11-2002

What Large-Scale, Survey Research Tells Us About Teacher Effects on Student Achievement: Insights From the Prospectus Study of Elementary Schools

Brian Rowan

Richard Correnti

Robert J. Miller

What Large-Scale, Survey Research Tells Us About Teacher Effects on Student Achievement: Insights From the Prospectus Study of Elementary Schools

Abstract
This report is about conceptual and methodological issues that arise when educational researchers use data from large-scale, survey research studies to investigate teacher effects on student achievement. In the report, we illustrate these issues by reporting on a series of analyses we conducted using data from Prospects: The Congressionally Mandated Study of Educational Opportunity. This large-scale, survey research effort gathered a rich store of data on instructional processes and student achievement in a large sample of U.S. elementary schools during the early 1990s as part of the federal government's evaluation of the Title I program. We use data from Prospects to estimate the "overall" size of teacher effects on student achievement and to test some specific hypotheses about why such effects occur. On the basis of these analyses, we draw some substantive conclusions about the magnitude and sources of teacher effects on student achievement and suggest some ways that survey-based research on teaching can be improved.

Disciplines
Curriculum and Instruction | Educational Administration and Supervision | Educational Assessment, Evaluation, and Research | Education Policy

Comments
View on the CPRE website.

What Large-Scale, Survey Research Tells Us About Teacher Effects On Student Achievement: Insights from the *Prospects* Study of Elementary Schools

Brian Rowan, Richard Correnti, and Robert J. Miller

CPRE Research Report Series
RR-051

November 2002

Consortium for Policy Research in Education
University of Pennsylvania
Graduate School of Education

Contents

Biographies

Brian Rowan is a professor of education at the University of Michigan and director of the Study of Instructional Improvement, conducted by the Consortium for Policy Research in Education. His scholarly interest focuses on the organizational analysis of schooling, paying special attention to the ways in which schools organize and manage instruction and affect student learning. Rowan's recent publications appear in W. Hoy and C. Miskel (Eds.), *Theory and Research in Educational Administration* (volume I) and the *Journal of Educational Change*.

Richard Correnti is a doctoral candidate in educational administration and policy at the University of Michigan, Ann Arbor. His research interests include the measurement of instruction, instructional effects on student learning, and program evaluation of educational reform interventions.

Robert J. Miller is a doctoral candidate in educational administration and policy at the University of Michigan, Ann Arbor. His main fields of interest are educational policy, organizational theory, and analysis of school effectiveness.

Acknowledgments and Authors' Note

This report is based on research conducted by CPRE, and was supported by grants from the Atlantic Philanthropies; the National Science Foundation's Interagency Educational Research Initiative (Grant No. REC 9979863); and the National Institute on Educational Governance, Finance, Policymaking, and Management; Office of Educational Research and Improvement; U.S. Department of Education (OERI Grant No. R308A960003). Opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the National Institute on Educational Governance, Finance, Policymaking, and Management; the Office of Educational Research and Improvement; the U.S. Department of Education; the Atlantic Philanthropies; the National Science Foundation; CPRE; or its institutional members. The authors thank Steve Raudenbush for advice and assistance at various stages of the work.

Ordering Information

Copies of this report are available for $5.00 each. Prices include book-rate postage and handling. Make checks payable to **Trustees of the University of Pennsylvania**. Sorry, we cannot accept returns, credit card orders, or purchase orders. Sales tax is not applicable. To obtain copies, write:

CPRE Publications
Graduate School of Education
University of Pennsylvania
3440 Market Street, Suite 560
Philadelphia, PA 19104-3325

Quantity discounts are available. For more information, please call (215) 573-0700.

Introduction

This report is about conceptual and methodological issues that arise when educational researchers use data from large-scale, survey research studies to investigate teacher effects on student achievement. In the report, we illustrate these issues by reporting on a series of analyses we conducted using data from *Prospects: The Congressionally Mandated Study of Educational Opportunity*. This large-scale, survey research effort gathered a rich store of data on instructional processes and student achievement in a large sample of U.S. elementary schools during the early 1990s as part of the federal government's evaluation of the Title I program. We use data from *Prospects* to estimate the "overall" size of teacher effects on student achievement and to test some specific hypotheses about why such effects occur. On the basis of these analyses, we draw some substantive conclusions about the magnitude and sources of teacher effects on student achievement and suggest some ways that survey-based research on teaching can be improved.[1]

We first illustrate the varying analytic procedures that researchers have used to estimate the overall magnitude of teacher effects on student achievement, showing why previous research has led to conflicting conclusions. This issue has gained special salience in recent years as a result of William Sanders' (1998, p. 27) claim that "differences in [the] effectiveness of individual classroom teachers…[are] the *single largest* [contextual] factor affecting the academic growth of…students" (emphasis added). Sanders' conclusion, of course, is sharply at odds with findings from an earlier generation of research, especially production function research, showing that home and social background effects are more important than classroom and school effects in explaining variance in student achievement. We also discuss the conceptual and methodological foundations that underlie various claims about the magnitude of teacher effects on student achievement, and we present some empirical results that explain why analysts have reached differing conclusions about this topic.

We then shift from examining the overall effects of teachers on student achievement to an analysis of *why* such effects occur. Here, we review some findings from recently conducted, large-scale research on U.S. schooling. This literature has examined a variety of hypotheses about the effects of teachers' professional expertise, students' curricular opportunities, and classroom interaction patterns on students' achievement. Decades of research suggests that each of these factors can have effects on student learning, but the research also suggests that such effects are usually small and often inconsistent across grade levels, types of pupils, and academic subjects (Brophy & Good, 1986). We also review some common hypotheses about teacher effects on student achievement and use *Prospects* data to empirically assess both the size and consistency of these effects.

Finally, we review what we learned from these analyses and suggest some strategies for improving large-scale, survey research on teaching. We argue that large-scale, survey research has an important role to play in contemporary educational research, especially in research domains where education policy debates are framed by questions about "what works" and "how big" the effects of specific educational practices are on student achievement. But we also argue that large-scale, survey research on

teaching must evolve considerably before it can provide accurate information about such questions. In particular, our position is that future efforts by survey researchers should: (a) clarify the basis for claims about "effect sizes"; (b) develop better measures of teachers' knowledge, skill, and classroom activities; and (c) take care in making causal inferences from non-experimental data.

Examining the Size and Stability of Teacher Effects on Student Achievement

Our discussion of large-scale, survey research on teaching begins with questions about the size of teacher effects on student achievement. Researchers can use a variety of analytic procedures to estimate the overall magnitude of teacher effects on student achievement, but as we demonstrate below, these alternative procedures produce markedly different conclusions about this question. The overall purpose of this section, then, is to carefully describe the conceptual and methodological underpinnings of alternative approaches to estimating the magnitude of teacher effects on student achievement, and to clarify why different approaches to this problem produce the results they do.

Variance Decomposition Models

In educational research, the overall importance of some factor in the production of student learning is often judged by reference to the percentage of variance in student achievement accounted for by that factor in a simple variance decomposition model.[2] With the widespread use of hierarchical linear models, a large number of studies (from all over the world) have decomposed the variance in student achievement into components lying among schools, among classrooms within schools, and among students within classrooms. In review of this literature, Scheerens and Bosker (1997, p. 182-209) found that when student achievement was measured at a single point in time (and without controlling for differences among students in social background and prior achievement), about 15-20% of the variance in student achievement lies among schools, another 15-20% lies among classrooms within schools, and the remaining 60-70% of variance lies among students. Using the approach suggested by Scheerens and Bosker (1997, p. 74), these variance components can be translated into what Rosenthal (1994) calls a *d*-type effect size. The effect sizes for classroom-to-classroom differences in students' achievement in the findings just cited, for example, range from .39 to .45, "medium-sized" effects by the conventional standards of social science research.[3]

Although the review by Scheerens and Bosker (1997) is a useful starting point for a discussion of the overall magnitude of teacher effects on student achievement, it does not illustrate the full range of empirical strategies that researchers have used to address this question. As a result, we decided to analyze data from *Prospects* in order to duplicate and extend that analysis. In the following pages, we illustrate several alternative procedures for estimating the percentages of variance in students' achievement lying among schools, among classrooms within schools, and among students within classrooms. The analyses were conducted using the approach to hierarchical linear modeling (HLM) developed by Bryk and Raudenbush (1992), and were implemented using the

statistical computing software HLM/3L, version 5.25 (Bryk, Raudenbush, Cheong, & Congdon, 2000).

Analysis of *Prospects* Data

As a first step in the analysis, we duplicated the approach to estimating teacher effects on student achievement reported by Scheerens and Bosker (1997). The analysis was conducted using data on two cohorts of students in the *Prospects* study, those progressing from first to third grade over the course of the study, and those progressing from third to sixth grade. In the analyses, we simply decomposed the variance in students' achievement at a single point, using students' Item Response Theory (IRT) scale scores on the Comprehensive Test of Basic Skills reading and mathematics batteries as dependent variables. The analyses involves estimation of a simple, three-level, "random effects" model that Bryk and Raudenbush (1992, p. 176-178) call an "unconditional" model; that is, a model in which there are no independent variables. For each cohort, we conducted variance decompositions at each grade level for reading and mathematics achievement, yielding a total of 12 separate analyses. Across these analyses, we found that between 12-23% of the total variance in reading achievement, and between 18-28% of the total variance in mathematics achievement was among classrooms. Thus, the classroom effect sizes in these analyses ranged from about .35 to about .53 using the *d*-type effect size metric discussed by Scheerens and Bosker (1997, p. 74).

While these results duplicate those reported by Scheerens and Bosker (1997), they are not very good estimates of teacher effects on student achievement. One problem is that the analyses look at students' achievement *status* — that is, achievement scores at a single point in time. However, students' achievement status results not only from the experiences students had in particular classrooms during the year of testing, but also from *all* previous experiences students had, both in and out of school, prior to the point at which their achievement was assessed. As a result, most analysts would rather *not* estimate the effect of teachers on cumulative measures of achievement status, preferring instead to estimate the effect teachers have on *changes* in students' achievement during the time when students are in teachers' classrooms.

A second problem with these estimates is that they come from a "fully unconditional" model; that is, a model that does not control for the potentially confounding effects of students' socio-economic status and prior achievement on classroom-to-classroom differences in achievement. For example, at least some of the classroom-to-classroom differences in students' achievement status resulted not only from some teacher effect, but also from differences in the socioeconomic background and prior achievement of the students in different classrooms. Most analysts are unwilling to attribute compositional effects on achievement to teachers, and they therefore estimate teacher effects on student achievement only after controlling for such effects in their models.

These clarifications have led to the development of what researchers call "value-added" analyses of teacher effects. Value-added models have two key features. First, the dependent variables in the analysis are designed to measure the amount of *change* that occurs in students' achievement during the year when students are in the classrooms under

study. Second, measures of change are adjusted for differences across classrooms in students' prior achievement, home and social background, and the social composition of the schools students attended. The purpose of value-added models is to estimate the proportions of variance in *changes* in student achievement lying among classrooms, after controlling for the effects of other confounding variables.

