Conclusion

Class Size Effects: Assessing the Evidence, its Policy Implications, and Future Research Agenda

David Grissmer

RAND

Where is there a school system having a sum of money with which to investigate and perfect a scheme experimentally, before putting it into general operation? And can we expect continuous and intelligent progress in school matters until the community adopts a method of procedure which is now a commonplace with every great industrial undertaking? Is not the existing method of introducing reforms into education a relic of an empirical cut-and-try method which has been abandoned in all other great organizations? And is not the failure to provide funds so that experts may work out projects in advance a penny-wise and pound foolish performance?

—John Dewey (1913)

Introduction

There is much new evidence in this issue regarding the cost and effect of class size reduction (CSR), and much remains to be learned. In the summary that follows, I incorporate this new evidence into the existing literature and suggest a future research agenda for providing a stronger empirical base for informing policy. First, I summarize the current evidence for the effect of CSRs from experimental and non-experimental measurements, address questions about the robustness of each type of estimate, and suggest hypotheses that could reconcile the conflicting evidence. Next, I discuss the evidence from inside the classroom about the processes that might account for achievement effects from smaller class sizes and outline a tentative theory directed toward explaining the effects from CSRs. I then discuss the cost of reductions and the implications the empirical evidence holds for policy formulation. Finally, I suggest a future research agenda for reconciling the experimental and non-experimental evidence and for supporting an expanded theory explaining why CSR produces achievement gains.

This article is organized around the following questions:

1. Results from experimental/quasi-experimental studies
   - What are the results from the recent experimental/quasi-experimental studies?
   - How robust are the results to deviations from ideal experimental design and execution?

2. Results from non-experimental studies
   - What are the results from non-experimental studies?
   - How robust are these results to assumptions inherent in the model specifications?

3. Reconciling the experimental and non-experimental evidence
   - What does experimental data tell us about the validity of model specifications?
   - What changes inside the classroom in large and small classes?
   - How might such changes inside the classroom affect short- and long-term student learning and development?

4. How much do CSRs cost?

5. What are the policy implications of the current results?

6. What are the implications for future research?
Results From Recent Experimental/Quasi-Experimental Studies

This section will summarize the results of the Tennessee STAR experiment and the Wisconsin SAGE quasi-experiment. The section draws from the results presented in this issue by Finn and Achilles; Nye, Hedges, and Konstantopoulos; and Molnar et al.; as well as results found by Nye, Hedges, and Konstantopoulos (1999) and Kneuer (1999).

The Tennessee STAR and Wisconsin SAGE Designs

The multi-district Tennessee STAR experiment randomly assigned a single cohort of kindergarten students in 79 participating schools to three treatment groups: large classes (approximate mean of 22-24 students) with or without an aide and small classes (approximate mean of 15-16 students). Those students entering at kindergarten were scheduled to maintain their treatment through first, second, and third grade. However, the treatment groups shifted in significant ways after kindergarten due to attrition and late-entering students.

Students newly entering a participating school in first, second, or third grade were also randomly assigned to treatment groups, but these late-entering students came from schools outside the sample and were in smaller classes in earlier grades. Many students entering the experiment between kindergarten and third grade moved away from a participating school after 1 or more years of participation, so students could have spent 1 to 4 years in small classes in different combinations of grades. For instance, the combination of late entry and attrition left the third grade small class sample with only one-half who had been in small classes all 4 years. The remaining part of the sample had spent 1 to 3 years in small classes. The experimental sample changed in another important way after kindergarten since late-entering students had lower average scores than beginning students, and those leaving the original sample had lower scores than those who remained all 4 years. The third grade small class sample then continued a higher scoring group who remained all 4 years, and a lower scoring group with fewer and later years in small classes.

The sample of participating students in any grade was over 6,000 students, but late entries and exits meant that about 12,000 students were included over the 4 years. The characteristics of the students were different than those of average Tennessee students. The experimental sample contained approximately 33% minority students and over 50% to 60% of all students were eligible for free or reduced-price lunch, compared to 23% minority students and about 45% free or reduced-price lunch students for Tennessee in 1986. The sample was also quite different from students nationwide in the U.S., where approximately 30% were minority students and 37% were eligible for free and reduced-price lunch in 1990.

The Wisconsin SAGE study differed in several important ways from the Tennessee STAR experiment. In the SAGE study, only schools with very high proportions of free-lunch students were eligible for inclusion. Assignments were not randomized within schools, but rather a pre-selected control group of students from different schools was matched as a group to the students in treatment schools. The treatment is more accurately characterized as pupil-teacher ratio reduction since a significant number of schools chose two teachers in a large class rather than one teacher in a small class. The size of the reduction in pupil-teacher ratio was slightly larger than CSR in Tennessee.

There were about 1,600 students in the small pupil-teacher treatment group in the Wisconsin study, compared to approximately 2,000 students in small classes in the Tennessee study. However, the size of control groups differed markedly, around 1,300 students in Wisconsin and around 4,000 in Tennessee, if both regular and regular with aide classes are combined. The SAGE sample had approximately 50% minority students with almost 70% eligible for free or reduced-price lunch.

Finally, there were other initiatives besides smaller pupil-teacher ratio initiated in Wisconsin that may have had some limited impact. A final important feature is that tests were given at the beginning and end of the first grade in SAGE, rather than at the end of consecutive years, as in Project STAR. Since achievement changes differently for advantaged and disadvantaged students over the summer, a beginning-year test for school effect is probably a better control than a previous-year test (Alexander & Entwisle, 1998; Entwisle & Alexander, 1992).

How Large Were the Effects?

The current results from experimental/quasi-experimental studies show statistically significant effects from large CSRs in early grades in all subjects tested from kindergarten through eighth grade. However, the size of the effects is surprisingly hard to pin down because it is dependent on student characteristics, the length of time and which grades were spent in small classes, the test and units of measure used to measure the effect, and whether the focus is short-term or long-term effects.

Finn and Achilles (this issue) restore the measured achievement differences for the Tennessee experiment between those in large and small classes at each K-3 grade in each subject tested. The results show how CSR had statistically significant effects in each K-3 grade and all subjects tested, showing achievement raised by 0.15-0.25 standard deviations. The size of the reported effects increased markedly from kindergarten to first grade, but remained fairly constant or even declined in Grades 2-3. Effects for math and reading were 0.15 and 0.16 in kindergarten, respectively, and 0.27 and 0.24 in first grade.

These basic results are also estimated by Kneuer (1999) and Nye et al. (1999) using models that control for school effects and teacher and classroom covariates. The results of these analyses show substantial agreement both in the magnitude of the effect by grade and its pattern. Thus, the use of more sophisticated models including covariates and school and classroom effects does not seem to substantially alter the reported effects.

