

4. Analysis of Large-Scale Datasets: Examples of NSF-Supported Research

NSF FUNDS A VARIETY OF STUDIES designed to investigate how to improve learning, especially in mathematics and science, including experiments, quasi-experiments, and secondary analyses of observational data.³³ Some of the studies focus on theory building, while others are evaluations where the researcher is interested in assessing the effectiveness of specific large-scale initiatives, such as systemic reform. We reviewed the research portfolio of NSF's Education and Human Resource Directorate and selected four quantitative studies that used different statistical techniques to investigate causal questions. These techniques serve as examples for investigators conducting secondary analyses of these and other large-scale datasets. All four studies address issues of causality. However, only the first two allow for causal inferences. We have included the other two studies because they can be used to generate causal hypotheses that can inform the design of experiments.

The first study, "How Large Are Teacher Effects?" examines teacher effects on student achievement based on data from the Tennessee Class Size Experiment, an experiment with random assignment (Nye, Konstantopoulos, & Hedges, 2004). Although the second study, "Effects of Kindergarten

Retention Policy on Children's Cognitive Growth in Reading and Mathematics," is based on observational data from the Early Childhood Longitudinal Study (ECLS), it approximates an experiment on the effects of kindergarten retention on children's literacy (Hong & Raudenbush, 2005). Using data from the Third International Mathematics and Science Study (TIMSS), the third study, *Why Schools Matter: A Cross-National Comparison of Curriculum and Learning*, employs multiple analytic approaches, including structural modeling, to uncover possible causal relationships between aspects of curriculum and achievement gains in mathematics and science (Schmidt et al., 2001). This study provides a strong foundation on which to construct an experiment with random assignment on the effects of curriculum on student learning. The last study, "The Role of Gender and Friendship in Advanced Course-Taking," uses standard regression techniques to examine the influence of friends on high school students' advanced course-taking in mathematics and science, using data from the National Longitudinal Study of Adolescent Health (Riegle-Crumb, Farkas, & Muller, 2006).

Case I. An Experiment With Random Assignment: "How Large Are Teacher Effects?"

To investigate teacher effects on student learning outcomes, Nye, Konstantopoulos, and Hedges (2004) conducted a secondary analysis of data from the Tennessee Class Size Experiment, or Project STAR (Student-Teacher Achievement Ratio), an experiment in which students and teachers were randomly assigned within each school to classes that varied in size and in the presence of a teacher aide (small classes, regular classes, and regular classes with a teacher aide). The original experiment was designed to compare the effects of class size and student-teacher ratios on student achievement (Finn & Achilles, 1990, 1999). In analyzing data from this experiment, Nye and

her colleagues take advantage of the study's large sample and use of random assignment to compare the learning outcomes of students in the same treatment condition who had different teachers. In contrast to most research on teacher effects, which tends to be based on observational data and relies on statistical controls to correct for selection bias, Nye et al.'s use of data from an experiment with random assignment of both students and teachers allows them to draw causal inferences about teacher effects on student achievement with a high degree of confidence.

Research Question and Theoretical Frame

Specifically, Nye and her colleagues use data from Project STAR to determine whether there are teacher effects on student achievement and to estimate the magnitude of these effects. If teacher effects are large, they argue, then identifying factors that contribute to teacher effectiveness would be important to both education researchers and reformers. If these effects are small, then finding ways to improve teacher effectiveness would be a less promising reform strategy.

Researchers have differed in their perspectives on what factors contribute to teacher effectiveness and whether differences in teacher quality have significant effects on student learning outcomes. Some have assumed that teacher and school characteristics such as teacher experience and education, class size, and school resources may affect the quality of teaching and in turn student achievement. Others argue that these measured characteristics have little effect on student learning, but acknowledge that there may be other observed or unobserved characteristics that have significant effects on student learning outcomes.