To see whether value-added models give different results than those previously discussed, we conducted further analyses using *Prospects* data. In these analyses, we used two of the most common empirical approaches to value-added estimates of teacher effects on student achievement. The first approach is often called a "covariate adjustment" model. Here, students' achievement status in a given year is adjusted for students' prior achievement, home and social background, and the social composition of schools, and the variance in students' "adjusted" achievement status is decomposed into school, classroom, and student components using the same three-level hierarchical linear model as before.[4] Using this approach with *Prospects* data, we found that roughly 4-16% of the variance in students' adjusted reading achievement was lying among classrooms (depending on the grade level in the analysis), and that roughly 8-18% of the variance in adjusted mathematics achievement was lying among classrooms (depending on the grade at which the analysis was conducted). In the covariate adjustment models, then, the *d*-type effect sizes for classrooms ranged between .21 and .42 depending on the grade level and subject under study, somewhat less than the effect sizes in the fully unconditional models.[5]

A second approach to value-added analysis uses students' *annual gains* in achievement as the criterion variable. In this approach, students' gain scores for a given year become the dependent variable in the analysis, where these gains are once again adjusted through regression analysis for the potential effects of students' socioeconomic status, family background, prior achievement, and school composition (using variables discussed in endnote 4). Using this approach with *Prospects* data, we found that somewhere between 3% and 10% of the variance in adjusted gains in students' reading achievement was lying among classrooms (depending on the grade being analyzed), and somewhere between 6% and 13% of the variance in adjusted gains in mathematics was lying among classrooms. The corresponding *d*-type effect sizes in these analyses therefore range from .16 to .36.

Problems with Conventional Analyses

Neither of the value-added analyses discussed indicates that classroom effects on student achievement are large. But each suffers from important interpretive and methodological problems warranting more discussion. Consider, first, some problems with covariate adjustment models. Several analysts have demonstrated that covariate adjustment models do not really model *changes* in student achievement (Rogosa, 1995; Stoolmiller & Bank, 1995). Instead, such analyses are simply modeling students' achievement status, which in a value-added framework has been adjusted for students' social background and prior achievement. When viewed in this way, it is not surprising to find that teacher effects are relatively small in covariate adjustment models. Such models, in fact,

are assessing teacher effects on achievement *status*, not change.

If one really wants to assess the size of teacher effects on changes in student achievement, models of annual gains in achievement are preferable. As Rogosa (1995) demonstrates, annual gains in achievement are unbiased estimates of students' "true" rates of achievement growth and are therefore preferable to covariate adjustment models in the analysis of change. However, simple gain scores suffer from an important methodological problem that researchers need to guard against. As Rogosa (1995) demonstrates, when there is little variance among students in true rates of academic growth, annual gains in achievement provide very unreliable measures of underlying differences among students in rates of change. In addition, in variance decomposition models using gain scores, measurement error due to unreliability in the gain scores will be reflected in student-level variance components, increasing the denominator in effect size formulas and thus reducing teacher effect size coefficients. In fact, as we discuss below, this problem is present in the *Prospects* data, where differences among students in true rates of academic growth are quite small. For this reason, the effect sizes derived from the gain score models discussed in this report are almost certainly *under*estimates of the overall effects that classrooms have on growth in students' achievement.

Improving Estimates of Teacher Effects

What can researchers do in light of the problems just noted? One obvious solution is to avoid the covariate adjustment and gains models used in previous research, and to instead use statistical models that directly estimate students' individual "growth curves" (Rogosa, 1995). In current research, the statistical techniques developed by Bryk and Raudenbush (1992, chap. 6), as implemented in the statistical computing package HLM/3L (Bryk, Raudenbush, Cheong, & Congdon, 2000) are frequently used for this purpose. For example, the HLM/3L statistical package can be used to estimate students' growth curves directly if there are at least three data points on achievement for most students in the data set. However, at the current time, this computing package cannot be used to estimate the percentages of variance in rates of achievement growth lying among classrooms within schools over time, for as Raudenbush (1995) demonstrated, estimation of these variance components within a growth modeling framework requires development of a "cross-classified" random effects model.[6]

Fortunately, the computer software needed to estimate cross-classified random effects models within the framework of the existing HLM statistical package is now under development, and we have begun working with Raudenbush to estimate such models using this computing package. A detailed discussion of the statistical approach involved here is beyond the scope of this report, but suffice it to say that it is an improvement over the simple gains models discussed earlier, especially since the cross-classified random effects model allows us to estimate the random effects of classrooms on student achievement within an explicit growth modeling framework.[7]

For this report, we developed a three-level, cross-classified, random effects model to analyze data on the two cohorts of students in the *Prospects* data set

discussed earlier. In these analyses, we decomposed the variance in students' growth in achievement (in mathematics and reading) into variance lying among schools, among students within schools, within students across time, and among students within classrooms. Two important findings have emerged from these analyses. One is that only a small percentage of variance in rates of achievement growth lies among students. In cross-classified random effects models that include all of the control variables listed in endnote 4, for example, about 27-28% of the reliable variance in reading growth lies among students (depending on the cohort), with about 13-19% of the reliable variance in mathematics growth lying among students. An important implication of these findings is that the "true score" differences among students in academic growth are quite small, raising questions about the reliability of the gain scores used in the analysis of *Prospects* data discussed above.

More important for our purposes is a second finding. The cross-classified random effects models produce very different estimates of the overall magnitude of teacher effects on *growth* in student achievement than do simple gain scores models. For example, in the cross-classified random effects analyses, we found that after controlling for student background variables, the classrooms to which students were assigned in a given year accounted for roughly 60-61% of the reliable variance in students' rates of academic growth in reading achievement (depending on the cohort), and 52-72% of the reliable variance in students' rates of academic growth in mathematics achievement. This yields *d*-type effect sizes ranging from .77 to .78 for reading growth (roughly two-to-three times what we found using a simple gains model), and *d*-type effect sizes ranging from .72 to .85 for mathematics growth (again, roughly two-to-three times what we find using a simple gains model).[8] The analysis also showed that school effects on achievement growth were substantial in these models (d = .55 for reading, and d = .53 for mathematics).[9]

The Consistency of Classroom Effects Across Different Academic Subjects and Pupil Groups

The analyses suggest that the classrooms to which students are assigned in a given year can have non-trivial effects on students' achievement growth in that year. But this does not exhaust the questions we can ask about such effects. An additional set of questions concern the consistency of these effects, for example, across different subjects (i.e., reading and mathematics) and/or for different groups of pupils. We have been unable to find a great deal of prior research on these questions, although Brophy and Good's (1986) seminal review of "process-product" research on teaching did discuss a few studies in this area. For example, Brophy and Good cite a single study showing a correlation of .70 for adjusted, classroom-level gains across tests of word knowledge, word discrimination, reading, and mathematics. They also cite correlations ranging from around .20 to .40 in the adjusted gains produced by the same teacher across years, suggesting that the effectiveness of a given teacher can vary across different groups of pupils. Both types of findings, it is worth noting, are comparable to findings on the consistency of school effects across subjects and pupil groups (see Scheerens & Bosker, 1997, chap. 3).

Given the sparseness of prior research on these topics, we turned to *Prospects* once again for relevant insights. To assess

whether classrooms had consistent effects on students' achievement across different academic subjects, we simply correlated the residuals from the value-added gains models for each classroom.[10] Recall that these residuals are nothing more than the deviations in actual classroom gains from the gains predicted for a classroom after adjusting for the student- and school-level variables in our models. In the analyses, we found only a moderate degree of consistency in classroom effects across reading and mathematics achievement, with correlations ranging from .30 to .47 depending on the grade level of the classrooms under study.[11] The results therefore suggest that a given teacher varies in effectiveness when teaching different academic subjects. In *Prospects* data, there was slightly less variation in teacher effects across academic subjects at later grades, but this could be a cohort effect, since different groups of pupils are in the samples in earlier and later grades.

A second question we investigated was whether classrooms had consistent effects on students from different social backgrounds. To investigate this issue, we changed the specification of the value-added regression models. In previous analyses, we were assuming that the effects of student-level variables on annual gains in achievement were the same in all classrooms. In this phase of the analysis, we allowed the effects of student socioeconomic status (SES), gender, and minority status on achievement gains to vary randomly across classrooms. Since the data set contains relatively few students per classroom, we decided to estimate models in which the effects of only one of these independent variables was allowed to vary randomly in this way in any given regression analysis.

Overall, the analyses showed that background variables had different effects on annual gains in achievement across classrooms, with these random effects being larger in lower grades (especially in reading) than at upper grades. Thus, in the *Prospects* study, students from different social backgrounds apparently did not perform equally well across classrooms within the same school. Moreover, when the variance components for these additional random effects were added to the variance components for the random effects of classrooms, the overall effects of classrooms on gains in student achievement became larger. In early grades reading, for example, the addition of random effects for background variables approximately doubles the variance in achievement gains accounted for by classrooms (the increase is much less, however, for early grades mathematics, and also less for upper grades mathematics and reading). For example, in a simple gains model where only the main effects of classrooms are treated as random, the *d*-type effect size was .26. When we also allowed background effects to vary across classrooms, however, the *d*-type effect sizes became .36 when the male effect was treated as random, .26 when the SES effect was allowed to vary, and .38 when the minority effect was allowed to vary.

Student Pathways through Classrooms

A third issue we examined was the consistency of classroom effects for a given student across years. We have seen that in any given year, students are deflected upward or downward from their expected growth trajectory by virtue of the classrooms to which they are assigned. This occurs, of course, because some classrooms are more effective at

producing academic growth for students, with the *d*-type effect size for annual deflections being around .72 to .85 in cross-classified random effects models (and around .16 to .36 when measured in terms of annual gains in achievement). In any given year, such effects may not seem especially sizeable. But if some students were consistently deflected upward as a result of their classroom assignments during elementary school, while other students were consistently deflected downward, the cumulative effects of classroom placement on academic growth could be quite sizeable, producing substantial inequality in student achievement in elementary schools.

Currently, we know very little about this process in U.S. elementary schools. Instead, the most important evidence comes from Kerckhoff's (1983) seminal study of schools in Great Britain. Kerckhoff tallied the accumulated deflections to expected academic growth for students as they passed through British schools and found that the accumulation of consistently positive or negative deflections was much greater in British secondary schools than in primary schools. A similar process might be occurring in the United States, where elementary schools have a common curriculum, classrooms tend to be heterogeneous in terms of academic and social composition, and tracking is not a part of the institutional landscape. Since this is the case, elementary schools do not appear to be explicitly designed to produce academic differentiation. As a result, we might expect the accumulation of classroom effects on student achievement to be fairly equal over the course of students' careers in elementary schools.

To get a sense of this issue, we analyzed the classroom-level Empirical Bayes (EB) residuals from the cross-classified growth models estimated above. Recall that these models control for a large number of student and school variables. In the analysis, we first calculated the classroom residuals for each student at each time point. We then correlated these residuals at the student level across time points. In the analysis, a positive correlation of residuals would indicate that students who experienced positive deflections in one year also experienced positive deflections in the following year, suggesting that classroom placements in elementary schools worked to the consistent advantage of some students and to the consistent disadvantage of others. What we found in the *Prospects* data, however, was that deflections were inconsistently correlated across successive years, sometimes being positive, sometimes being negative, and ranging from -.30 to +.18. Overall, this pattern suggests that, on average, within a given school, a student would be expected to accumulate no real learning advantage by virtue of successive classroom placements.

Note, however, that these data are not showing that students *never* accumulate successively positive (or negative) deflections as a result of their classroom placements. In fact, some students do experience consistent patterns. But in these data, such patterns should be exceedingly rare. For example, assuming that classroom effects are uncorrelated over time, we would expect about 3% of students to experience positive deflections one standard deviation or more above their expected gain for two years in a row, and less than 1% to receive such positive deflections three years in a row. Another 3% of students in a school would receive two straight years of negative deflections of this magnitude, with less than 1% receiving three straight

negative deflections. Obviously, students who experience consistently positive or negative deflections will end up with markedly different cumulative gains in achievement over the years (Sanders, 1998, p. 27). But the data analyzed here suggest that such differences arise almost entirely by chance, not from a systematic pattern of academic differentiation through successively advantaging or disadvantaging classroom placements.