The effects from the Wisconsin study for two consecutive first-grade classes also show statistically significant effects on achievement in all subjects. The effect sizes in the first grade are in the range of 0.1-0.3 standard deviations. The lower estimates between 0.1-0.2 occur in regression estimates, while the raw effects and hierarchical linear modeling (HLM) estimates are in the 0.2-0.3 range. While the estimates seem consistent with the Tennessee study at first grade, more analysis is needed before the results can be compared.

The patterns in the Tennessee study of increasing effects from kindergarten to first grade, followed by stable or even declining effects, raises the question whether the resources invested in each year after first grade were justified. However, the consistency across grades may be partly as artifact of the standard deviation measure. The original effect sizes were reported in standard deviation units assuming a uniform standard deviation across grades. Finn and Achilles (this issue) use a different measure-grade-equivalents—that shows an expanding gap by grade. While the grade equivalent measure cannot be strictly interpreted as the additional months of schooling required to close the gap, this measure suggests that the amount of resources needed to close the gap does expand across grades.

However, regardless of the measure used, the reported experimental effects should not be compared across grades since the small class sample past kindergarten changed in each year, containing children who had spent different numbers of years in small classes and had different characteristics. The reason for the stable or declining effect may be that students in later grades had spent fewer average years in small classes. So answering the question about the size of effects requires determining whether effect size changes for different types of students and for students having more years in small classes. But first we address the most important question for policy.

Do the Effects Persist?

All students were in large classes in Grades 4-7, but Finn and Achilles (this issue) report statistically significant effects between 0.10 and 0.16 standard deviations in each subject through seventh grade. Nye et al. (this issue) also report effects for fourth, sixth, and eighth grade using two-level HLM including student and school characteristics that shows statistically significant effects in three subjects for each grade. Their estimates are between 0.11 and 0.20 standard deviations. Since the effects at third grade tend to be in the range of 0.15-0.20, the long-term effects show little sign of declining from the time students left smaller classes. Finn and Achilles also show that the achievement gap between large and small classes expands through seventh grade if the grade equivalent measure is utilized. The gap in kindergarten is less than 1 month, while the gap in seventh grade is 4-8 months, depending on subject. So, using this measure shows an expanding gap.

Do the Effects Grow With More Years in Small Classes?

The reported results using the entire sample are increasingly hard to interpret past kindergarten since any changes in the reported effects across grades might be due to either different average duration in small classes or different student characteristics. So the key question is whether achievement rises with more years in small classes. The three measurements that address this question have some inconsistency between them.

232

233
Nye et al. (1999) analyze the short-term effects through third grade with HLM comparing results for four groups: students in small classes in any grade, two or more grades, three or more grades, and four grades. Each estimate is made using the remaining part of the sample as the contrasting group. Their results show effects for those in small classes in all 4 years versus the remaining sample to be 0.35–0.39 standard deviations. Corresponding estimates for the group with 3 or more years, 2 or more years, and at least 1 year in small classes are 0.26–0.27, 0.19–0.24 and 0.15–0.16, respectively. All of these results are statistically significant.

Nye et al. (this issue) do similar estimates for long-term effects at fourth, sixth, and eighth grades. Their results show continuing statistically significant effects in three subjects and all grades, with increasing effects for longer time spent in small classes. The estimated effect sizes are around 0.30–0.41 for those in small classes all 4 years, 0.22–0.29 for 3 or more years in small classes, 0.16–0.24 for 2 or more years in small classes, and 0.06–0.20 for 1 or more years in small classes. These appear to be no differences between subjects, nor significant trends from fourth through eighth grade. So the pattern and magnitude of effect size appear to remain fairly stable from third to eighth grade for each subject.

Krueger (1999) made separate short-term estimates of the effects by year of entry and grade. These estimates essentially contrast groups having different durations and student characteristics. His results for the group entering in kindergarten and staying all 4 years show a gain in kindergarten, but show no gains in subsequent years for each additional year in small classes. However, for groups entering in first and second grade, there is evidence of significant gains from each additional year, with somewhat smaller first year gains.

Krueger (1999) also estimates a pooled model that assumes equal first year gains regardless of year of entry, and equal incremental gains for each additional year, regardless of year of entry or duration. His estimates show statistically significant gains for both first year and incremental effects, but larger first year gains with only small incremental gains. For instance, estimates made from his equations predict a cumulative effect from 4 years in small classes of 0.19–0.25 standard deviations with approximately two-thirds of the effect in the first year.

Krueger's (1999) results and those of Nye et al. (1999) on the short-term effects of more years in smaller classes seem inconsistent. Krueger's predicted results for 4 years in small classes at third grade (0.19–0.25) are significantly smaller than Nye et al.'s (0.35–0.39). Nye et al.'s estimates also underscore why group effects would be for students in small classes all 4 years. However, unlike Krueger, the results in Nye et al. do not imply that most of the short-term effect happens in the first 2 years. In fact, the long-term effects measured by Nye et al. imply that the sustained effects only occur if students are in small classes for 3 or 4 years. More years in small classes is most important for sustaining long-term effects. This explanation would be consistent with previous early intervention research that shows short duration interventions produce significant short-term effects that usually fade over time, but long duration interventions produce more long-term effects (Barnett, 1995; Ramey & Ramey, 1985).

A common interpretation of the results of the Tennessee experiment is that effect sizes are around 0.2 standard deviations for 4 years in small classes. The results on duration would suggest that this common interpretation of the Tennessee results may substantially underestimate the sustained class size effect for those in small classes all 4 years where long-term effects are approximately 0.4 standard deviations. From a policy perspective, the differences in magnitude and the difference in duration effects are critical since significant funding is associated with CSR for each grade, and the cost-effectiveness of CSR would change significantly.

Are the Effects Larger for Minority and Disadvantaged Students?

Finn and Achilles (this issue) report minority effects that are approximately double those for White students in Grades K–3. Krueger (1999) also reports larger effects for minority students. Estimated short-term effects for minority students from 4 years in small classes would be 0.3 standard deviations or larger with effects for remaining students closer to 0.2 standard deviations or greater. Mohan et al. (this issue) also show that in Wisconsin minority students have significantly larger effects than non-minority students. Current analysis also indicates that free-lunch students have larger short-term effects than their counterparts. Estimates from Krueger for free-lunch and non-free-lunch students are 0.27 and 0.21, respectively. No separate long-term estimates have been made for minority or free-lunch students. No separate long or short-term estimates have been made for advantaged White students. It is not clear whether a sufficient sample exists to do this or whether such students could not be identified from the data.

