0	−0.02	0.51	−0.5
5	5—proficient +	4	2	0.35	1.17	0.13	−0.14
5(b)	3—unsatisfactory	2	2	1.04	3.45	0.001	0.55
5(b)	4—unsatisfactory	3	2	0.59	1.94	0.03	0.09
5(b)	5—unsatisfactory	4	2	0.27	0.9	0.19	−0.23
6	3	K	7	0.79	2.36	0.01	0.24
7	3–5, AA to AA	k	4	0.29	1.99	0.02	0.05
7	3–5, Hisp. to Hisp.	k	4	0.2	1.55	0.06	−0.01
7	3–5, Low Inc. to LI	k	4	0.18	1.99	0.02	0.03
7	3–5, Excp. to Excp.	k	4	0.27	1.46	0.07	−0.03
7	3–5, Cauc. to Cauc.	k	4	0.23	1.74	0.04	0.01
7	3–5, total to total	k	4	0.09	1.25	0.11	−0.03
7	3–5, AA to Total	k	4	0.34	2.33	0.01	0.1
7	3–5, Hisp. to Total	k	4	0.28	2.2	0.01	0.07
7	3–5, LI to Total	k	4	0.22	2.36	0.01	0.07
7	3–5, Excp. to Total	k	4	0.23	1.25	0.11	−0.07
7	3–5, Cauc. to Total	k	4	0.18	1.39	0.08	−0.03
10	4	K	9	0.19	2.18	0.01	0.05

Table B-4 continued

Site #	Grade and comparison	Year began RM	Years DI at school	Effect size	z	Prob.	Min. for CI
11	4	K	7+	0.22	1.37	0.09	−0.04
13	4	3	2	1.04	4.40	<.0001	0.65
17	3—sch. A	1	3	0.53	5.57	<.0001	0.37
17	4—sch. A	2	3	0.35	3.71	0.0001	0.19
17	5—sch. A	3	3	0.22	2.36	0.01	0.07
17	3—sch. B	1	3	0.36	2.06	0.02	0.07
17	4—sch. B	2	3	0.33	1.86	0.03	0.04
17	5—sch. B	3	3	0.01	0.05	0.67	−0.28
18 (a)	4	k	8	0.63	3.46	0.0005	0.33
18	4 (alternate years)	k	7	0.94	5.17	<.0001	0.64
19	3—Total to total	1	3	0.65	5.72	<.0001	0.47
19	3—Hisp. To Hisp.	1	3	0.80	4.96	<.0001	0.53
19	3—Hisp. To Total	1	3	0.81	5.03	<.0001	0.55
23	4	K	7	0.03	0.20	0.42	−0.12
24	4	K	7	0.56	2.87	0.0025	0.24
Min.				−0.26			−0.58
Max.				1.04			0.65
Aver.				0.36			0.07

Site #	Average effect size	Aver. Min. for CI
Site-Level Summaries		
1	0.21	−0.09
2	0.04	−0.25
5	0.48	−0.02
6	0.79	0.24
7	0.23	0.01
10	0.19	0.05
11	0.22	−0.04
13	1.04	0.65
17	0.30	0.08
18	0.79	0.49
19	0.75	0.52
23	0.03	−0.12
24	0.56	0.24
Min.	0.03	−0.25
Max.	1.04	0.65
Aver.	0.43	0.13

level (0.23 to 1.14). The average effect size was, however, slightly larger with the analysis with sites as the unit of analysis (0.66 vs. 0.71). On average, the minimum value for the 95% (unidimensional) confidence interval was 0.29 for the set of individual comparisons and 0.39 for the site level analysis.

B.3 CCG-H

Table 5 in the text also summarizes characteristics of the sites that had data appropriate for analysis with the CCG-H design. There were 13 sites and 40 separate comparisons. As with the CCG analyses, these sites represent schools and districts from around the country and schools with varying grades. Most of the analyses involve students in grades three and four, with a few also including students in grades 5 and 6. This no doubt reflects the grades at which state assessments are first routinely given. Table B-3 summarizes the socio-demographic characteristics of the students. There was substantial variation across the sites in racial-ethnic composition, in the percentage of students eligible for free or reduced lunch, and in the percentage of ELL students.