The following results further illustrate this point. Using the EB residuals, we classified students according to whether (in a given year) they were in classrooms that were one standard deviation above the mean in effects on achievement growth, one standard deviation below the mean, or somewhere in between. Overall, when data on both cohorts and for both academic subjects are combined, we found that 3.4% of the students were in classrooms one standard deviation above the mean in two consecutive years, while 2.4% of the students were in classrooms one standard deviation below the mean for two consecutive years. Across three years, .45% of students were in classrooms one standard deviation above the mean for three consecutive years, and .32% were in classrooms one standard deviation below the mean for three consecutive years. To be sure, students accumulated different classroom deflections to growth over time, and this produced inequalities in achievement among students. But the pattern of accumulation here appears quite random, and not at all the result of some systematic process of social or academic differentiation.

Summary

What do the findings suggest about the overall size and stability of teacher effects on student achievement? On the basis of the analyses reported, it seems clear that assertions about the magnitude of teacher effects on student achievement depend to a considerable extent on the methods used to estimate these effects and on how the findings are interpreted. With respect to issues of interpretation, it is not surprising that teacher effects on students' achievement status are small in variance decomposition models, even in the earliest elementary grades. After all, status measures reflect students' cumulative learning over many years, while teachers have students in their classrooms only for a single year. In this light, the classroom effects on students' achievement status found in *Prospects* data might be seen as surprisingly large. In elementary schools, *Prospect*s data suggest that after controlling for student background and prior achievement, the classrooms to which students are assigned account for somewhere between 4-18% of the variance in students' cumulative achievement status in a given year, which translates into a *d*-type effect size of .21 to .42.

As we have seen, however, most analysts do not want to analyze teacher effects on achievement status, preferring instead to examine teacher effects on students' academic growth. Here, the use of gain scores as a criterion variable is common. But analyses based on gain scores are problematic. While annual gains provide researchers with unbiased estimates of true rates of change in students' achievement, they can be especially unreliable when true differences among students in academic growth are small. In fact, this was the case in *Prospects* data, and the resulting

unreliability in achievement gains probably explains why we obtained such small effect size coefficients when we used gain scores to estimate teacher effects. Recall that in these analyses, only 3-13% of the variance in students' annual achievement gains was found to be lying among classrooms.

One clear implication of these analyses is that researchers need to move beyond the use of both covariate adjustment models (which estimate effects on students' adjusted achievement status) and annual gains models if they want to estimate the overall magnitude of teacher effects on growth in student achievement. A promising strategy here is to use a cross-classified random effects model, as Raudenbush (1995) and Raudenbush and Bryk (2002, chap. 12) discuss. The preliminary analysis of *Prospects* data reported here suggests that cross-classified random effects models will lead to findings of larger *d*-type teacher effects. For example, in the cross-classified random effects analysis discussed in this report, we reported *d*-type effect sizes of .77 to .78 for teacher effects on students' growth in reading achievement, and *d*-type effect sizes of .72 to .85 for teacher effects on students' growth in mathematics achievement. These are roughly three times the effect size found in other analyses.

In this report, we also presented findings on the consistency of teacher effects across academic subjects and groups of pupils. Using a gains model, we found that the same classroom was not consistently effective across different academic subjects or for students from different social backgrounds. We also used a cross-classified random effects model to demonstrate that cumulative differences in achievement among students resulting from successive placements in classrooms could easily have resulted from successive chance placements in more and less effective classrooms. This latter finding suggests that elementary schools operate quite equitably in the face of varying teacher effectiveness, allocating pupils to more and less effective teachers on what seems to be a chance rather than a systematic basis.

While the equity of this system of pupil allocation to classrooms might be comforting to some, the existence of classroom-to-classroom differences in instructional effectiveness should not be. As a direct result of teacher-to-teacher differences in instructional effectiveness, some students make *less* academic progress than they would otherwise be expected to make simply by virtue of chance placements in ineffective classrooms. All of this suggests that the important problem for U.S. education is not simply to demonstrate that differences in effectiveness exist among teachers, but rather to explain why these differences occur and to improve teaching effectiveness broadly.

What Accounts for Classroom-to-Classroom Differences in Achievement?

Up to this point, we have been reviewing evidence on the overall size of teacher effects on student achievement. But these estimates, while informative about how the educational system works, do not provide any evidence about why some teachers are more instructionally effective than others. In order to explain this phenomenon, we need to inquire about the properties of teachers and their

teaching that produce effects on students' growth in achievement.

In this section, we organize a discussion of this problem around Dunkin and Biddle's (1974) well-known scheme for classifying types of variables in research on teaching. Dunkin and Biddle were working within the "process-product" paradigm and discussed four types of variables of relevance to research on teaching. *Product* variables were defined as the possible outcomes of teaching, including student achievement. *Process* variables were defined as properties of the interactive phase of instruction; that is, the phase of instruction during which students and teachers interact around academic content. *Presage* variables were defined as properties of teachers that can be assumed to operate prior to, but also to have an influence on, the interactive phase of teaching. Finally, *context* variables were defined as variables that can exercise direct effects on instructional outcomes and/or condition the effects of process variables on product variables.

Presage Variables

The process-product paradigm discussed by Dunkin and Biddle (1974) arose partly in response to a perceived over-emphasis on presage variables in early research on teaching. Among the presage variables studied in such work were teachers' appearance, enthusiasm, intelligence, and leadership — so-called "trait" theories of effective teaching (Brophy & Good, 1986). Most of these trait theories are no longer of interest in research on teaching, but researchers have shown a renewed interest in other presage variables in recent years. In particular, researchers increasingly argue that teaching is a form of expert work that requires extensive professional preparation, strong subject-matter knowledge, and a variety of pedagogical skills, all of which are drawn upon in the complex and dynamic environment of classrooms (for a review of conceptions of teachers' work in research on teaching, see Rowan, 1999). This view of teaching has encouraged researchers once again to investigate the effects of presage variables on student achievement.

In large-scale, survey research, teaching expertise is often measured by reference to teachers' educational backgrounds, credentials, and experience. This is especially true in the so-called "production function" research conducted by economists. Since employment practices in U.S. education entail heavy reliance on credentials, with more highly educated teachers, those with more specialized credentials, or those with more years of experience gaining higher pay, economists have been especially interested in assessing whether teachers with different educational backgrounds perform differently in the classroom. In this research, teachers' credentials are seen as "proxies" for the actual knowledge and expertise of teachers, under the assumption that teachers' degrees, certification, or experience index the instructionally relevant knowledge that teachers bring to bear in classrooms.

In fact, research on presage variables of this sort has a long history in large-scale studies of schooling. Decades of research have shown, for example, that there is no difference in adjusted gains in student achievement across classes taught by teachers with a Master's or other advanced degree in education compared to classes taught by teachers who lack such degrees. However, when large-scale research has focused in greater detail on the academic majors of teachers and/or

on the courses teachers have taken, results have been more positive. For example, several large-scale studies (reviewed in Rowan, Chiang, & Miller, 1997, and Brewer & Goldhaber, 2000) have tried to assess the effect of *teachers' subject-matter knowledge* on student achievement by examining differences in student outcomes for teachers with different academic majors. In general, these studies have been conducted in high schools and have shown that in classes where teachers have an academic major in the subject area being tested, students have higher adjusted achievement gains. In the *NELS:88* data, for example, the *r*-type effect sizes for these variables were .05 for science gains, and .01 for math gains.[12] Other research suggests an extension of these findings, however. At least two studies, using different data sets, suggest that the gains to productivity coming from increases in high school teachers' subject-matter coursework occur mostly when advanced material is being taught (see, for example, Monk, 1994 and Chiang, 1996).[13] Fewer production function studies have used teachers' professional preparation as a means of indexing teachers' *pedagogical knowledge*, although a study by Monk (1994) is noteworthy in this regard. In Monk's study, the number of classes in subject-matter pedagogy taken by teachers during their college years was found to have positive effects on high school students' adjusted achievement gains. Darling-Hammond, Wise, and Klein (1995) cite additional, small-scale studies supporting this conclusion.

Analyses of Presage Variables

As a follow-up to this research, we examined the effects of teachers' professional credentials (and experience) on student achievement using *Prospects* data. In these analyses, we developed a longitudinal data set for two cohorts of students in the *Prospects* study: students passing from grades one through three over the course of the study, and students passing from grades three through six. Using these data, we estimated an explicit model of students' growth in academic achievement using the statistical methods described in Bryk and Raudenbush (1992, p. 185-191) and the computing software HLM/3L, version 5.25 (Bryk, Raudenbush, Cheong, & Congdon, 2000). Separate growth models were estimated for each cohort of students, and for each academic subject (reading and mathematics). Thus, the analyses estimated four distinct growth models: (a) a model for growth in reading achievement in grades one through three, (b) a model for growth in mathematics achievement in grades one through three, (c) a model for growth in reading achievement in grades three through six, and (d) a model for growth in mathematics achievement in grades three through six.

In all of these analyses, achievement was measured by the IRT scale scores provided by the test publisher. The reader will recall that these are equal-interval scores (by assumption), allowing researchers to directly model growth across grades using an equal-interval metric. In all analyses, students' growth in achievement was modeled in quadratic form, although the effect of this quadratic term was fixed. In the early grades cohort, the results showed that students' growth in both reading and mathematics was steep in initial periods but decelerated over time. In the upper grades, academic growth in reading was linear, while growth in mathematics achievement accelerated at the last point in the time series. Average growth rates for both reading and mathematics were much

lower in the upper grades than in the lower grades.

In all of the models, we estimated the effects of home and social background on both achievement status and achievement growth, where the variables included: (a) gender, (b) SES, (c) minority status, (d) number of siblings, (e) family marital status, and (f) parental expectations for a student's educational attainment. In general, these variables had very large effects on students' achievement status, but virtually no effects on growth in achievement. We also controlled for school composition and location in these analyses, where the social composition of schools was indexed by the percentage of students in a school eligible for the federal free lunch program, and where location was indexed by whether or not a school was in an urban location. Here, too, the school-level variables had large effects on intercepts but not on growth. All of these results are important — suggesting that when the analysis shifts from concern with students' achievement status to concern with students' growth in achievement, home and social background, as well as school composition and location, become relatively insignificant predictors of academic development.

In our analysis of presage variables using *Prospects* data, we focused on three independent variables measuring teachers' professional background and experience. One was a measure of whether or not a teacher had special certification to teach reading or mathematics. The second was a measure of whether or not a teacher had a Bachelor's or Master's degree in English (when reading achievement was the dependent variable) or in mathematics (when mathematics was tested). Third, we reasoned that teacher experience could serve as a proxy for teachers' professional knowledge, under the assumption that teachers learn from experience about how to represent and teach subject-matter knowledge to students. The reader is cautioned that very few teachers in the *Prospects* sample (around 6%) had special certification and/or subject-matter degrees. For this (and other reasons), we used the robust standard errors in the HLM statistical package to assess the statistical significance of the effects of these variables on growth in student achievement.