Possible Bias Due to Deviations From Ideal Experimental Conditions

We now turn to questions concerning whether deviations from ideal experimental design and execution could affect the present results. Well-designed and well-executed experimental studies are those in which data are accurately analyzed and interpreted and in which we can come to causal evidence in social science. The basic premise of experimentation—choosing two groups of subjects through randomization or pre-selection such that the only difference between them is the variable of interest—remains the ideal method of building social science knowledge. While experiments are potentially capable of providing the most compelling evidence, they often fall far short of achieving this objective (Bucholz, 1994; Heckman & Smith, 1995; Manis, 1996).

An experimental design may fail to make the treatment large enough to dominate the inevitable sources of "noise" in the experiment. Thus a choice in this issue is between either an effect from 21 rather than from 24 to 16–17 would most likely not have produced compelling results. The flaws in execution inevitable in experiments with actual people can be severe enough to discredit the results. These can arise from differential attrition from test and control samples, failure to maintain control over students, exclusion of students, leakage between test and control groups, and the presence of Hawthorne type effects. Another problem is that experimental results can depend on specific contextual factors such as types of students, teachers, grades included, and curriculum, so that results might not be generalizable.

Hambrook (this issue) summarizes potential issues of similar treatment effect at attention when analyzing and interpreting the Tennessee results. He includes differences between random and actual classroom assignment, non-random test-taking and attrition, lack of randomization of schools, potential problems in teacher assignments, lack of knowledge of prior class sizes for students entering after kindergarten, and Hawthorne effects. The impact of most of these deviations has been tested, but some either have not or cannot be tested directly. However, even if the effects are directly tested and found not to introduce bias into the measured effects, their presence alone will introduce additional variance not directly measured by an analysis. So the standard errors will be underestimated.

Not all students initially assigned to small or large classes remained in these classes. Krueger (1999) and Nye et al. (1999, this issue) have performed analyses comparing results using the actual and assigned treatments. No significant differences in effects are discovered in these studies. The issue of differential attrition in test and control groups is also addressed in these studies. All studies, including Nye et al. (this issue), find significant differences between students staying in the experiment (continuers) and those who either enter after kindergarten (late entries) or leave (leavers), with continuers having higher scores than leavers and late entries. However, none of the studies find significant differences in the characteristics of entering and exiting students between small and large class groups. Its only differential effects that would bias effects.

These studies have not addressed the issue of non-test taking for students present in the classes. About 88%–116% of children in classes did not take tests due to absence or other reasons. Hannushek (this issue) reports that 14% of the first-grade sample excluded is in test and control groups for each grade, so the characteristics of the students excluded would have to vary widely in small and large classes to have any effect on the results. This certainly needs to be analyzed further.

Krueger (1999) analyzes available teacher characteristics (experience, education, and non-evidence of non-random teacher assignment. While other teacher characteristics (teacher test scores and verbal ability) may be more closely related to teacher quality, it would be highly unlikely that randomization did not distribute these characteristics nearly equally between small and large classes (Lentz & Brewer, 1995, 1993; Ferguson, 1991, 1989). School size is not significantly related and finds no evidence of non-random teacher assignment. While other teacher characteristics (teacher test scores and verbal ability) may be more closely related to teacher quality, it would be highly unlikely that randomization did not distribute these characteristics nearly equally between small and large classes (Lentz & Brewer, 1995, 1993; Ferguson, 1991, 1989). School size is not significantly related and finds no evidence of non-random teacher assignment.
representative of Tennessee schools, and were even less representative of students in the nation.

A second concern for the generalizability of the results may be the variance of the results across schools. Hauhshek (this issue) presents evidence of a distribution of class size effects across schools in kindergarten that suggests that small class achievement exceeded both large treatment groups in less than one-half of the schools. This may indicate the presence of intensive effects below the school level that are important. However, Nye et al. (1999) utilize a three-level HLM that suggests the effects were fairly uniform across schools.

Hauhshek (this issue) suggests that if variance in the results is due to intensive effects within schools, teachers may be more effective at utilizing the additional time small classes provide. Another possible source of variance is of the composition of the classes. Kuegler (1999) provides evidence that having more classmates who attended kindergarten and were not eligible for free lunch affected achievement in the Tennessee study. Discovering the extent of variance across classes and what might explain the variance is important since it could lead to enhancing the effects and would help inform policy about the context where results are larger.

For students entering the experiment after kindergarten, the composition of the class size in their previous schools. Since most migrating students would be from Tennessee, presumably most would have been in large classes in previous grades. The randomization process would in all likelihood account for any differences in previous class sizes for late-entering students. Most analyses of the difference from experimental design and implementation is desirable. It is possible for further analysis to find a flaw in the experiment that significantly affects the results, but extensive analysis to date has eliminated most of the potential problems.

Results From Non-Experimental Studies

Hauhshek (1995, this issue) presents a review of the non-experimental literature specifically directed toward measurements of class size or pupil-teacher ratio that shows little support for significant and positive, class size effects. He presents two other arguments that seem consistent with this evidence. The first argument is that long-term trends in national achievement from 1950-1990 have been flat, although sizable national reductions in class size and/or pupil-teacher ratio occurred that were equal to or larger in size than those in Tennessee. The second argument comes from international studies that he interprets as showing that countries with larger class size do as well or better in scores than with smaller class size.

Hauhshek (this issue) reviews the non-experimental studies from over 250 measurements from studies of effects of class size and/or pupil-teacher ratio. He provides separate results by elementary and secondary school and for "value-added" models within and across states. The results show many studies with statistically significant positive and negative effects, but approximately equal numbers of positive and negative results. It is not clear whether a meta-analysis taking the sample sizes into account would shift the results significantly. 6

National Assessment of Educational Progress (NAEP) scores for 17-year-olds from 1970-1990 do not show large gains even though pupil-teacher ratio declined significantly. However, there is a hypothesis consistent with significant class size effects that may explain this data. Long-term NAEP scores for 17-year-old students are flat only for average and higher scoring White students. Black 17-year-old students made gains of 0.6-0.8 standard deviations, Hispanic students made gains in the 0.2-0.3 range and non-Hispanic White students made gains in the 0.2-0.5 range. However, for White students, the gains are smaller than those for Black students. It is possible that the gains for White students are due to school effects rather than changes in class size. However, even if the gains for White students are due to school effects, it is still possible that the gains are due to changes in class size. It is possible that the gains for White students are due to school effects rather than changes in class size. However, even if the gains for White students are due to school effects rather than changes in class size, it is still possible that the gains for White students are due to school effects rather than changes in class size.