Table B-4 summarizes the results. The first panel of the table gives statistics for each of the 40 separate analyses, and the bottom panel aggregates the information across the thirteen sites. As would be expected, the range of effect sizes is larger with the group of individual comparisons (−0.26 to 1.04) than with the values averaged to the site level (0.03 to 1.04). As with the analyses with the CCG designs, the average effect size is slightly larger with the analysis with sites as the unit of analysis (0.43 vs. 0.36). On average, the minimum value for the 95% (unidimensional) confidence interval was 0.07 for the set of individual comparisons and 0.13 for the site level analysis.

References

Agaodino, R., Dynarski, M.: Are experiments the only option? A look at dropout prevention programs. Rev. Econ. Stat. **86**, 180–194 (2004)

Boruch, R.: Better evaluation for evidence-based policy: place randomized trials in education, criminology, welfare, and health. Ann. AAPS **599**(May), 6–18 (2005)

Brunswick, E.: Perception and the Representative Design of Psychological Experiments, 2nd edn. University of California Press, Berkeley (1956)

Campbell, D.T.: Relabeling internal and external validity for applied social scientists. In: Trochim, W.M.K. (ed.) Advances in Quasi-Experimental Design and Analysis, pp. 67–77. Jossey-Bass, San Francisco (1986)

Campbell, D.T., Stanley, J.C.: Experimental and Quasi-Experimental Designs for Research. Rand McNally, Chicago (1963)

Cartwright, N.: Are RCTs the gold standard? BioSocieties **2**, 11–20 (2007)

Clay, R.A.: More than one way to measure. Monit. Psychol. **41**(September), 52–55 (2010)

Cook, T.D: The generalization of causal connections: multiple theories in search of clear practice. In: Sechrest, L., Perrin, E., Bunker, J. (eds.) Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data (DHHS Publication No. PHS 90-3454)., pp. 9–31. Department of Health and Human Services, Rockville (1990)

Cook, T.D: Clarifying the warrant for generalized causal inferences in quasi-experimentation. In: McLaughlin, M.W., Phillips, D.C. (eds.) Evaluation and education: At Quarter-Century, pp. 115–144. National Society for the Study of Education, Chicago (1991)

Cook, T.D.: Randomized experiments in educational policy research: a critical examination of the reasons the educational evaluation community has offered for not doing them. Educ. Eval. Policy Anal. **24**, 175–199 (2002)

Cook, T.D.: Why have educational evaluators chosen not to do randomized experiments. Ann. AAPS **589**, 114–149 (2003)

Cook, T.D.: Emergent principles for the design, implementation, and analysis of cluster-based experiments in social science. Ann. AAPS **599**(May), 176–198 (2005)

Cook, T.D., Campbell, D.T.: Quasi-Experimentation: Design and Analysis Issues for Field Settings. Rand McNally, Chicago (1979)

Cook, T.D., Steiner, P.M.: Case matching and the reduction of selection bias in quasi-experiments: the relative importance of pretest measures of outcome, of unreliable measurement, and of mode of data analysis. Psychol. Methods **15**, 56–58 (2010)