The analyses were conducted using a three-level hierarchical linear model of students' growth in academic achievement, where classroom variables are included at level one of the model as time-varying covariates.[14] The results of these analyses were reasonably consistent across cohorts in the *Prospects* data, but differed by academic subject. In reading, neither teachers' degree status nor teachers' certification status had statistically significant effects on growth in students' achievement, although we again caution the reader about the small number of teachers in this sample who had subject-matter degrees or special certification. In reading, however, teacher experience was a statistically significant predictor of growth in students' achievement, the *d*-type effect size being $d = .07$ for early grades reading and $d = .15$ for later grades reading.[15] In mathematics, the results were different, and puzzling. Across both cohorts of students, there were no effects of teachers' mathematics certification on growth in student achievement. There was a positive effect of teachers' experience on growth in mathematics achievement, but only for the later grades cohort ($d = .18$).[16] Finally, in mathematics and for both cohorts, students who were

taught by a teacher with an advanced degree in mathematics did worse than those who were taught by a teacher not having a mathematics degree (d = -.25).[17]

It is difficult to know how to interpret the *negative* effects of teachers' mathematics degree attainment on students' growth in mathematics achievement. On one hand, the negative effects could reflect selection bias (see also endnote 13, where this is discussed in the context of high school data). In elementary schools, for example, we might expect selection to *negatively* bias estimated teacher effectiveness, especially if teachers with more specialized training work in special education and/or compensatory classroom settings. In a subsidiary analysis, we re-specified the regression models to control for this possibility (by including measures of students' special education, compensatory education, or gifted and talented classification), but the effects remained unchanged. The other possibility is that this is a real effect, and that advanced academic preparation is actually negatively related to students' growth in achievement in elementary schools. Such an interpretation makes sense only if one assumes that advanced academic training somehow interferes with effective teaching, either because it substitutes for pedagogical training in people's professional preparation, or because it produces teachers who somehow cannot simplify and clarify their advanced understanding of mathematics for elementary school students.

Discussion of Presage Variables

What is interesting about production function studies involving presage variables is how disconnected they are from mainstream research on teaching. Increasingly, discussions of teachers' expertise in mainstream research on teaching have gone well beyond a concern with proxy variables that might (or might not) index teachers' expertise. Instead, researchers are now trying to formulate more explicit models of what teaching expertise looks like. In recent years, especially, discussions of expertise in teaching often have been framed in terms of Shulman's (1986) influential ideas about pedagogical content knowledge. Different analysts have emphasized different dimensions of this construct, but most agree that there are several dimensions involved. One is teachers' knowledge of the content being taught. At the same time, teaching is also expected to require knowledge of how to represent that content to different kinds of students in ways that produce learning, and that, in turn, requires teachers to have a sound knowledge of the typical ways students understand particular topics or concepts within the curriculum, and of the alternative instructional moves that can produce new understandings in light of previous ones.

None of this would seem to be well measured by the usual proxies used in production function studies, and as a result, many researchers have moved toward implementing more direct measures of teachers' expertise. To date, most research of this sort has been qualitative and done with small samples of teachers. A major goal has been to describe in some detail the pedagogical content knowledge of teachers, often by comparing the knowledge of experts and novices. Such work aims to clarify and extend Shulman's (1986) original construct. One frustrating aspect of this research, however, is that it has been conducted in relative isolation from large-scale, survey research on teaching, especially the long line of production

function studies just discussed. Thus, it remains to be seen if more direct measures of teachers' knowledge will be related to students' academic performances.

It is worth noting that prior research has found positive effects of at least some direct measures of teachers' knowledge on student achievement. For example, large-scale research dating to the Coleman report (Coleman et al., 1966) suggests that verbal ability and other forms of content knowledge are significantly correlated to students' achievement scores as the meta-analysis reported in Greenwald, Hedges, and Laine (1996) shows. This is complemented by more recent work showing that teachers' scores on teacher certification tests and college entrance exams also affect student achievement (for a review, see Ferguson & Brown, 2000). It should be noted, however, that Shulman's (1986) original conception of "pedagogical" content knowledge was intended to measure something other than the "pure" content knowledge measured in the tests just noted. As Shulman (1986) pointed out, it would be possible to know a subject well, but lack the knowledge to translate this kind of knowledge into effective instruction for students.

Given the presumed centrality of teachers' pedagogical expertise to teaching effectiveness, a logical next step in large-scale, survey research is to develop *direct* measures of teachers' pedagogical and content knowledge and to estimate the effects of these measures on growth in students' achievement. In fact, along with colleagues, we are currently taking steps in this direction.[18] Our efforts originated in two lines of work. The first was the Teacher Education and Learning to Teach (*TELT*) study conducted at Michigan State University. The researchers who conducted this study developed a survey battery explicitly designed to assess teachers' pedagogical content knowledge in two areas — mathematics and writing (Kennedy, 1993). Within each of these curricular areas, a battery of survey items was designed to assess two dimensions of teachers' pedagogical content knowledge: (a) teachers' knowledge of subject matter, and (b) teachers' knowledge of effective teaching practices in a given content area. As reported in Deng (1995), the attempt to construct these measures was more successful in the area of mathematics than in writing, and more successful in measures of content knowledge than pedagogical knowledge.

An interesting offshoot of this work is that one of the items originally included as a measure of pedagogical content knowledge in the *TELT* study was also included in the *NELS:88* teacher questionnaire. As a result, we decided to investigate the association between this item and student achievement in the *NELS:88* data on 10th-grade math achievement. As reported in Rowan, Chiang, and Miller (1997), we found that in a well-specified regression model predicting adjusted gains in student achievement, the item included in the *NELS:88* teacher questionnaire had a statistically significant effect on student achievement. In this analysis, a student whose teacher provided a correct answer to this single item scored .02 standard deviations higher on the *NELS:88* mathematics achievement test than did a student whose teacher did not answer the item correctly. The corresponding *r*-type effect size for this finding is $r = .03$, and $R^2 = .0009$.[19]

Although the effect sizes in the *NELS: 88* analysis are tiny, the measurement

problems associated with an *ad hoc*, one-item scale measuring teachers' content knowledge are obvious. Moreover, the effect of this *ad hoc* measure of teachers' knowledge was assessed in Rowan, Chiang, and Miller's (1997) analysis by reference to a covariate adjustment model of students' 10th-grade achievement status. As a result, one should not expect large effects from such an analysis. For this reason, our colleagues are now developing an extensive battery of survey items to directly assess teachers' pedagogical content knowledge in the context of elementary schooling. Our development work to date is promising. For example, we have found that we can construct highly reliable measures of teachers' pedagogical content knowledge within fairly narrow domains of the school curriculum using as few as six-to-eight survey items. Our goal in the future is to estimate the effects of these measures on growth in students' achievement in our own study of school improvement interventions.[20]

Teaching Process Variables

Although presage variables of the sort just discussed, if well-measured, hold promise for explaining differences in teacher effectiveness, quantitative research on teaching for many years has focused more attention on process-product relationships than on presage-product relationships. In this section, we discuss prior research on the effects of teaching process variables on student achievement and describe how we examined such effects using *Prospects* data.

Time-on-Task/Active Teaching

One aspect of instructional process that has received a great deal of attention in research on teaching is "time-on-task." A sensible view of this construct, based on much previous process-product research, would refer not so much to the amounts of time allocated to learning a particular subject, which has virtually no effect on achievement, nor even to the amount of time in which students are actively engaged in instruction, for high inference measures of student engagement during class time also have only very weak effects on achievement (Karweit, 1985). Rather, process-product research suggests that the relevant causal agent producing student learning is how teachers *use* instructional time.

Brophy and Good's (1986) review of process-product research on teaching suggests that effective use of time involves "active" teaching. In their view, active teaching occurs when teachers spend more time in almost any format that directly instructs students, including lecturing, demonstrating, leading recitations and discussions, and/or frequently interacting with students during seatwork assignments. This kind of teaching contrasts with a teaching style in which students frequently work *independently* on academic tasks and/or are engaged in non-academic work. Active teaching also involves good classroom management skills, for example, the presence of clear rules for behavior with consistent enforcement, close and accurate monitoring of student behavior, and the quick handling of disruptions and/or transitions across activities.

There are several interesting points about these findings. The most important is that the concept of active teaching is generic. That is, research shows that active teaching looks much the same across academic subjects and positively affects student achievement across a range of grade levels and subjects. At the

same time, the concept does *not* imply that a particular instructional format (e.g., lecture and demonstration, recitation, or other forms of guided discussion) is generally more effective than another across academic subjects and/or grade levels. In fact, the findings presented in Brophy and Good (1986) suggest that what is important is not how a teacher is active (i.e., the activities he or she engages in) as much as that the teacher is — in fact — an active *agent* of instruction. Thus, we can expect to find variability in the frequency and effectiveness of various instructional formats, but in virtually all settings, high achievement growth is expected to occur when the teacher is actively carrying the material to students as opposed to allowing students to learn without scaffolding, supervision, and feedback.

Analysis of Time-on-Task/Active Teaching Measures

To see if patterns of active teaching help explain classroom-to-classroom differences in students' academic growth, we analyzed the effects on growth in achievement of several measures of active teaching available for upper grades classrooms in *Prospects* data.[21] The measures were taken from three types of questions on the teacher questionnaire. One question asked teachers to report on the average minutes per week spent in their classrooms on instruction in reading and mathematics. The second asked teachers to rate the percentage of time they spent engaged in various active teaching formats, including time spent: (a) presenting or explaining material, (b) monitoring student performance, (c) leading discussion groups, and (d) providing feedback on student performance. The third asked teachers to rate the percentage of time that students in their classrooms spent in "individualized" and "whole-class" instruction.

Following the review of evidence on active teaching mentioned earlier, we reasoned that what would matter most to student achievement was not the amount of time teachers spent on instruction, nor even how teachers distributed their time across various active teaching behaviors. Instead, we hypothesized that the important variable would be how much active teaching occurred. From this perspective, we predicted that there would be no effect of minutes per week of instruction in reading or math on student achievement, and no effect of the instructional format variables (a-d above). What would matter most, we reasoned, was the extent to which the teacher was operating as an active agent of instruction. From this perspective, we predicted that the percentage of time students spent in individualized instruction (where students work alone) would indicate a *lack* of active teaching and would have negative effects on students' growth in achievement. By contrast, we reasoned that the percentage of time spent in whole-class instruction (where teachers are the active agents of instruction) would have positive effects.

To conduct this analysis, we simply re-specified the HLM growth analyses used in estimating the effects of teacher certification and experience so that it now included the active teaching variables. As expected, teachers' reports about minutes per week spent in instruction, and their reports on the teaching format variables, did not have statistically significant effects on students' growth in reading or mathematics achievement. The results for time spent on individualized instruction were mixed, but generally supportive of our hypotheses. For reading, the data were consistent with the prediction that

more time spent by students in individualized settings translated into less academic growth, the effect size here being d = -.09.[22] In mathematics, however, time spent on individualized instruction had no significant effect. The data on percentage of time spent in whole-class instruction were consistently supportive of our hypothesis. In both reading and mathematics, this variable was statistically significant. In reading, the effect size was d = .09. In mathematics, the effect size was d = .12.[23]

Discussion of Time-on-Task/Active Teaching Variables

The results from the *Prospects* analyses appear remarkably consistent with previous process-product research and confirm that active teaching (as carried out in a whole-class setting) can have a positive effect on students' growth in achievement. However, the results reported here probably do not provide a very accurate indication of the *magnitude* of this effect for several reasons. For one, items in the *Prospects* teacher questionnaire forced teachers to report on their use of different instructional behaviors and settings by averaging across all of the academic subjects they taught. Yet Stodolsky (1988) has found that the mix of instructional activities and behavior settings used by the same teacher can differ greatly across subjects. Moreover, a great deal of research on the ways in which respondents complete questionnaires suggests that the kinds of questions asked on the *Prospects* teacher questionnaire — questions about how much time was spent in routine forms of instructional activities — cannot be responded to accurately in "one-shot" questionnaires. This lack of accuracy probably introduces substantial error into our analyses, biasing all effect sizes downward and perhaps preventing us from discovering statistically significant relationships among teaching processes and student achievement.

Opportunity-to-Learn/Content Covered

In addition to active teaching, process-product research also consistently finds a relationship between the curricular content covered in classrooms and student achievement. However, definitions and measures of curricular content vary from study to study, with some studies measuring only the content that is covered in a classroom, and other studies measuring both the content covered and the "cognitive demand" of such content.