It occurs for subgroups of studies selected by level of schooling (elementary vs. secondary) or particular methodology (product function using previous year's score for controls). However, since some non-experimental studies produce statistically significant positive results, it is possible that there is a set of non-experimental studies with certain characteristics that agree with the experimental data. 6

The conflict may arise simply from the assumption that weighting of non-experimental studies is a valid exercise. Weighing studies assumes that each study represents an independent observation and each measurement accurately measures a class size effect with some standard error for the sample involved. The non-experimental measurements are not independent for two reasons. Many come from different specifications involving the same data set. But more importantly, virtually all of our analyses of non-experimental data rest on a few, often fragile assumptions that, if proven false, can distort large parts of the literature. These commonalities make assumptions inherent in most model specifications make non-experimental studies across different data sets non-independent.

Moreover, different specifications and assumptions made across measurements are not equally accurate. So equal weighting inevitably involves combining inaccurate and inaccurate measurements. Weighting of measurements simply avoids the critical problem of discovering why the results differ among the non-experimental studies and what the most accurate specifications and assumptions are. It is possible that identifying the best set of assumptions and specifications could produce a subset of studies that are more consistent with experimental data.

If all measurements were independent and accurate within the estimated standard errors, then meta-analysis would be able to arrive at a better estimate. However, the high degree of dependency among studies and the combining of measurements that have different specifications and assumptions—when there are significant differences in the quality of specifications and assumptions—analytically meta-analysis problematic. It also avoids the problem of determining which specifications and assumptions are accurate.

There are several reasons why experimental and non-experimental results can differ, and why non-experimental results differ with each other. One possibility is that class size effects vary with characteristics that are not accounted for in the experimental studies. These characteristics may be educationally related or population-related.

There are many reasons why experimental and non-experimental results can differ, and why non-experimental results differ with each other. One possibility is that class size effects vary with characteristics that are not accounted for in the experimental studies. These characteristics may be educationally related or population-related.
ables (Ferguson & Ladd, 1986). Only non-experimental measurements with similar grades, similar class size ranges, and similar teacher characteristics should be grouped. Hanushek (this issue) moves in this direction by separating the measurements by elementary and secondary schools with either group showing consistent positive effects. But it may be that further isolating studies involving earlier grades would show closer agreement.

Flaws in model specifications and assumptions inherent in analyses with non-experimental data can also explain differences. Models having different specifications or making different assumptions are essentially incomparable and should be independently grouped since only one specification and set of assumptions can be accurate. Since there are probably almost as many model specifications and assumptions as there are measurements, it means that most specifications are flawed.

One common source of model specification differences is missing variables and poor quality data. In education, a particular problem is that family variables reflecting both inputs and environmental factors explain most of the variance, and schooling variables are correlated with these family variables. The effect of any single school factor such as class size is dwarfed by the set of family variables, making it very sensitive to the specific specification and quality of these variables. Rarely have data sets included such variables and rarely have the specification of these variables included the interaction and non-linearities now known to be present in family effects.

Use of previous year’s test scores as a proxy for all earlier experience and missing family variables has been considered to be among the best specifications. But the previous year’s test scores are a direct function of the test used, and the use of previous test scores as a control and shows that these produce no consistent positive effects. However, a previous year’s test score cannot be a poor control because differential declines occur over the summer in achievement (Alexander & Entwisle, 1989). Thus a test score used for control at the beginning of the school year may produce a different result than at the end of the school year for different students. Krueger (1999) has also suggested that the pattern of results in Tennessee cannot be obtained with production function techniques using a previous year’s score as control except in the first year of the intervention since much of the effect occurs in the first year.

Another reason, suggested by the Tennessee data, why a single year of previous scores may be inadequate is that CSR has long-term effects. Grismer and Flanagan (in press) have suggested that multiple year effects imply that the single year of previous scores is insufficient for control of previous school experience. The fact that the substantial effect remains at eighth grade, and that changes occur between fourth and eighth grade for some students, can only be explained by knowing that students in smaller classes during particular years in K–3. The effect in any year is the summation of conditions from several previous years that requires controlling for previous scores for as many years as effects last. Since data has generally not been available from kindergarten at the individual level, few if any previous studies can account for all previous grades. Ironically, aggregate data analysis generally thought to be more subject to bias—might be able to provide previous historical data better than currently available individual-level data, and thus could provide better results (Hanushek, Rivkin, & Taylor, 1996).

While model specification issues are generally thought to be a criterion for judging the quality of non-experimental studies, it is harder to develop criteria that can span experimental and non-experimental studies to equivalent scrutiny. One method might be to compute the robustness factors in classroom-level selection and more experienced teachers may choose smaller classes. In other districts, newer teachers may choose smaller classes in order to ease the entry into teaching. In rural districts, random influences may be at work that make these differences more unanalyzable to analysis.

Despite the potential flaws in non-experimental data, in the long run, policy analysis will largely be dependent on improving non-experimental analysis. Large-scale experiments such as the Tennessee STAR experiment can be costly and take considerable time to plan, implement, and analyze. While more experimentation seems essential to making progress in educational research, experiments can never be depended on to solve all the complex and contextual effects. Educational research will probably never follow health research where trials are needed for every new intervention before implementation.

However, one of the key contributions of experimental data is that it can be used to help "benchmark" specifications and assumptions used in non-experimental data. Accurate model specification should also work on experimental data, replicating the experimental results. Using experiments to guide non-experimental analysis can be a powerful tool in improving analysis with non-experimental data. These improvements can lead to appropriate specifications that can provide more credible results for variables not included in experimental studies.

Besides recomposing the results from experimental and non-experimental data, there is a third path to building research consensus. Confidence will eventually arise primarily from associated analysis that builds a theory of classroom and family processes that explains why class size may affect achievement and predicts the effects.

How Can Smaller Class Sizes Raise Achievement?

Besides replicating experiments and improving specifications and data in non-experimental analysis, a third key element is needed to provide the kind of research consensus desired in all science. The results of either experimental or non-experimental analysis are meant to provide the basis for developing theories of educational processes and student learning that gradually incorporate wider phenomena within their purview. These theories should eventually accurately predict the results of empirical work and be able to make new predictions to guide future classroom research. Theories are necessary when their very nature are more robust than any set of experimental or non-experimental studies since they incorporate results of multiple measurements and incorporate research across levels of aggregation. However, little theory-building has been done in education.

In the case of class size we need a theory of classroom and home behavior of teachers, students, and parents that explains why smaller classes might produce higher achievement in both the short- and long-term. Initially we need to understand what teachers and students are differently in large and small classes, and then determine whether these differences can be related to the size of short-term achievement gains. Perhaps the more difficult area of theory will be to explain gains lost after the end of an intervention. An early intervention either has to change cognitive, psychological, or social development in important ways, or change the future environment (peers, family) that affects the individual. Possibilities range from changes in brain development to learning different ways of interaction with teachers and peers to developing differ-
ent study habits to bring in different poor groups years later.