Problems With Studies of Teacher Effectiveness

Although a considerable number of studies have been conducted on teacher effects, results have been mixed. Some studies indicate that teacher effects are negligible; others suggest that characteristics such as teacher experience and education have significant effects on student achievement (for reviews of the literature, see Hanushek, 1986, and Greenwald, Hedges, & Laine, 1996). However, reviewers of these studies generally agree that it is difficult to draw causal inferences about the relationship between measures of teacher quality and student achievement because of the exclusion of potentially relevant variables such as teacher instructional practices. Nye and her colleagues are able to avoid these problems of selection bias by using data from an experiment with random assignment.

Nye et al. identify two traditions of research on teacher effectiveness and describe the limitations of each with respect to making causal inferences. One tradition, referred to as education production-function studies, examines the relationship between specific teacher or school characteristics (e.g., teacher experience, teacher education, class size) and student achievement. These school resource variables tend to be associated with student and family characteristics because parents typically choose the neighborhoods they live in (and the schools within them) based on particular preferences and resources. Although production-function studies attempt to adjust statistically for these associations by including student and family characteristics in the analyses, they fail to take into account the possible influences of unmeasured characteristics on student learning outcomes (e.g., instructional practices) and often include measures that may have no relation to student achievement (e.g., teacher salary). In many studies, determining the direction of causality between teacher effectiveness and student achievement is also problematic because the assignment of students to classes is often based on student and

teacher characteristics. For example, more experienced teachers may be assigned to classes of high-achieving students as a reward for seniority, making it difficult to determine whether learning outcomes are due to teacher effects or students' prior achievement.

Studies in the second tradition examine variations in student achievement across classrooms, adjusting for student background characteristics. These studies typically include a prior measure of student achievement, making the focus of the analysis variation in student achievement gains across classrooms. It is assumed that between-classroom variation in student achievement gains is caused by differences in teacher effectiveness. These analyses, however, may fail to include adequate adjustments for preexisting differences between students assigned to different classrooms (i.e., selection bias), including unobserved differences related to achievement growth (e.g., differences in the quality of instruction students received in prior years).

Experimental Design: Avoiding Problems of Selection Bias

The randomized assignment of students and teachers to treatment conditions, and to classrooms within treatment conditions, ensures that any differences between groups in participants' pretreatment characteristics occur only by chance. Because both teachers and students were randomly assigned to treatment conditions, Nye and colleagues can assume that, barring any difficulties in implementing the experiment, any significant differences in student achievement across treatment conditions can be attributed to either treatment effects (class size and the presence or absence of a teacher aide) or teacher effects; within classrooms of the same type (e.g., small), these differences can be attributed to teacher effects.

Even though experiments are designed to produce valid evidence of causal effects, they are not always implemented

with fidelity (e.g., students may move between treatment groups after being randomly assigned, or there may be differential attrition across treatment groups). Nye and her colleagues therefore investigated deviations from the study design and their potential effects on study outcomes. They also conducted tests to determine whether randomization had been effective in eliminating systematic differences between treatment groups. Although randomized assignment tends to produce treatment groups that are, on average, balanced with respect to pretreatment characteristics, within any single trial, randomization may result in groups that systematically differ with respect to certain pretreatment characteristics.

Study Design, Data, and Approach

The Tennessee Class Size Experiment was a 4-year longitudinal study, initially fielded in 1985, that was funded by the Tennessee legislature and conducted by the state department of education. More than 6,000 students from 79 schools and 42 school districts in Tennessee participated in the first year of study, and almost 12,000 students participated over the course of the 4-year experiment (Finn & Achilles, 1999). The policy issue addressed by the study was the effect of class size on student learning. Specifically, the study examined whether reducing the number of students in a single classroom, or reducing class size by having two adults in the classroom, improved students' mathematics and reading achievement more than "regular-sized" classes. Unlike previous studies of the effects of class size on student achievement, this study was a controlled experiment with random assignment.

Within each school, entering kindergarten students were randomly assigned to one of three types of classrooms: small classes (13–17 students), regular classes (22–26 students), or regular classes with a full-time teacher aide. Teachers were also randomly assigned to these three treatment conditions.