Any serious attempt to measure content coverage begins with a basic categorization of curriculum topics in a particular subject area (e.g., math, reading, writing, etc.). Such categorization schemes have been derived from many different sources, including curriculum frameworks or standards documents, textbooks, and items included in the achievement test(s) being used as the dependent variable(s) in a process-product study. In most research on content coverage, teachers are asked to rate the amount of emphasis they place on each topic in the content list developed by researchers. Across all such studies, the procedures used to measure content coverage vary in two important respects. First, some surveys list curriculum content categories in extremely fine-grained detail while others are more course-grained. Second, teachers in some studies fill out these surveys on a daily basis, while in most studies, they fill out an instrument once annually, near the end of the year.

Obviously, measures of content coverage can serve either as dependent or independent variables in research on teaching for it is as interesting to know why content coverage differs across teachers as it is to know about the effects of content coverage on student achievement. When the goal of research is to predict student achievement, however, a common approach has been to measure the amount of overlap in content covered in a classroom with the content assessed in the achievement test serving as the dependent measure in a study. A great deal of research, ranging from an early study by Cooley and Leinhardt (1980) to more recent results from the Third International Mathematics and Science Study assessments (Stedman, 1997), have used this approach. These studies uniformly show that students are more likely to answer items correctly on an achievement test when they have received instruction on the topics assessed by that item. In fact, the degree of overlap between content covered in a classroom and content tested is a consistent predictor of student achievement scores.[24]

In addition to measuring topics covered, it can be useful to examine the cognitive objectives that teachers are seeking to achieve when teaching a given topic. In research on teaching, the work of Andrew Porter and colleagues is particularly noteworthy in this regard. In Porter's work, curriculum coverage is assessed on two dimensions — what topics are covered *and* for each topic, the level of cognitive demand at which that topic is covered where cognitive demand involves rating the complexity of work that students are required to undertake in studying a topic. Recently, Porter and colleagues have found that the addition of a cognitive demand dimension to the topic coverage dimension increases the power of content measures to predict gains in student achievement (Gamoran, Porter, Smithson, & White, 1997).

Analysis of Content Covered

To examine the effects of content coverage on student achievement, we conducted an analysis of *Prospects* data. In the *Prospects* study, teachers filled out a questionnaire near the end of the year in which they were asked to rate the amount of emphasis they gave to several broad areas of the reading and mathematics curricula using a three-point rating scale (ranging from no emphasis, to moderate emphasis, to a great deal of emphasis). From these data, we were able to construct two measures of content coverage — one in reading for the lower grades cohort (sufficient items for a scale were not available for the upper grades), and one for mathematics. Below we discuss how these items were used to assess the effects of content coverage on student achievement.

For lower grades reading, we developed a set of measures intended to reflect students' exposure to a *balanced* reading curriculum. Such a curriculum, we reasoned, would include attention to three broad curricular dimensions — word analysis, reading comprehension, and writing. We measured students' exposure to word analysis through a single item in which the teacher reported the amount of emphasis placed on this topic. We measured students' exposure to reading comprehension instruction by combining eight items into a single Rasch scale, where the items were ordered according to the cognitive demand of instruction in this area. In the scale, items ranged in order from the lowest cognitive demand to the highest cognitive demand as follows: identify main ideas, identify sequence of events, comprehend facts and

details, predict events, draw inferences, understand author's intent, differentiate fact from opinion, and compare and contrast reading assignments. The scale had a person reliability (for teachers) of .73.[25] A third measure was a single item in which teachers reported the emphasis they placed on the writing process. In assessing the effects of these variables on growth in students' reading achievement, we simply expanded the HLM growth models for the early grades cohort used in previous analyses. In the analyses, each of the curriculum coverage variables had a positive and statistically significant effect on students' growth in reading. The effect of a teachers' emphasis on word analysis skills was d = .10. The effect of the reading comprehension measure was d = .17. The effect of a teacher's emphasis on the writing process was d = .18.[26]

For mathematics, we used a single, multi-item scale measuring content coverage. Data for this measure were available for both cohorts of students in the *Prospects* data. For both cohorts, the measure can be thought of as indexing the *difficulty* of the mathematics content covered in a classroom, where this is assessed using an equal-interval Rasch scale in which the order of difficulty for items (from easiest to most difficult) was: whole numbers/whole number operations, problem solving, measurement and/or tables, geometry, common fractions and/or percent, ratio and proportions, probability and statistics, and algebra (formulas and equations). In both scales, a higher score indicated that a student was exposed to more difficult content. For the early elementary cohort, the scale had a person reliability (for teachers) of .77; in the upper elementary sample, the person reliability (for teachers) was .80. Once again, this measure was simply added as an independent variable into the HLM growth models used in earlier analyses. When this was done, the effect of content coverage on early elementary students' growth in mathematics achievement was *not* statistically significant. However, there was a statistically significant relationship for students in the upper elementary grades, the effect size being d = .09.[26]

Discussion of Content Covered

In general, the d-type effect sizes reported for the association of content coverage measures and growth in student achievement are about the same size as d-type effect sizes for the other variables measured here. This should give pause to those who view opportunity-to-learn as the *main* explanation for student-to-student differences in achievement growth. In fact, in one of our analyses (lower grades mathematics), the opportunity-to-learn variable had no statistical effect on student achievement.[28]

Moreover, the positive effects of curriculum coverage should be interpreted with caution for two reasons. One problem lies in assuming that opportunity-to-learn is "causally prior" to growth in student achievement and is therefore a causal agent, for it is very possible that instead, a student's exposure to more demanding academic content is endogenous — that is, results from that student's achievement rather than causing it. To the extent that this is true, we have overestimated curriculum coverage effects.[27] On the other hand, if curriculum coverage is relatively independent of past achievement, as some preliminary results in Raudenbush, Hong, and Rowan (2002) suggest, then our measurement procedures could be leading us to underestimate its effects on student achievement. This is because the measures of curriculum coverage used in

our analyses are very course-grained in their descriptions of instructional content, and because teachers are expected to accurately recall their content coverage patterns across an entire year in responding to a one-shot questionnaire. Once again, the findings just discussed seem plagued by unreliability in measurement, and in this light, it is somewhat remarkable that crude measures of the sort developed for the *Prospects* study show any relationship at all to achievement growth.

Context Variables

As a final step in our analysis of instructional effects on student achievement, we examined the extent to which the relationships of presage and process variables to student achievement were stable for different kinds of students. This analysis was motivated by data from the random effects models estimated in the first section of this report, which showed that the same classroom could have different effects on growth in achievement for students from different social backgrounds. In this section, we have shifted from estimating random effects models to estimating mixed models in which instructional effects are "fixed"; that is, assumed to have the same effects in all classrooms for students from all social backgrounds. In this section, we relax this assumption in order to examine interactions among presage and process variables and student background.

The HLM statistical package being used here allows researchers to examine whether presage and process variables have the same effects on growth in achievement for students from different social backgrounds, but it can do so only when there are sufficient data. In the analyses conducted here, for example, students' achievement is measured only at three or (in the best case) four time points. With this few number of time points, the program has insufficient data to estimate the extremely complex models that would be required to test for interactions among social background and instructional process variables. But there are some ways around this problem.[28] In addition, if one proceeds with such an analysis, as we did for exploratory purposes, interactions *can* be found. For example, in an exploratory analysis, we specified a statistical model for growth in early reading achievement in which we assumed that the effects of the instructional variables discussed earlier would be conditioned by students' gender, SES, or minority status. In the analysis, we found some evidence for the kinds of interactions being modeled, but it was far from consistent. For example, the data suggested that whole-class instruction was more effective for males, and less effective for higher SES students. The analysis also suggested that teachers' emphasis on the writing process was more effective for males, and that teacher experience was less effective for minority students. Thus, one can find evidence that the effectiveness of particular teaching practices varies for different groups of pupils.

But there are problems with this kind of analysis that extend far beyond the fact that there are insufficient data for such an analysis in *Prospects*. Equally important, there is little strong theory to use when formulating and testing such hypotheses. Thus, while research on teaching suggests that the effects of instructional variables can vary across different groups of pupils, it provides little guidance about what — exactly — we should predict in this regard. Consider, for example, the findings just discussed. What instructional theory predicts that the

effect of whole-class teaching is more effective for males than females, or for lower SES rather than higher SES students? More importantly, while it would be possible to formulate an elaborate *post hoc* explanation for why more experienced teachers appear to be less effective in promoting early reading growth among minority students (e.g., cohort differences in teacher training or in attitudes might explain the finding), should we interpret this finding knowing that it occurs in the context of several other findings that are completely unpredicted by any theory? We would argue that we should not, and that we keep our statistical analyses simple, at least until theory catches up with our power to analyze data statistically.

The main point about context effects, then, is that educational researchers have a long way to go in modeling context effects, both in terms of having the requisite data available for modeling complex, multilevel statistical interactions, or in having the kinds of theories that would make attempts to do so justifiable. As a result, we recommend that large-scale research on teaching limit itself for now to an examination of fixed effects models, where theoretical predictions are stronger and more straightforward.

Summary

The analyses in this section illustrate that large-scale research *can* be used to examine hypotheses drawn from research on teaching. The results also suggest that such hypotheses can be used to at least partially explain why some classrooms are more instructionally effective than others. The analyses presented in this report, for example, showed that classroom-to-classroom differences in instructional effectiveness in early grades reading achievement, and in mathematics achievement (at all grades) could be explained by differences in presage and product variables commonly examined in research on teaching. In the analyses, several variables had *d*-type effect sizes in the range of .10 to .20, including teacher experience, the use of whole-class instruction, and patterns of curriculum coverage in which students were exposed to a balanced reading curriculum and to more challenging mathematics.

At the same time, these results suggest that we probably should not expect a single instructional variable to explain the classroom-to-classroom differences in instructional effectiveness found in the first section. Instead, the evidence presented here suggests that many small instructional effects would have to be combined to produce classroom-to-classroom differences in instructional outcomes of the magnitude found in the first section. At the same time, the distribution of classroom effectiveness within the same school suggests that very few classrooms in the same school present an optimal *combination* of desirable instructional conditions. Instead, the majority of classrooms probably present students with a mix of more *and* less instructionally effective practices simultaneously. This scenario is made all the more plausible by what we know about the organization and management of instruction in the typical U.S. school. Research demonstrates that U.S. teachers have a great deal of instructional autonomy within their classrooms, producing wide variation in instructional practices within the same school. Variations in instructional practices, in turn, produce the distribution of classroom effects that we discovered in our variance decomposition models, with

a lack of real coordination across classrooms probably accounting for students' movement through more and less effective classrooms over the course of their careers in a given school.

If there is a "magic bullet" to be found in improving instructional effectiveness in U.S. schools, it probably lies in finding situations in which many instructionally desirable conditions co-exist in classrooms and in situations where students experience such powerful combinations of instructional practice across their careers in school. In fact, this is one reason we and our colleagues have become so interested in studying instructional *interventions*. By design, these interventions seek to smooth out classroom-to-classroom differences in instructional conditions, and to encourage the implementation of instructional conditions that combine to produce fairly powerful effects on student learning across all classrooms within a school. This insight suggests a real limitation to research on teaching that looks exclusively at *natural variations* in instructional practice, as did the research presented in this report (and as much other large-scale, survey research tends to do). If we look only at natural variation, we will find some teachers who work in ways that combine many desirable instructional conditions within their classrooms and others who don't. But if we rely solely on a strategy of looking at naturally occurring variation to identify "best" practice, we have no way of knowing if the "best" cases represent a truly optimal combination of instructional conditions or whether even the best classrooms are operating below the real (and obtainable) production frontier for schooling. In our view, it would be better to shift away from the study of naturally occurring variation in research on teaching and to instead compare alternative instructional interventions that have been designed — a priori — to implement powerful *combinations* of instructionally desirable conditions across classrooms in a school. In this case, we would no longer be studying potentially idiosyncratic variations in teacher effectiveness, but rather the effects of well-thought-out instructional designs on student learning.[29]

How to Improve Large-Scale, Survey Research on Teaching

The discussions presented in this report show how large-scale, survey research has been used to estimate classroom-to-classroom differences in instructional effectiveness and to test hypotheses that explain these differences by reference to presage, process, and context variables commonly used in research on teaching. Throughout this report, however, we have pointed out various conceptual and methodological issues that have clouded interpretations of the findings from prior research on teaching or threatened its validity. In this section, we review these issues and discuss some steps that can be taken to improve large-scale research on teaching.