Four articles in this issue present evidence relating to the processes that may be producing either short- or long-range achievement effects. In both the Tennessee and Wisconsin experiments, data were collected from teachers about their own behavior and the behavior of students. In addition, Betts and Shkolnik (this issue) and Rice (this issue) utilize other data sets to determine whether class size affects the ways teachers spend time in the classroom, assign homework, plus classes, and complete textbook material in middle and high school math and science classrooms.

Finn and Achilles (this issue) provide a literature review and report findings from teacher data in Tennessee’s Project STAR. They stress the finding that teachers report that students exhibit more “on-task” behavior and engagement in learning, not only when in small classes in second grade, but after being returned to large classes in fourth grade. These on-task behavior effects in K-3 may be due to more teacher attention, greater opportunity to participate, or other reasons. But the fourth grade effect is a student-generated behavior internalized earlier that affects later behavior or environment in favorable ways—perhaps generating more teacher attention even within the large classes. It is interesting to note the results show higher achievement for those in smaller K-3 classes even in subjects not taught in K-3 (social science and science). While increased reading and math achievement might account for these spillover gains, different engagement in all learning is a possibility.

Molnar et al.’s (this issue) Wisconsin data suggests teachers in small classes spend less time on discipline and more time on instruction. Teachers report more individualized instruction and greater knowledge of each student’s strength and weakness. Also reported are more hands-on activities, more small group discussion, and more content covered.

Betts and Shkolnik (this issue) and Rice (this issue) have analyzed middle and high school data with the objective of estimating how various teacher behaviors change in large and small classes. They add an interesting dimension because their samples, unlike the experimental samples, contain large numbers of classes across student ability and socioeconomic status. SES data. Betts and Shkolnik use an economic production function framework in modeling teacher time allocation, where teachers maximize achievement under constraints of total time and teach-offs in utilizing the time in different ways. Such an approach provides testable hypotheses about teacher time allocation.

Betts and Shkolnik (this issue) estimate the shifts in teacher time allocation as a function of class size, while controlling for many teacher, student, and classroom characteristics. The results indicate statistically significant effects on a number of time variables. Teachers in smaller classes have more instructional time due to spending less time on discipline and administrative routines, and shift to more individual instruction and less lecture time. No significant main effects for time per new material or for percentage of text covered. However, the size of those effects seems too small to account for significant achievement changes. In fact, class size increases of 10 reduced instructional time by only 25%.

The shift away from lecture and toward individual instruction and increased instructional time in small classes was much larger in lower scoring classes and middle school classes. It is still not clear whether the shifts are large enough to account for significant gains in achievement. However, no reliable class size effects have been measured at the middle and high school level, so similar analysis is needed at Grades K-3.

In a companion study, Shkolnik and Betts (1998) develop models that estimate the dependence of achievement scores on teacher time allocation. Both more lecture time and more individual student time raise average scores, but their model indicates that relative shifts to more lecture time and less individual time would raise average scores. With lower class size, there is much less need for one-on-one individual instruction too much if strict achievement maximization is the goal. Rather, teachers seem to have a combined goal of reducing the variance in scores as well as lifting class averages.

Rice (this issue) uses the National Educational Longitudinal Survey (NELS) to estimate the effects of class size on instructional time, non-instructional time (discipline and administration), teacher planning time, and homework assigned in high school math and science classes. The findings generally show more sensitivity in math classes than in science classes to differences in class size and other classroom and teacher characteristics. In math classes, lower class size had statistically significant and positive effects on individual small-group time, innovative practices, and whole group discussion, but not on homework assigned.

The effects showed much greater sensitivity for class sizes less than 20 than for those greater than 20. This suggests that "thresholds" of class size may exist that allow more choice and teacher discretion in time use. It seems reasonable to assume that larger classes make teachers more aware of the opportunity costs of individualized instruction, and so less individual instruction takes place in these classes. However, similar to Betts and Shkolnik’s findings (this issue), the size of the shifts in time allocation due to CSRs is small in typical high school classrooms.

Finding similar to Betts and Shkolnik, Rice (this issue) finds that small classes have statistically significant and positive effects increasing instructional time and reducing time spent on discipline in math classes with increased sensitivity for class sizes of less than 20. The models also show more disciplinary time and more sensitivity to class size effects in lower scoring classes.

Both Rice (this issue) and Betts and Shkolnik (this issue) show many significant and often larger differences in time usage for different teacher characteristics than for different class sizes. Their results indicate that female teachers tended to spend more time in individual/small-group instruction, while more experienced teachers and those certified in math devoted more time to whole-group instruction. More experienced and better certified teachers and those who can plan more also use less time in discipline. While Betts and Shkolnik and Rice usually have similar findings with respect to the effects of smaller class size, there are some significant differences in the effects of some teacher characteristics. However, this type of analysis can suggest interesting trade-offs between effective teacher characteristics, class size, and grouping of students. It is also clear from these analyses that classroom instruction cannot be fully understood without including time spent outside the classroom by the teacher on planning and by the student and parent on homework. Teacher planning time appears to have positive effects on a number of classroom variables. It is unclear whether more planning time reflects a more dedicated or higher quality teacher, or reflects more planning time provided through less classroom time.

Homework assigned may be a key intervening variable between school and home. Having more assigned homework was significantly related to higher scoring students, advanced courses, and male teachers. Lower scoring classes get less homework assigned and more individual instruction, with those in lower scoring small classes receiving the least homework but the most individual instruction. This may suggest that teachers’ assignment of homework may substitute for individualized instruction. It also suggests the possibility that teachers vary their time usage dependent on whether parents spend time with students on homework. Teachers can provide a fairly small amount of individual instruction time per pupil. The Betts and Shkolnik (this issue) sample gives an average of about 6 minutes per week. Thirty minutes per week provides a 50% increase in individual time. Even if parental time is less effective than teacher individual time, the substitution potential is large. Many parents do far more than 5 minutes per week.

The less time spent on individual instruction in higher ability classes may reflect teachers’ experience that parents will provide this time, allowing more group instruction, more new material covered, and more of the textbook covered in these classes. On the other hand, teachers may assign less homework and may try to provide more individualized instructional time to those students with less able and less available parents who do not give much individual instruction at home. However, other students who receive parental time but are in classes with high proportions of children needing individual time will likely receive less group instruction and homework that will lower their achievement.

Another possible substitution occurring between parent and teacher time is disciplinary time. More time is spent on discipline in lower ability classes, allowing less instructional time. Long-term parental involvement with children probably reduces disciplinary time in classes, allowing more instructional time.