Student assignments by classroom type were maintained throughout the day and throughout the school year. Students who entered a school in first grade or in subsequent grades were randomly assigned to classroom type upon entry. As students in the experimental cohort progressed through subsequent grade levels, teachers at each grade level were randomly assigned to one of the three types of classrooms each year. Agreements were obtained from school districts to remain in the study for 4 years and to maintain the random assignment of students to classroom type from kindergarten through third grade.³⁴

In analyzing data from the Tennessee Class Size Experiment, Nye et al. focused on differences in the mathematics and reading achievement of students in different classrooms within the same treatment condition. Due to the large size of the dataset, the researchers were able to select a subset of schools that had at least two classrooms assigned to the same treatment condition.³⁵ In analyzing variations in math and reading achievement, they examined differences in both achievement status (e.g., achievement measured at a particular grade level) and achievement gains (e.g., achievement growth from one grade to the next).³⁶

Analyses and Results

In the original experiment, the fidelity with which the study was implemented was somewhat compromised. In a small number of cases there was overlap in the sizes of the classes categorized as large and small. In kindergarten and later grades, there was also a small amount of crossover of students between classroom types. There was some student attrition between kindergarten and third grade as well. Preliminary analyses were therefore conducted to investigate deviations from the study design; based on these analyses, the investigators concluded that none of the deviations invalidated the

results of the original experiment.³⁷ Additional checks were conducted to ensure that randomization had been effective in eliminating preexisting differences between students and teachers assigned to different types of classrooms. Results of these checks were consistent with successful randomization. For example, no differences were found in the SES, ethnicity, or age of students across treatment conditions. Analyses also revealed no systematic differences in these characteristics across classrooms within the same treatment condition within schools.³⁸

Because Nye et al.'s study focuses on variation in student achievement across teachers within the same treatment group (small class, regular class, regular class with teacher aide), a method of analysis was needed that took into account the clustering of students within classrooms, treatment groups, and schools. Hierarchical linear modeling (Bryk & Raudenbush, 2002) was therefore used in analyzing teacher effects on student achievement. This method of analysis allowed the investigators to examine between-classroom but within-school-and-treatment variation in reading and mathematics achievement. Within a school, systematic variation in student achievement between classes in the same treatment condition could be attributed to teacher effects.³⁹

To estimate teacher effects on student achievement, Nye and her colleagues developed separate analytic models to examine teacher effects on achievement gains and on achievement status. Separate models were also constructed for reading and mathematics achievement for each grade level. Results of these analyses showed that variations in student achievement gains between classrooms (and thus teachers) within the same treatment condition were significantly larger than variations in student achievement gains between schools, indicating that the teacher a student is assigned to may be more important for that student's achievement than the school the student attends. This pattern of results was similar for reading and mathematics

achievement and was consistent across grades, indicating that teachers had substantial effects on student learning growth from one year to the next.⁴⁰ Teacher effects were found to be much larger in mathematics than in reading, regardless of the grade attended.⁴¹ Nye et al. suggest that mathematics is more likely to be learned in school and thus to be influenced by teachers, whereas reading is often learned in contexts other than school; alternatively, there may be more variation (either in quantity or quality) of mathematics instruction than in reading instruction. Teacher effects on student achievement status were found to be similar in magnitude to those for achievement gains.⁴²

Additional analyses were conducted to determine whether teacher effects might be explained by differences in teacher experience or education and whether these effects varied with school or student SES. Results indicated that teacher experience and education explained very little of the variance in teacher effects (never more than 5%). However, teacher effects did vary significantly by school SES; there was more variation in teacher effects in low-SES schools than in high-SES schools. The proportion of total variance in student achievement accounted for by teacher effects was also higher in low-SES schools.⁴³ These findings suggest that teacher effects are much more uneven in low-SES versus high-SES schools. Thus, in low-SES schools, which teacher a student is assigned to has a greater impact on average classroom achievement than it does in high-SES schools. In analyzing the relationship between teacher effects and student SES, the investigators found that although teacher effects vary by student SES, this variation does not help to explain variation in teachers' effectiveness across schools.