Cook, T.D., Scriven, M., Coryn, C.L.S., Evergreen, S.D.H.: Contemporary thinking about causation in evaluation: a dialogue with Tom Cook and Michael Scriven. Am. J. Eval. **31**, 105–117 (2010)
Collins, M., Carnine, D.: Evaluating the field test revision process by comparing two versions of a reasoning skills CAI program. J. Learn. Disabil. **21**, 375–379 (1988)
Couglin, C.: (forthcoming) A review of the direct instruction literature: a four decade program of research. In: John, L. (ed.). Direct Instruction and Evidence-Based Practice. ADI Press, Eugene
Cronbach, L.J.: Designing Evaluations of Educational and Social Programs. Jossey-Bass, San Francisco (1982)
Cronbach, L.J., Meehl, P.E.: Construct validity in psychological tests. Psychol. Bull. **52**, 281–302 (1955)
Engelmann, S.: Teaching Needy Kids in Our Backward System: 42 Years of Trying. ADI Press, Eugene (2007)
Engelmann, Z.: Socrates on gold standard experiments. http://www.zigsite.com/PDFs/socrates5mm2.pdf (2009). Accessed 8 Dec 2011
Engelmann, S.E., Carnine, D.: Theory of Instruction: Principles and Applications. Irvington Publishers, New York (1982)
Engelmann, S.E., Engelmann, K.E.: Impediments to scaling up effective comprehensive school reform models. In: Glennan, T.K. Jr.., Bodilly, S.J., Galegher, J.R., Kerr, K.A. (eds.) Expanding the Reach of Education Reforms: Perspectives from Leaders in the Scale-up of Educational Interventions, pp. 107–133. Rand, Santa Monica (2004)
Fayer, H.: Place randomized trials: experimental tests of public policy. Ann. AAPS **599**(May), 272–291 (2005)
Fuchs, L.S., Fuchs, D., Hosp, M.K., Jenkins, J.R.: Oral reading fluency as an indicator of reading competence: a theoretical, empirical, and historical analysis. Sci. Stud. Read. **5**(3), 239–256 (2001)
Glazerman, S., Levy, D.M., Myers, D.: Nonexperimental versus experimental estimates of earnings impacts. Ann. AAPSS **589**, 63–93 (2003)
Good, R.H., Simmons, D.C., Kame'enui, E.J.: The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. Sci. Stud. Read. **5**(3), 257–288 (2001)
Hattie, J. (2009). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. London and New York: Routledge.
Heinsman, D.T., Shadish, W.R.: Assignment methods in experimentation: when do nonrandomized experiments approximate answers from randomized experiments?. Psychol. Methods **1**, 154–169 (1996)
Huitt, W.G., Monetti, D.M., Hummel, J.H.: Direct approach to instruction. In: Reigeluth, C., Carr-Chellman, A. (eds.) Instructional-Design Theories and Models: Volume III, Building a Common Knowledge Base, pp. 73–98. Lawrence Erlbaum, Mahwah (2009)
Julnes, G., Rog, D.J.: Current federal policies and controversies over methodology in evaluation. N. Dir. Eval. **113**(spring), 1–12 (2007)
Maxwell, S.E.: Introduction to the special section on Campbell's and Rubin's conceptualizations of causality. Psychol. Methods **15**, 1–2 (2010)
McMillan, J.H.: Randomized field trials and internal validity: not so fast my friend. Pract. Assess. Res. Eval. **12**(15), http://pareonline.net/pdf/v12n15.pdf (2007). Accessed 8 Dec 2011
Millenson, M.: Demanding Medical Excellence: Doctors and Accountability in the Information Age. University of Chicago Press, Chicago (1997)
Miron, G.: The constructive use of existing data and research for evaluating charter schools. Paper prepared for the Symposium on the Use of School-Level Data for Evaluating Federal Education Programs, December 8–9, 2005, Washington (2005)
Morgan, S.L., Winship, C.: Counterfactuals and Causal Inference: Methods and Principles for Social Research. Cambridge University Press, New York (2007)
National Mathematics Advisory Panel (NMAP): *Foundations for Success: The Final Report of the National Mathematics Advisory Panel*. (US Department of Education, Washington 2008)
Noble, J.H. Jr.: Meta-analysis: methods, strengths, weaknesses, and political uses. J. Lab. Clin. Med. **147**, 7–20 (2006)
Odom, S.L., Brantlinger, E., Gersten, R., Horner, R.H., Thompson, B., Harris, K.R.: Research in special education: scientific methods and evidence-based practices. Except. Child. **71**, 137–148 (2005)
Raudenbush, S.W.: Advancing educational policy by advancing research on instruction. Am. Educ. Res. J. **45**, 206–230 (2008)
Robinson, D.H.: Scientific research is programmatic. In: Scientifically-based Education Research and Federal Funding Agencies: The Case of the No Child Left Behind Legislation, pp 121–128. Information Age Publishing, Charlotte (2004)
Rubin, D.B.: Reflections stimulated by the comments of Shadish (2010) and West and Thoemmes (2010). Psychol. Methods **15**, 38–46 (2010)
Sampson, R.J.: Gold standard myths: observations on the experimental turn in quantitative criminology. J. Quant. Criminol. **26**, 489–500 (2006)