Costs of Class Size Reductions

Obtaining good estimates of class size effects is only the first step needed for policy guidance. The second is to accurately estimate the costs of CSRs and the third is to compare the cost-effectiveness of CSRs to that of other policy initiatives (Levin, 1985). These would include increased resources for teacher professional development, higher teacher salaries, more and better-quality pre-school...
programs, more and better technology, and even increases in parental resources.

Brewer, Knop, Gill, and Reckhardt (this issue) take a first step by providing cost estimates for CSRs in Grades 1–3 nationally under a variety of scenarios and assumptions. They take into account the size of reductions, the effect of targeting reductions to disadvantaged students, and the effect of how rules are specified for reductions. Their cost estimates are for national reductions under various scenarios from 1998 through 2007. Their estimates include the operational costs of teacher salaries and benefits and associated teacher resources. They do not include capital costs associated with providing new class rooms. Overall, their estimates suggest that overall teacher salary differences due to decrease in demand or longevity/education increases. The exclusion of capital costs is mainly due to the unavailability of data regarding staff classroom capacity and exactly how schools would meet the demand for additional classroom space in the short and long run. To these estimates, Brouse et al. (this issue) find that the way in which rules for reductions are specified can make significant differences in costs. The most stringent rules—where every classroom in each grade nationwide must be at or below a given limit—can raise costs by over 55% more than a policy that specifies targets in averages across grades at state levels. This study policy allows some classrooms to be over limits within the state while other classrooms are below the limit. However, it is important to note that average class size would be less under the first set of rules than under the latter set. Since rural, less densely populated areas are more likely to have class sizes already below the limits, such a policy might create inequity between class sizes in more versus less densely populated areas.

Costs are significantly reduced by targeting reductions to students in lower income families as measured by free-lunch participation. Costs in 1998–99 fall from $5 billion to $2 billion for CSRs to 18 when targeted at schools having at least 50% of students eligible for free and reduced-price lunch. Since the results from experimental data consistently show larger short-term effects for minority and low-income students, CSRs will be significantly more cost-effective when targeted at these students.

Policy Implications

Several clear policy implications emerge if we assume that CSRs will occur. The evidence shows significant gains in test scores (Karcher et al., 1998) after three or four years in small classes in the K–3 period. The evidence is less clear and somewhat inconsistent as to whether these gains result primarily in the first and second year in small classes, or whether a significant part occurs in the third and fourth year. However, the evidence on long-term gains—the most important measure—indicates significant gains only for the third and fourth year in small classes in all subjects. So current evidence would support reductions in all grades from K–3 if higher sustained achievement were the objective.

Targeting CSRs toward minority and low-income students substantially reduces the cost and raises the test scores of those students. So reductions should focus on those schools with large proportions of minority and low-income students first. Although short-term gains in achievement are smaller for remaining students, it is not clear how the effects change as we move into schools with higher proportions of high-income students. Long-term NAEP trends would indicate little evidence for gains among more advantaged students. So the return on investing in smaller class sizes for these students is certainly more uncertain and more risky.

In implementing class size policies, more rigid rules imposed at the grade and school level significantly raise costs. Decisions at the margin on class size can probably best be done locally, so local discretion with policies targeted broadly across grades and school district seems more sensible than imposing them by grade and school.

Research currently has little to say about the broader decisions concerning reducing class size versus making alternative investments such as increased public pre-kindergarten, higher teacher salaries, providing better facilities and more resources for teachers, and investing in summer and after-school programs. There is almost certainly greater reliability in the evidence that CSRs will bring achievement gains. However, greater reliability does not mean that it is the best investment.

Justifying CSRs strictly on achievement may considerably underestimate benefits. Primary effects may be to decrease anxiety, costs of future education, future employment, and welfare utilization. These effects also carry direct costs to society that can ultimately be related to the cost of the intervention. One study done of a pre-kindergarten program and randomized early intervention for young toddlers showed significant long-term returns to society from targeted investment in the program close to the above effects (Karcher et al., 1998). Some caution is warranted in moving too far too fast in CSRs for at least two reasons. First, the opportunity costs of CSRs are high because they are expensive and because they are hard to reverse if other investments are later found to be more cost-effective. The irreversibility stems from the visibility and political popularity of CSRs, which make it hard to backtrack. The second reason is that there will presumably be diminishing returns as class sizes are reduced to lower levels. It is not clear at present how quickly returns might decline, so it is possible to make too much reduction.

Future Research

Enhancing the Tennessee and Wisconsin Studies and Future Experimentation

Ritter and Borch (this issue) trace the somewhat tortuous path of the Tennessee experiment from inception, to legislative compromise, to key research decisions, and ultimately through analysis and assessment of its impact. The original demonstration (at no point in the legislative history was the word experiment recorded) arose from strong advocacy of statewide CSRs. Eventually a political compromise was struck between two opposing factions: those who wanted widespread CSRs and those who considered them too expensive with insufficient evidence of their effects. A social scientist turned legislator played a key role in shaping the compromise. The legislative specifications for such a demonstration were sufficiently imprecise that the research teams from several Tennessee universities need not have designed an experiment at all—it was one that needed randomization and/or large CSRs. These decisions critical to its future influence were made by members of the research teams who carried out its largely successful implementation.

Only years after the end of the K–3 portion of the experiment did it begin to receive the recognition it deserved. A review of the experiment by Mosteller (1995) helped to widen the awareness and provide an external, independent assessment of its validity. Ritter and Borch (this issue) cite descriptions of the Tennessee experiment as "one of the great experiments in education in our history" and "the most significant educational research done in the U.S. during the past 25 years."

Much has been learned from the Tennessee study despite fairly minimal levels of resources devoted to the experiment and continuation. Less attention is also fairly mismanaged. Underestimating often results in significant delays in reporting results, inability to fully address analytical issues, and failure to take maximum advantage of the opportunity afforded by experimentation. Additional resources can considerably enhance the value of these data collections and their analysis, even today. There are short-term secondary analyses needed to resolve inconsistencies in effect sizes and the pattern of gains by duration, to estimate gains more precisely for more advanced students, and to compare the emerging Wisconsin results with the results from Tennessee.