"Effect Sizes" in Research on Teaching

One issue that has clouded research on teaching is the question of: "how big" are instructional effects on student achievement? As we tried to show in previous sections, the answer one gives to the question of how much of the variance in student achievement outcomes is accounted for by students' locations in particular classrooms depends in large

part on how the criterion outcome in an analysis of this problem is conceived and measured. Research that uses achievement *status* as the criterion variable in assessing teacher effects is looking at how much a single year of instruction (or exposure to a particular instructional condition during a single year) affects students' cumulative learning over many years. Obviously, the size of the instructional effect that one obtains here will differ from what would be obtained if the criterion variable assessed instructional effects on *changes* in student achievement over a single year. In fact, in analyses of achievement status, home background variables and prior student achievement will account for larger proportions of variance than variables indexing a single year of teaching. That said, it is worth noting that analyses using covariate adjustment models to assess instructional effects on students' achievement status *can* identify both the random effects of classroom placement on students' achievement and the effects of specific instructional variables. However, the effect sizes resulting from such analyses will be relatively small for obvious reasons.

A shift to the analysis of instructional effects on *growth* in achievement presents different problems, especially if gain scores are used to measure students' rates of academic growth. To the extent that the gain scores used in analysis are unreliable, estimates of the overall magnitude of instructional effects on student achievement will be biased downward. As the literature on assessing change suggests, it is preferable to begin any analysis of instructional effects by first estimating students' "true" rates of academic growth and then assessing teacher effects on growth within this framework. Unfortunately, computing packages that allow for such analyses are not yet commercially available, although preliminary results obtained while working with a developmental version of such a program (being developed by Steve Raudenbush) suggests that effect size estimates from such models will be very different from those obtained using covariate adjustment and gains models.

All of this suggests that there might be more smoke than fire in discussions of the relative magnitude of instructional effects on student achievement. Certainly, the discussion to this point suggests that "all effect sizes are *not* created equally." In fact, the same instructional conditions can be argued to have large or small effects simply on the basis of the analytic framework used to assess the effects (i.e., a covariate adjustment model, a gains model, or an explicit growth model). Thus, while there is much to be said in favor of recent discussions in educational research about the over-reliance on statistical significance testing as the single metric by which to judge the relative magnitude of effects — especially in large-scale, survey research, where large numbers of subjects almost always assure that very tiny effects can be statistically significant — the discussion presented in this report also suggests that substantively important instructional effects can indeed have very small effect sizes when particular analytic frameworks are used in a study. Moreover, when this is the case, large sample sizes and statistical significance testing turns out to be an advantage, for it works against having insufficient statistical power to identify effects that are substantively important when the dependent variable is measured differently. In particular, to the extent that researchers are using covariate adjustment or gains models to assess instructional effects, large sample sizes and statistical significance tests would

seem to be an important means for locating substantively meaningful effects, especially since these models present analytic situations in which the decks are stacked against finding large effect sizes.[30]

A final point can be made about efforts to estimate the magnitude of teacher effects on student achievement. In our view, the time has come to move beyond variance decomposition models that estimate the random effects of schools and classrooms on student achievement. These analyses treat the classroom as a "black box," and while they can be useful in identifying more and less effective classrooms, and in telling us how much of a difference natural variation in classroom effectiveness can make to students' achievement, variance decomposition models do not tell us *why* some classrooms are more effective than others, nor do they give us a very good picture of the potential improvements in student achievement that might be produced if we combined particularly effective instructional conditions into powerful instructional programs. For this reason, we would argue that future large-scale research on teaching move to directly measuring instructional conditions inside classrooms and/or to assessing the implementation and effectiveness of deliberately designed instructional interventions.

The Measurement of Instruction

As the goal of large-scale, survey research on teaching shifts from estimating the random effects of classrooms on student achievement to explaining why some classrooms are more instructionally effective than others, problems of measurement in survey research will come to the fore. As we discussed earlier, there is a pervasive tendency in large-scale, survey research to use proxy variables to measure important dimensions of teaching expertise, as well as an almost exclusive reliance on one-shot questionnaires to crudely measure instructional process variables. While the findings presented here suggest that crude measures of this sort *can* be used to test hypotheses from research on teaching, and that crude measures often show statistically significant relationships to student achievement, it is also true that problems of measurement validity and reliability loom large in such analyses.

What can be done about these problems? One line of work would involve further studies of survey data quality — that is, the use of a variety of techniques to investigate the validity and reliability of commonly used survey measures of instruction. There are many treatments of survey data quality in the broader social science literature (Biemer et al., 1991; Groves, 1987, 1989; Krosnick, 1999; Scherpenzeel & Saris, 1997; Sudman & Bradburn, 1982; Sudman, Bradburn, & Schwarz, 1996), and a burgeoning literature on the quality of survey measures of instruction in educational research (Brewer & Stasz, 1996; Burstein et al., 1995; Calfee & Calfee, 1976; Camburn, Correnti, & Taylor, 2000, 2001; Chaney, 1994; Elias, Hare, & Wheeler, 1976; Fetters, Stowe, & Owings, 1984; Lambert & Hartsough, 1976; Leighton et al. 1995; Mayer, 1999; Mullens, 1995; Mullens et al., 1999; Mullens & Kasprzyk, 1996, 1999; Porter et al., 1993; Salvucci et al., 1997; Shavelson & Dempsey-Atwood, 1976; Shavelson, Webb, & Burstein, 1986; Smithson & Porter, 1994; Whittington, 1998). A general conclusion from all of this work seems to be that the survey measures of instruction used in educational research suffer from a variety of methodological and conceptual

problems that can only be addressed by more careful work during the survey development stage.

The work that we are doing with colleagues to address these problems deserves brief mention here. As we discussed at an earlier point in this report, we have become keenly interested in assessing the effects of teachers' pedagogical content knowledge on students' achievement, but rather than rely on the kinds of indirect "proxy" measures that typify much previous research in this area, we have instead begun a program of research designed to build direct measures of this construct from scratch. To date, we have completed one round of pre-testing in which we have found that it is possible to develop highly reliable measures of teachers' content and pedagogical knowledge in very specific domains of the school curriculum using as few as six-to-eight items (Rowan, Schilling, Ball, & Miller, 2001). We also have begun to validate these measures by looking at "think aloud" protocols in which high- and low-scoring teachers on our scales talk about how and why they answered particular items as they did. Finally, in the near future, we will begin to correlate these measures to other indicators of teachers' knowledge and to growth in student achievement. The work here has been intensive (and costly). But it is the kind of work that is required if survey research on instruction is to move forward in its examination of the role of teaching expertise in instructional practice.[31]

We also have been exploring the use of instructional logs to collect survey data on instructional practices in schools. In the broader social science research community, logs and diaries have been used to produce more accurate responses from survey respondents about the frequency of activities conducted on a daily basis. The advantage of logs and diaries over one-shot questionnaires is that logs and diaries are completed frequently (usually on a daily basis) and thus avoid the problems of memory loss and mis-estimation that plague survey responses about behavior gathered from one-shot surveys. Here, too, we have engaged in an extensive development phase. In spring 2000, we asked teachers to complete daily logs for a 30-60 day time period, and during this time, we conducted independent observations of classrooms where logging occurred, conducted "think alouds" with teachers after they completed their logs, and administered separate questionnaires to teachers designed to measure the same constructs being measured by the logs. To date, we have found that teachers will complete daily logs over an extended period of time (if given sufficient incentives); that due to variation in daily instructional practice, roughly 15-20 observations are needed to derive reliable measures of instructional processes from log data; that log and one-shot survey measures of the same instructional constructs often are only moderately correlated; and that rates of agreement among teachers and observers completing logs on the same lesson vary depending on the construct being measured.[32] In future work, we will be correlating log-derived measures with student achievement and comparing the relative performance of measures of the same instructional construct derived from logs and from our own one-shot questionnaire.

The point of all this work is not to trumpet the superiority of our measures over those used in other studies. Rather, we are attempting to take seriously the task of improving survey-based measures of instruction so that we can better test

hypotheses derived from research on teaching. Without such careful work, estimates about "what works" in terms of instructional improvement, and "how big" the effects of particular instructional practices are on student achievement will continue to be plagued by issues of reliability and validity that currently raise doubts about the contributions of past survey research to broader investigations of teaching and its consequences for student achievement.

Problems of Causal Inference in Survey Research

If the goal of survey research is to test hypotheses about the effects of teachers and their teaching on student achievement, then more is needed in addition to appropriate interpretation of differing effect size metrics and careful development of valid and reliable survey instruments. To achieve the fundamental goal of assessing the effects of teachers and their teaching on students' achievement, researchers must also pay attention to problems of causal inference in educational research. That large-scale survey research confronts tricky problems of causal inferences in this area is demonstrated by some of the results we reported earlier in this report. Consider, for example, the findings we reported about the effects of teacher qualifications and students' exposure to advanced curricula on students' achievement. A major problem in assessing the effects of these variables on student achievement is that students who have access to differently qualified teachers or to more and less advanced curricula are also likely to differ in many other ways that also predict achievement. These other factors are confounding variables that greatly complicate causal inference, especially in non-experimental settings.

For several decades, educational researchers assumed that multiple regression techniques could resolve most of these problems of causal inference. But this is not always the case. For example, some analysts have noted that strategies of statistical control work effectively to reduce problems of causal inference only under limited circumstances. These include circumstances where all confounding variables are measured without error and included in a regression model, when two-way and higher-order interactions between confounding variables and the causal variable of interest are absent or specified in a model, when confounding variables are not also an outcome in the model, and when confounding variables have the same linear association with the outcome that was specified by the multiple regression model (Cohen, Ball, & Raudenbush, in press). Other researchers have taken to using instrumental variables and two-stage least squares procedures to simulate the random assignment of experiments, or they have employed complex selection models to try and control for confounding influences across treatment groups formed by non-random assignment, or they have advocated for "interrupted time series" analyses in which data on outcomes are collected at multiple time points before and after exposure to some "treatment" of interest. All of these approaches are useful, but they also can be difficult to employ successfully, especially in research on teaching, where knowledge of confounding factors is limited and where at least one of the main confounding variables is also the outcome of interest (students' achievement levels). In fact, difficulties associated with effectively deploying alternatives to random assignment in non-experimental research might account for the finding that non-experimental data are less

efficient than experimental data in making causal inferences. For example, Lipsey and Wilson (1993) reported on 74 meta-analyses that included both experimental and non-experimental studies of psychological, educational, and/or behavioral treatment efficacy. Their analysis showed that *average* effect sizes for various causal hypotheses did not differ much between experiments and non-experimental studies, but that *variation in effect sizes* was much larger for the non-experimental studies. All of this suggests that the typical — non-experimental — survey study of instructional effects on student achievement probably builds knowledge more slowly, and more tenuously, than experimental research.