Schwandt, T.A.: A diagnostic reading of scientifically based research for education. Educ. Theory **55**, 285–305 (2005)
Scriven, M.: The logic of causal investigations. Unpublished paper, Western Michigan University (n.d.)
Scriven, M.: Can we infer causation from cross-sectional data? National Academy of Sciences. http://www7.nationalacademies.org/bota/School-Level%20Data_Michael%20Scriven-Paper.pdf. (2005). Accessed 8 Dec 2011
Scriven, M.: A summative evaluation of RCT methodology: and an alternative approach to causal research. J. MultiDiscipl. Eval. **5**, 11–24 (2008)
Shadish, W.R.: Campbell and Rubin: a primer and comparison of their approaches to causal inference in field settings. Psychol. Methods **15**, 3–17 (2010)
Shadish, W.R., Cook, T.D.: The renaissance of field experimentation in evaluating interventions. Ann. Rev. Psychol. **60**, 607–629 (2009)
Shadish, W.R., Cook, T.D., Campbell, D.T.: Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin, Boston (2002)
Shadish, W.R., Clark, M.H., Steiner, P.M.: Can randomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. J. Am. Stat. Assoc. **103**, 1334–1343 (2008)
Sherman, L.W.: Misleading evidence and evidence-led policy: making social science more experimental. Ann. AAPSS **589**(September), 6–19 (2003)
Slavin, R.E.: What works? Issues in synthesizing educational program evaluations. Educ. Res. **37**, 5–14 (2008)
Sloane, F.: Randomized trials in mathematics education: recalibrating the proposed high watermark. Educ. Res. **37**, 624–630 (2008a)
Sloane, F.: Through the looking glass: experiments, quasi-experiments, and the medical model. Educ. Res. **37**, 41–46 (2008b)
St. Pierre, E.A.: Scientifically based research in education: epistemology and Ethics. Adult Educ. Quart. **56**, 239–266 (2006)
Towne, L., Wise, L.L., Winters, T.M.: Advancing Scientific Research in Education. National Academies Press, Washington (2005)
U. S. Department of Education: Improving Teacher Quality State Grants, ESEA Title II, Part A, Non-Regulatory Guidance. US Department of Education, Washington, October 5 (2006) http://www2.ed.gov/programs/teacherqual/guidance.pdf. Accessed 8 Dec 2011
Weisburd, D., Lum, C.M., Petrosino, A.: Does research design affect study outcomes in criminal justice?. Ann. AAPSS **578**, 50–70 (2001)
West, S.G., Thoemmes, F.: Campbell's and Rubin's perspectives on causal inference. Psychol. Methods **15**, 18–37 (2010)
What Works Clearinghouse: Procedures and Standards Handbook (Version 2.0). http://ies.ed.gov/ncee/wwc/pdf/wwc_procedures_v2_standards_handbook.pdf (2008). Accessed 8 Dec 2011
Whitehurst, G.J.: The Institute of Education Sciences: new wine, new bottles. Paper presented at the 2003 Annual Meeting of the American Educational Research Association, April 22 (2003)
Williams, B.A.: Perils of Evidence-Based Medicine. Perspect. Biol. Med. **53**, 106–120 (2010)
Winship, C., Morgan, S.L.: The estimation of causal effects from observational data. Ann. Rev. Sociol. **25**, 659–706 (1999)
Yin, R.K., Davis, D.: Adding new dimensions to case study evaluations: the case of evaluating comprehensive reforms. N. Dir. Eval. **10**, 75–93 (2007)
Zdep, S.M., Irvine, S.H.: A reverse Hawthorne effect in educational evaluation. J. School Psychol. **8**, 89–95 (1970)