More importantly, expanded and additional follow-up data collection should be considered. Since recent research has suggested that raised achievement may not be the primary benefit from early interventions, wider and continued data collection from the Tennessee STAR sample is crucial. The STAR sample is now finishing high school. Delinquency, labor force characteristics, years of education completed, welfare utilization, test performance, and college performance data are all much more important to collect than achievement data. Such follow-up should also collect data that is directed at identifying what is different within individuals or their environment that can explain the sustained gains and/or changed behavior. A clear opportunity will be raised without significant future investment in this sample. In Wisconsin or in future such experiments, it is clearly desirable in early data collection to make assessments of learning processes, social and psychological development, peer relations, and other factors that may later account for sustained gains. We have a very narrow set of measures from the current studies with which to build theories.
Grissmer

More experimentation seems important, but some feasibility studies are needed to identify specific candidates, their potential for successful experimentation, their costs, and their potential benefits. But until more trust is developed in non-experimental methods, experiments will remain a key source of reliable evidence for policymakers. Such experimentation need not be confined to large-scale experiments like Tennessee's Project STAR. Systematic small-scale experimentation directed toward theory building can have substantial payoffs.

There are several key policy areas in education that seem likely candidates for experimentation. Many class size issues remain—especially concerning later grades. But experimentation with hypotheses designed to enhance class size effects appears to be a fertile area. More widely, experimentation appears possible with curriculum, instructional techniques, summer and after-school programs, tutoring, parent involvement, teacher professional development, teacher salary, and incentives and technology. Currently, it may be the case that the limitation on experimentation is the capacity of the research community to support such experimentation.

In designing future experiments, one important objective should be specifically to test assumptions commonly made in non-experimental analyses. In the long run, it is infeasible in non-experimental research that is needed for policy guidance, since there will be only a limited number of experiments possible, and contextual effects will likely be important influences in education. Thus the generalizability of experimental data may always be limited and we will have to depend on non-experimental analysis.

Making Non-Experimental Analysis More Reliable

There are several research directions to improve the reliability of non-experimental analysis. Use of longitudinal data beginning at school entry can sort out many of the specification problems that may exist in previous analyses. There are two new sources of such longitudinal data that will have school, teacher, and family characteristics and achievement data. First, there are newly emerging longitudinal state databases that link student achievement across years. Such data has very large sample sizes and is possible with teacher data and school characteristics.29 This data should be able to sort out many of the potential specification issues involving dependence of later achievement on previous year’s class size and thresholds and interactions with teacher characteristics. It will also likely be possible to determine class size effects for various combinations of large and small classes in early and later grades and the importance of small classes in later grades. The second source will be the Early Childhood Longitudinal Study (ECLS) funded by the U.S. Department of Education that will collect very detailed data on children, their families, and their schools. This data will be much richer in variables, but much smaller in sample size.

A second approach to improving the reliability of non-experimental analysis is to use empirical analysis to test and better understand the assumptions upon which such analysis depends. Why do students in large and small classes have different characteristics? How important are parent and teacher selection processes in determining class size? Do more senior teachers choose smaller classes? Are assumptions more valid in some kinds of schools? Are class sizes in rural areas randomly determined, whereas more selection occurs in cities? Are there strong random elements across states causing the large variance in class sizes across states? There are many empirical approaches to address these kinds of questions that would give us a better idea whether assumptions made in specification are reasonable.

A third approach is to utilize current experimental data and design future experiments specifically to test non-experimental specification assumptions. The Tennessee data is providing important evidence about the dynamic characteristics of class size effects, the differences by student type, and the dependence on duration and possibly other covariates. All this information can be used to determine whether assumptions made in non-experimental analysis are likely to be valid.

Developing Integrated Theories of Classroom and Family Resources

Time on task still appears to be a central organizing concept in learning. A secondary concept involves the productivity and optimal division of that time among the different alternatives: presentation of new material through lecture; supervised and unsupervised practice; periodic repetition and review; and testing.30 Students have a wide variety of ways they spend time in school. Understanding the variance appears to depend on teacher characteristics, characteristics of other students in the class, and the amount of time parents spend at home instructing children. A theory of learning needs to be developed that incorporates school and home time and the various tradeoffs and differences that exist across teachers, classrooms, and SES levels. Such a theory would generate a number of testable hypotheses for research that would then allow better and probably more complex theories to be developed. Such theories then provide guidance as to which research is important to undertake.

Such theory building would mandate linking several disparate and isolated fields of research in education. There is microresearch involving time, repetition, and the like in class learning specific tasks. There is research on teachers in classrooms. There is research on homework and tutoring. There is research on specific reading and math instructional techniques. There is research on class size and teacher characteristics. Theorists can begin to try to understand these disparate areas and suggest theories that can explain the empirical work across these areas. Such linkages seem essential to future progress.

Finally, cognitive development may have patterns similar to development in other areas for children, since brain development seems to be central to each type of development. There is a great deal of research on patterns of physical, emotional, and social development in children from birth through the compulsory school age areas such as differences across children, delays in development, and dependence on previous mastery. Studies involving long-term developmental outcomes—especially for children at risk—are key frontiers that enable development to occur even in highly risky situations. Much can be learned from the literature to help prevent the use of poor modeling assumptions.

Cost Research

Cost research needs to improve in several ways. The most important way is to develop cost-effectiveness measures across interventions. This will require better cost estimates from such interventions. Capital costs for CSER can be estimated with additional research in states and districts that have reduced class sizes. Estimates can be formulated to determine how much modification of existing space, how much new construction, and how many temporary classrooms were utilized, as well as how many teachers were hired in the past. Understanding the reasons why each alternative was utilized will allow better estimates to be calculated nationwide.

Costs also need to include the significant variation in construction and land costs in different states and localities. Urban expansion may be significantly more expensive than rural expansion. Teacher data can be used to obtain better long-term estimates of the incremental costs of hiring additional teachers due to their increasing education and experience over time.

Notes

This research has been supported by the Center for Educational Diversity and Excellence (CREDE), the Department of Education funded Research Center, and the Exxon Education Foundation. This article benefited greatly from comments made by the authors of articles in this issue, by Alan Kreuger and Ron Ferguson, and by Jane Hannum, who initiated this special issue on class size.

1 Earlier small-scale experimentation occurred in the 1950s and 1970s and is summarized in Glass and Smith (1978). This summary was influential in the design of the Tennessee-experiment. Specifically, the summary supported a significant reduction in class sizes that was crucial to successful experimentation. Nye, Hedges, and Kazanas (1999) also show that estimates made from these earlier studies had reductions of the magnitude of class sizes in Tennessee would be near the lower limit of current experimental estimates.

2 The means difference in class sizes changed in each year of the experiment with the last year having a larger gap between large and small classes by almost one student per teacher.

3 Many entering students would be from Tennessee where average class sizes probably were similar to the large-class sample. Some entering from other states could have been in either larger or smaller classes.

4 Many who entered in first grade did not attend kindergarten and those entering and leaving were from families who moved, both of which are associated with lower scores (Kawczak, 1989; Raudenbush, 1995; Clogg in press). Some of the newly entering were also students retained from later grades, which is another factor that would lower scores for entrant students.