The argument we are making should not be considered an unambiguous call for experimental studies of teaching, however. While there is growing consensus among researchers in many disciplines — including economics, political science, and the applied health sciences fields — that experiments are the most desirable way to draw valid causal inferences, it is the case that educational experiments will suffer from a number of shortcomings, especially when they are conducted in complex field settings, over long periods of time, where treatments are difficult to implement, where attrition is pervasive, where initial randomization is compromised, where crossover effects frequently occur, and where complex organizations (like schools) are the units of treatment. Much has been learned about how to minimize these problems in experimental studies (e.g., Boruch, 1997), but in the real world of educational research, complex and larger-scale experiments seldom generate unassailable causal inferences. Thus, scrupulous attention to problems of causal inference seems warranted not only in non-experimental, but also in experimental, research.

Moreover, even when experiments (or various quasi-experiments that feature different treatment and/or control groups) are conducted, there is still an important role for survey research. While policymakers may be interested in the effects of "intent to treat" (i.e., mean differences in outcomes among those assigned to experimental and control groups), program developers are usually interested in testing their own theories of intervention. They, therefore, want to know whether the conditions they think should produce particular outcomes do indeed predict these outcomes. The usual "black box" experiment, which examines differences in outcomes across those who were and were not randomly assigned to the treatment — regardless of actual level of treatment — is fairly useless for this purpose. Instead, measures of treatment implementation and its effects on treatment outcomes are what program developers usually want to see. They recognize that treatments are implemented variably, and they want to know how — and to what effect — their treatments have been implemented. Thus, even in experimental studies of teaching effects on student achievement, there is an important need for careful measurement of instruction, and the larger the experiment, the more likely that surveys will be employed to gather the necessary data for such measures.

Conclusion

All of this analysis suggests that there is a continuing role for survey research in the study of instructional effects on student achievement. It also shows the critical interdependence among the three problems that must be confronted if survey research is to inform research on teaching. We cannot interpret the results of large-scale, survey research on teaching very sensibly if we do not have a clear understanding of what constitutes a *big* or *small* effect, but no matter what method we choose to develop effect size metrics, we will not have good information from survey research about these effects if we fail to pay attention to issues of measurement and causal inference. Without good measures, no amount of statistical or experimental sophistication will lead to valid inferences about instructional effects on student achievement, but even with good measures, sound causal inference procedures are required. The comments and illustrations presented in this report therefore suggest that while large-scale, survey research has an important role to play in research on teaching and in policy debates about "what works," survey researchers still have some steps to take if they want to improve their capacity to contribute to this important field of work.

References

Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., & Sudman, S. (1991). *Measurement errors in surveys*. New York: John Wiley & Sons.

Boruch, R. F. (1997). Randomized experiments for planning and evaluation: A practical guide. *Applied Social Research Methods Series*, *44*, 1-263.

Brewer, D. J., & Goldhaber, D. D. (2000). Improving longitudinal data on student achievement: Some lessons from recent research using NELS:88. In D. W. Grissmer & J. M. Ross (Eds.), *Analytic issues in the assessment of student achievement*. Washington, DC: U.S. Department of Education.

Brewer, D. J., & Stasz, C. (1996). Enhancing opportunity to learn measures in NCES data. In G. Hoachlander, J. E. Griffith, & J. H. Ralph (Eds.), *From data to information: New directions for the National Center for Education Statistics* (pp. 3-1 – 3-28) (NCES 96-901). Washington, DC: National Center for Education Statistics.

Brophy, J. E., & Good, T. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd edition). New York: Macmillan.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage Publications, Inc.

Bryk, A. S., Raudenbush, S. W., Cheong, Y. F., & Congdon, R. (2000). *HLM 5: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Burstein, L., McDonnell, L., Van Winkle, J., Ormseth, T., Mirocha, J., & Guiton, G. (1995). *Validating national curriculum indicators*. Santa Monica, CA: RAND.

Calfee, R., & Calfee, K. H. (1976). *Beginning teacher evaluation study: Phase II, 1973-74, final report: Volume III.2. Reading and mathematics observation system: Description and analysis of time expenditures.* Washington, DC: National Institute of Education. (ERIC Document Reproduction Service No. ED127367).

Camburn, E., Correnti, R., & Taylor, J. (2000, April). *Using qualitative techniques to assess the validity of teachers' responses to survey items*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Camburn, E., Correnti, R., & Taylor, J. (2001, April). *Examining differences in teachers' and researchers' understanding of an instructional log*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Chaney, B. (1994). *The accuracy of teachers' self-reports on their postsecondary education: Teacher transcript study, Schools and Staffing Survey* (NCES 94-04). Washington, DC: National Center for Education Statistics.

Chiang, F. S. (1996). *Teachers' ability, motivation, and teaching effectiveness*. Unpublished doctoral dissertation, University of Michigan, Ann Arbor.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (in press). Resources, instruction, and research. In R. F. Boruch & F. W. Mosteller (Eds.), *Evidence matters: Randomized trials in educational research*. Washington, DC: Brookings Institution.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., et al. (1966). *Equality of educational opportunity*. Washington, DC: National Center for Education Statistics.

Cooley, W. W., & Leinhardt, G. (1980). The instructional dimensions study. *Educational Evaluation and Policy Analysis*, *2*(1), 7-25.

Darling-Hammond, L., Wise, A. E., & Klein, S. P. (1995). *A license to teach: Building a profession for 21st-century schools*. San Francisco: Westview Press.

Deng, Z. (1995). *Estimating the reliability of the teacher questionnaire used in the teacher education and learning to teach (TELT) study*. East Lansing, MI: National Center for Research on Teacher Learning, Michigan State University. (ERIC Document Reproduction Service No. ED392750).

Dunkin, M., & Biddle, B. (1974). *The study of teaching*. New York: Holt, Rhinehart, & Winston.

Elias, P. J., Hare, G., & Wheeler, P. (1976). *Beginning teacher evaluation study: Phase II, 1973-74, final report: Volume V.5. The reports of teachers about their mathematics and reading instructional activities.* Washington, DC: National Institute of Education. (ERIC Document Reproduction Service No. ED127374).

Ferguson, R. F., & Brown, J. (2000). Certification test scores, teacher quality, and student achievement. In D. W. Grissmer & J. M. Ross (Eds.), *Analytic issues in the assessment of student achievement*. Washington, DC: U.S. Department of Education.

Fetters, W., Stowe, P., & Owings, J. (1984). *High school and beyond, a national longitudinal study for the 1980s: Quality of responses of high school students to questionnaire items* (NCES 84-216). Washington, DC: National Center for Education Statistics.

Gage, N. L., & Needels, M. C. (1989). Process-product research on teaching: A review of criticisms. *Elementary School Journal*, *89*, 253-300.

Gamoran, A., Porter, A.C., Smithson, J., & White, P. (1997). Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. *Educational Evaluation and Policy Analysis*, *19*(4) 325-338.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research, 66*(3), 361-396.

Groves, R. M. (1987). Research on survey data quality. *Public Opinion Quarterly*, *51*(2), S156-172.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: John Wiley & Sons.

Karweit, N. (1985). Should we lengthen the school term? *Educational Researcher*, *14* (6), 9-15.

Kennedy, M. (1993). *A guide to the measures used in the Teacher Education and Learning to Teach study*. East Lansing, MI: National Center for Research on Teacher Education, Michigan State University.

Kerckhoff, A. C. (1983). *Diverging pathways: Social structure and career deflections.* New York: Cambridge University Press.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537-567.

Lambert, N., & Hartsough, C. (1976). *Beginning teacher evaluation study: Phase II, 1973-74, final report: Volume III.1. APPLE observation variables and their relationship to reading and mathematics achievement*. Washington, DC: National Institute of Education. (ERIC Document Reproduction Service No. ED127366).

Leighton, M., Mullens, J., Turnbull, B., Weiner, L., & Williams, A. (1995). *Measuring instruction, curriculum content, and instructional resources: The status of recent work* (NCES 95-11). Washington, DC: National Center for Education Statistics.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*(12), 1181-1209.

Mayer, D. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis*, *21*(1), 29-45.

Monk, D. H. (1994). Subject area preparation of secondary mathematics and science teachers and student achievement. *Economics of Education Review, 13*(2), 125-145.

Mullens, J. (1995). *Classroom instructional processes: A review of existing measurement approaches and their applicability for the teacher follow-up survey* (NCES 95-15). Washington, DC: National Center for Education Statistics.

Mullens, J., & Gayler, K. (1999). *Measuring classroom instructional processes: Using survey and case study field test results to improve item construction* (NCES 1999-08). Washington, DC: National Center for Education Statistics.

Mullens, J., & Kasprzyk, D. (1996). Using qualitative methods to validate quantitative survey instruments. In *1996 Proceedings of the Section on Survey Research Methods* (pp. 638-643). Alexandria, VA: American Statistical Association.

Mullens, J., & Kasprzyk, D. (1999). *Validating item responses on self-report teacher surveys*. Washington, DC: U.S. Department of Education.

Porter, A. C., Kirst, M., Osthoff, E., Smithson, J., & Schneider, S. (1993). *Reform up close: An analysis of high school mathematics and science classrooms*. Madison, WI: Consortium for Policy Research in Education, University of Wisconsin-Madison

Raudenbush, S. W. (1995). Hierarchical linear models to study the effects of social context on development. In J. M. Gottman (Ed.), *The analysis of change*. Mahwah, NJ: Lawrence Erlbaum Associates.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Raudenbush, S. W., Hong, G. L., & Rowan, B. (2002). *Studying the causal effects with application to primary school mathematics*. Ann Arbor, MI: Consortium for Policy Research in Education, University of Michigan, Ann Arbor. Unpublished Manuscript.

Rogosa, D. (1995). Myths and methods: Myths about longitudinal research plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change*. Mahwah, NJ: Lawrence Erlbaum Associates.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis*. New York: Russell Sage Foundation.

Rowan, B. (1999). The task characteristics of teaching: Implications for the organizational design of schools. In C. N. Hedley, R. Bernhardt , G. Cattaro, & V. Svplopoulos (Eds.), *Curriculum leadership: Rethinking schools for the 21st century*. Cresskill, NY: Hampton Press.

Rowan, B., Chiang, F. S., & Miller, R. J. (1997). Using research on employees' performance to study the effects of teachers on students' achievement. *Sociology of Education*, *70(4)*, 256-284.

Rowan, B., Schilling, S., Ball, D. L., & Miller, R. (2001). *Measuring teachers' pedagogical content knowledge in surveys: An exploratory study.* Ann Arbor, MI: Consortium for Policy Research in Education, University of Michigan, Ann Arbor. Unpublished manuscript.

Salvucci, S., Walter, E., Conley, V., Fink, S., & Mehrdad, S. (1997). *Measurement error studies at the National Center for Education Statistics* (NCES 97-464). Washington, DC: National Center for Education Statistics.

Sanders, W. (1998). Value-added assessment. *The School Administrator*, *15*(11), 24-32.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, *8*, 299-311.

Scheerens, J., & Bosker, R. (1997). *The foundations of educational effectiveness*. New York: Pergamon.

Scherpenzeel, A., & Saris, W. E. (1997). The validity and reliability of survey questions: A meta-analysis of MTMM studies. *Sociological Methods & Research*, *25*(3), 341-383.

Shavelson, R. J., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research*, *46*, 553-611.

Shavelson, R. J., Webb, N. M, & Burstein, L. (1986). Measurement of teaching. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd edition). Washington, DC: American Educational Research Association.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4-14.

Smithson, J. L., & Porter, A. C. (1994). *Measuring classroom practice: Lessons learned from efforts to describe the enacted curriculum — The reform up close study*. New Brunswick, NJ: Consortium for Policy Research in Education, Rutgers University.

Stedman, L. C. (1997). International achievement differences: An assessment of a new perspective. *Educational Researcher*, *26(*3), 4-15.