5 Even the effects for the group remaining all 4 years may be different in the last year if class sizes were reduced statewide in Tennessee, because students entering in later grades would have been in smaller classes in previous schools under statewide reductions. In the experiment, small classes in Grades 1, 2, and 3 had mixes of combining students who had higher scores in kindergarten and even higher scores from being in smaller classes and migrating students who would have lower baseline scores and no advantage of small classes. This introduces a widening variance in small classes in later grades that would not be present.
with statewide reductions. If more homogeneous classes are easier to teach, then these findings might be expected from statewide reductions.

There are differences in the minority samples as well. The minority sample in the Boston study was primarily Black, whereas the minority sample nationwide had approximately equal proportions of Black students and other minority students.

In the Wisconsin study, the minority percentages were quite different from those in the Tennessee study. Only 25% of the students in the Wisconsin sample were Black, while 7% were Hispanic, 12% were Native American, and 6% were Asian.

Knecht estimates the effects using a variety of models starting with simple ordinary least squares (OLS) without covariates, to including covariates and school fixed effects, to two-stage estimates. I have made adjustments to the results in Knecht's work as well as Finn and Achilles' (this issue) to make them more comparable. I adjusted the effects based on assuming a uniform class size difference of 7.6 students across grades since the average difference in class size between large and small classes changed in each grade. Knecht's data showed differences of 7.6, 7.3, 8.2, and 8.5 from K-2. I have estimated the effect size per pupil each year and multiplied by 7.6 to standardize the estimates. I have also adjusted Finn and Achilles' estimates to eliminate the effect of year-to-year changes in the estimates. I utilized Knecht's estimates of the effect of aides by year to adjust Finn and Achilles' estimates. This results in a decrease in the estimates of years 0.00 to 0.02 standard deviations. While there is substantial agreement among all estimates, Knecht reports somewhat higher than Finn and Achilles' in K-2, but generally lower in third grade with the most significant discrepancies in first grade estimates. However, both show a similar pattern across grades with neither showing significant upward or downward trends. Nye et al. (1999) utilized 2- and 2-stage HLM to estimate effects. Their results generally show a similar magnitude and pattern of effects. Their effects are from 0.15 to 0.30 with generally stable or smaller effects in the national sample.

In the Wisconsin study, unlike Tennessee's Project STAR, a different proportion of children may have attended kindergarten, and the two first grade groups had significantly different kindergarten class sizes. The SES and demographic characteristics of the Tennessee and Wisconsin samples also differed considerably. Furthermore, the model specifications currently differ, with Molner et al. explicitly utilizing beginning-year test scores for controls, whereas their methodology was not used in Tennessee. Knecht (1995) suggests that the use of a previous year's scores in models with the Tennessee data would not pick up the tail effects. Finally, it is not clear whether effects are similar for a single teacher with 14 students versus two teachers with 25 students. Further analysis with more similar specifications and accounting for differences in design and student population is needed.

Knecht (personal communication, May 1999) points out that the sample used as the basis for estimating the effects of duration is not fully experimentally designed. An experimental design would have randomly assigned a portion of the small class sample to begin in kindergarten to large classes in each succeeding grade. Rather, this was done to randomly assign incoming students who had different characteristics to large and small classes.

These assumptions of equal first-year gains regardless of entry year or equal increments regardless of entry year or duration terms at odds with the results Knecht presents earlier where distinctly different first year and incremental patterns are shown.

 Policymakers would like to know the differences between those spending all 4 years in small classes versus those spending all 4 years in large classes. The comparisons in Nye et al. compare those in 4 years in small classes with the remaining sample of students, some of whom had 1-3 years in small classes. About one-third of the comparison group had some time in small classes and their average gain was about 0.20 standard deviations, so estimates of effects could increase by about 0.05 standard deviations when compared to students in large classes all 4 years. This means that comparisons of those with no time in small classes versus those with 4 years in large classes would likely be in the 0.35-0.45 standard deviation range.

A conversation with Jeremy Finn about the differences between long- and short-term effects provided the citations from Ramsey and Ramsey and other research on this topic.

A non-experimental analysis of large CSRAs in 17 of the poorest school districts in Tennessee measured effects from 0.4-0.6 standard deviations (Achilles, Nye, & Zabarrick, 1995; Nye, Achilles, Zabarrick, & Falcon, 1995). While non-experimental, the reductions were started in one year and continued for 4 years. Comparisons were made to scores prior to reductions and to similar districts without CSRAs. These data suggest that effects continue to grow for the most disadvantaged students.

Meta-analyses of studies previously reviewed by Hanushek generally provide somewhat more positive appraisal, although still bearing a fairly wide variance in results (Hedges, Laine, & Greenland, 1992; Hedges & Greenwald, 1996).

A conversation with Ron Ferguson helped to elucidate this point.

Even these studies are likely not comparable because each uses different covariates and different specifications for variables.

Ferguson and Ladd (1992) have an excellent discussion of the assumptions often made in these models leading to a discussion of whether a "gold standard" model is possible (i.e., a model specification and data set that accurately describe the underlying phenomena).

One can make the inclusion of teacher education in both the Boston and Oregon studies that generally show the expected sign of more instructional time and less discipline with more experience. Since education and experience are correlated positively, leaving teacher education out would not affect the coefficients of experience.

Part of the reason for this is the lack of access to the data until after the study's completion.

Many have called for more experimentation in educational research, dating from the 1970s and 1980s. Recent papers addressing the subject include: Hanushek (1994), Boruch (1994), Boruch and Foley (in press), Jenkins and Phillips (1996), and Grissmer and Fanagan (in press). However, as the quote at the beginning of the article suggests, the need for experimentation has been recognized for at least 80 years.

See Sanders and Horn (1994, 1998) and Wright, Horn, and Sanders (1997) for an example of such data and the analytical potential for such data. These studies utilize a very sophisticated set of computational algorithms (mixed model methodology) that can accommodate data very large, linked data sets. However, significant questions exist about the basic model specifications used, especially involving SES variables. Part of the larger teacher effects research may reflect the positive correlation between teacher and student SES characteristics. These model specifications also do not take account of the specification issues raised by the Tennessee experiment.

This approach has been followed in Bates (1997), Betsa, and Shkolnik (this issue), and Shkolnik and Bates (1998).

References


Learning about treatment effects from experiments with random assignment of treatments. The Journal of Human Resources, 3(4), 705–733.


Author

DAVID GRISWOLD is senior management scientist at RAND, 1333 H Street, NW, Suite 800, Washington, DC 20005; e-mail: davis@rand.org. He specializes in issues of teacher supply and demand, class size, and analysis of achievement scores.