Stodolsky, S. S. (1988). *The subject matters: Classroom activity in math and social studies*. Chicago: University of Chicago Press.

Stoolmiller, M., & Bank, L. (1995). Autoregressive effects in structural equation models: We see some problems. In J. M. Gottman (Ed.), *The analysis of change*. Mahwah, NJ: Lawrence Erlbaum Associates.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions*. San Francisco: Jossey-Bass.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996*). Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement*, *58*(1), 21-37.

Endnotes

[1] Our interest in this problem derives from current work on a large-scale, survey study of instruction and student achievement in elementary schools. This project is being conducted by Deborah Ball, David K. Cohen, and Brian Rowan under the auspices of the Consortium for Policy Research in Education. Known as the *Study of Instructional Improvement*, the project is investigating the design, implementation, and effects on student learning of three of the most widely disseminated comprehensive school reform programs in the United States (the Accelerated Schools Program, America's Choice, and Success for All). As part of this project, we have worked with other researchers to develop a variety of innovative survey research instruments to study teacher effects on student achievement in over 100 elementary schools across the United States. The research we are reporting on here, which used *Prospects* data, was conducted in preparation for this study. *Prospects* data were used to "test out" various analytic models that might be used in our research, and to investigate various survey measures of teaching. Readers interested in learning more about the *Study of Instructional Improvement* can consult the project's website at www.sii.soe.umich.edu.

[2] Of course, variance decomposition models do not unambiguously demonstrate "causal" effects. However, they are useful as a first step in detecting the potential effects some factor might have on an outcome in question. So, in the first section of this report, where we estimate a variety of variance decomposition models, the reader is duly cautioned about the ambiguities of causal inference associated with this approach to estimating the overall magnitude of teacher "effects" on student achievement.

[3] An effect size can be calculated from a three-level, hierarchical, random effects model as: d = [√(variance in achievement lying among classrooms)/√(total student + classroom + school variance in student achievement)]. Effect size metrics in what Rosenthal (1994) calls the *d*-type family of effect sizes are designed to express differences in outcomes across two groups (e.g., an experimental and control group) in terms of standard deviations of the outcome variable. In the current analysis, we are analyzing data from more than two groups, however. In fact, in the "random effects" models estimated here, the variance components are calculated from data on all of the classrooms in a data set, with the assumption that all schools have equal variance among students and classrooms. In this case, we can develop a *d*-type effect size metric by comparing outcomes across two groups arbitrarily chosen from among the larger sample of classrooms. The two groups chosen for comparison here are classrooms within the same school that differ in their effects on student achievement by one standard deviation. Using this approach, the resulting "effect size" of .45 can be interpreted as showing the difference in achievement that would be found among two students from the same school if they were assigned to classrooms one standard deviation apart in effects on student achievement. For example, if the effect size is .45, we would conclude that two students from the same school assigned to classrooms a standard deviation apart in effectiveness would differ by .45 standard deviations in achievement.

[4] The student-level variables controlled for in these "value-added" analyses include prior achievement on the outcome variable; SES; gender; race; whether the student participated in special education, a gifted and talented program, or compensatory education; the student's age; the number of months between test administrations; the educational expectations of the student's parent(s); whether both parents live in the household; and the number of school-age siblings in the household. The school-level variables controlled for included the percentage of students at a school receiving free lunches, school enrollment, number of days a school was in session, and whether the school was located in an urban, suburban, or rural location.

[5] The effect size d in this case is: [√(adjusted variance among classrooms)/√(total adjusted

variance in achievement)], where the variance components have been adjusted through HLM regression analysis for the student background and school composition variables discussed in endnote 4.

[6] For this reason, Sanders and colleagues (e.g., Sanders & Horn, 1994) have used their own statistical computing package and "mixed model" methodology to perform variance decompositions with a similar aim.

[7] A detailed discussion of cross-classified random effects models can be found in Raudenbush and Bryk (2002, chap. 12). Like other HLM growth models discussed in this report, this model allows analysts to directly model individual growth curves for students in ways that separate "true" score variance in growth rates from "error" variance. As a result, as we discuss *d*-type effect sizes in the context of these models, we are able to ignore error variance, which improves our estimates of teacher effects over those derived from simple gains models. Another advantage of the "cross-classified" random effects model being discussed here is that it allows researchers to appropriately model the cross-nested nature of students passing through different classrooms in the same school over time.

[8] The *d*-type effect size here is: [√(variance in achievement growth lying among classrooms)/√(total school + class + student variance in achievement growth)]. The growth models estimated here were quadratic in form, but we "fixed" the non-linear term in this model. The main reason effect size coefficients are so much larger in the cross-classified random effects models than in the gain scores models is that the cross-classified random effects models provide a direct estimate of both student growth rates and "errors" of measurement, whereas gain scores models do not. As a result, effect size estimates based on gain scores include error variance, whereas explicit growth models do not include error variance. To see how the inclusion of measurement error affects "effect size" coefficients, we can use the variance components from the cross-classified random effects models to estimate the teacher effect size as [√(variance in achievement growth lying among classrooms)/√(total school + class + student + error variance in achievement growth)]. If we used this formula, we find effect sizes of .37 to .38 for reading and .32 to .45 for math, remarkably close to what we find using the gains models.

[9] The *d*-type effect size here is: [√(variance in achievement growth lying among schools)/√(total school + class + student variance in achievement growth)]. The growth models estimated here were quadratic in form, but we "fixed" the non-linear term in this model.

[10] The statistical computing package, HLM/3L version 5.25, calculates two kinds of residuals, ordinary least squares residuals and empirical Bayes residuals. For our purposes, the empirical Bayes residuals seem preferable, and it is these that are being correlated here. For a discussion of these different residuals, see Bryk and Raudenbush (1992, chap. 10).

[11] The careful reader might wonder whether the low correlations among residuals is produced by the unreliability of gain scores. This is probably *not* the case since the classroom-level residual scores being reported here are relatively free of this kind of measurement error. This is because variance due to the measurement errors that afflict gain scores is reflected in the within-class part of the model, but our residuals reflect variance among classrooms. As further evidence that this is the case, consider the residuals from a covariate adjustment model — where the dependent variable (achievement status) is measured very reliably. When the residuals from covariate adjustment models are correlated as in the examples above, the results are almost identical. Thus, the instability of residuals reported here does not appear to be due to the unreliability of gain scores as measures of growth in student achievement. For a further discussion of this issue, see Bryk and Raudenbush (1992, p. 123-129).

[12] The effect sizes quoted here come from Brewer and Goldhaber (2000, Table 1, p. 177).

The effect size we are using is what Rosenthal (1994) calls an *r*-type effect size. Effect sizes in the *r*-family are designed to express the strength of linear relationships among variables and are suitable for assessing effect sizes in models like linear regression which assume such relationships. Rosenthal's (1994) formula for deriving R^2 from the t-tests in a regression table is the one used here. The formula for deriving *r* (the correlation among two variables) from a t-test statistic is: $r = \sqrt{(t^2/(t^2\text{-df})}$. We simply square this to estimate R^2.

[13] These studies suffer from an important shortcoming, however — the strong possibility that selection effects are operating. In secondary schools especially, teachers with advanced degrees often teach the most advanced courses so that even after controlling for obvious differences among students enrolled in more- and less-advanced classes (e.g., their prior achievement, prior coursework, motivation, and home background), uncontrolled selection variables, rather than teachers' subject-matter training, could explain the results here.

[14] The reader is cautioned that the analyses conducted here did not use the cross-classified random effects hierarchical model discussed earlier and, as a result, do not take into account the nesting of pupils within classrooms across years. Additionally, the analyses reported here do not take into account possible complications in causal inference that arise when the kinds of teaching that students receive in a given year result from the kinds of instruction they received in previous years, or as a result of their prior achievement. The ways in which both the "cross-nesting" of students in different classrooms over time, and the endogeneity of instructional practices affect causal inference in research on teaching, and some newly developing strategies for coping with these problems, are discussed in Raudenbush, Hong, and Rowan (2002).

[15] The *d*-type effect size reported here is, in effect, a standardized regression coefficient. It expresses the difference among students in annual growth (expressed in terms of standard deviations in annual growth) that would be found among students whose teachers are one standard deviation apart in terms of experience. In this analysis, the standard deviation of teachers' experience is 8.8 years, the unstandardized regression coefficient for the effect of experience on achievement growth is .18, and the standard deviation in "true" rates of annual growth among students is 21.64. Thus, $d = [(8.8 * .18)/21.64]$.

[16] The effects sizes are calculated as in endnote 11.

[17] Ibid.

[18] This is the work of a team of researchers headed by Deborah Ball and Brian Rowan and including Sally Atkins-Burnett, Heather Hill, Robert Miller, P. David Pearson, Geoff Phelps, and Steve Schilling.

[19] The *r*-type effect size here is tiny, but it should be pointed out that it is based on a covariate adjustment model in which we are modeling students' achievement status (controlling for prior achievement and many other variables). Effect size metrics expressing the relationship of this measure to "true" rates of growth in student achievement might be much higher, as the analyses of teachers' certification status and degree attainment just above demonstrate.

[20] A report on this work can be found in Rowan, Schilling, Ball, and Miller (2001) and accessed at www.sii.soe.umich.edu.

[21] Relevant data were unavailable in the lower grades cohort.

[22] The effect sizes here are as in endnote 11.

[23] Ibid.

[24] In research on high schools, curriculum content is often indexed by course enrollment. For example, in earlier research, we used *NELS:88* data to assess the effects of mathematics content coverage on student achievement in high schools. In these data,

variations in the content covered by students were assessed at the course level. Even at this very broad level of analysis, however, the effects of content coverage on achievement are evident in the data. For example, in a well-specified covariate adjustment model controlling for students' home background, prior achievement, and motivation, we found that an additional course in mathematics during 9th and/or 10th grade results in a .13 standard deviation effect on students' achievement status in the *NELS:88* data (see Rowan, 1999). However, these findings could reflect selection bias, since course placement in high schools does not occur from random assignment.

[25] The benefit of a Rasch model is that it produces an equal interval scale that can be used with all teachers.

[26] The effect sizes here are as in endnote 11.

[27] This problem could be pervasive in non-experimental research on instructional effects, as Cohen, Raudenbush, and Ball (in press) discuss. As a result, Raudenbush, Hong, and Rowan (2002) are developing analytic procedures to take this problem into account in estimating instructional effects.

[28] For example, one can estimate interactions of the sort being discussed here without first testing the assumption that the effects of instructional variables are random. In such models, one is therefore treating the interactions under analysis as "fixed effects."

[29] One way to further illustrate this point is to compare the "effect sizes" from the random effects (i.e., variance decomposition) models discussed in this report with the effect sizes reported in experiments where the effects of deliberately designed teaching interventions are studied. Gage and Needels (1989), for example, reported the effect sizes for 13 field experiments designed to test the effects of interventions based on teacher behaviors found to be effective in process-product research. In these experiments, multiple instructional dimensions were altered through experimental manipulation. When these interventions worked, the experiments produced effect sizes ranging from .46 to 1.53. These effect sizes compare more than favorably to the kinds of effect sizes we reported from the random effects models estimated here, especially when one considers that the effect sizes reported in the intervention studies come from studies where achievement status and/or gains were used to calculate "effect size" metrics.

[30] In fact, one possible explanation for the "inconsistent" findings in prior process-product research might be that researchers using gains models or covariate adjustment models to assess instructional effects sometimes lacked sufficient statistical power to identify the effects of instructional variables on student achievement.

[31] Information on this work can be found at www.sii.soe.umich.edu.

[32] Information on this work can be found at www.sii.soe.umich.edu. Our results are similar to those reported in other studies of these same issues, especially Burstein et al. (1995) and Smithson and Porter (1994).