Implications for Estimating Causal Effects

This study analyzes data from a randomized controlled experiment in which students and teachers within each school were randomly assigned to treatment conditions (small class, regular class, regular class with teacher aide). Because random assignment was used, all observed or unobserved differences in teacher and student characteristics across treatment conditions occur by chance alone, making it unnecessary to adjust for specific student or family characteristics or to specify in advance teacher characteristics that are related to student achievement. Checks of differences between treatment groups confirmed that randomization was effective in eliminating systematic differences in the pretreatment characteristics of students (and teachers) assigned to different treatment conditions. Differences in student achievement across treatment conditions could thus be attributed to treatment effects rather than to the pretreatment characteristics of students or teachers.

By focusing on schools in which different teachers were assigned to the same treatment condition, Nye and her colleagues were able to differentiate between treatment effects and teacher effects. Because random assignment was used, within any given school, systematic variation in achievement between classrooms within the same treatment condition could be attributed to teacher effects. The investigators were thus able to draw causal inferences about teacher effects on student achievement.

As Nye and her colleagues observe, their results suggest that

teacher effects are real and are of a magnitude that is consistent with that estimated in previous studies. However, we would argue that, because of random assignment of teachers and students to classrooms in this experiment, our results provide stronger evidence

about teacher effects. The results of this study support the idea that there are substantial differences among teachers in the ability to produce achievement gains in their students . . . [suggesting] that interventions to improve the effectiveness of teachers or identify effective teachers might be promising strategies for improving student achievement. (p. 253)

The authors acknowledge that “this design cannot identify the specific characteristics that are responsible for teacher effectiveness” (p. 239). Although both teacher education and teacher experience were examined, they explained virtually none of the variance in teacher effects. Because Nye et al. were analyzing data that were collected for a different purpose (i.e., to examine the relationship between class size and student achievement), their analysis was constrained by the available data on teacher characteristics.

That there are teacher effects on student achievement may seem obvious. However, demonstrating these effects empirically using data from an experimental study is an important contribution. We can be confident that there are substantial teacher effects and that they vary by school SES. These findings suggest that interventions to replace less qualified teachers or to improve teacher quality would be more promising in low-SES schools than in high-SES schools. Overall, the study addresses issues of data quality, provides stronger grounds on which to base policy decisions, and suggests strategies for designing future intervention studies. It also suggests possibilities for conducting analyses of data from experimental studies. As randomized controlled experiments become more common in education, data from these studies will provide additional opportunities for secondary analyses.

Case II. Approximating a Randomized Experiment: “Effects of Kindergarten Retention Policy on Children’s Cognitive Growth in Reading and Mathematics”

To investigate the causal effects of kindergarten retention policies on children’s cognitive growth in mathematics and reading, Hong and Raudenbush (2005) use observational data from the Early Childhood Longitudinal Study (ECLS-K), to approximate a randomized controlled experiment. Using propensity score matching, they construct treatment groups from this national dataset that are comparable with respect to students’ probabilities of being retained and balanced with respect to students’ pretreatment characteristics. A similar analysis is conducted at the school level to examine the effects of school retention policies (allowing or banning retention) on student learning outcomes.

Research Questions and Theoretical Frame

The purpose of this study is to determine whether kindergartners who were retained would have had higher growth rates in reading and mathematics if they had been promoted to first grade. In other words, if an experiment could be conducted in which kindergartners were randomly assigned to treatment groups (retention and promotion), would the growth trajectories of retained students differ significantly from those of promoted students? Similarly, if schools could be randomly assigned to policy conditions (allowing or banning retention), would the learning outcomes of students in retention schools differ significantly from those in nonretention schools?

Developmental psychologists differ in their perspectives on the potential benefits of kindergarten retention. Proponents of retention argue that children develop at different rates; kindergartners who have trouble keeping up academically may need additional time to mature socially and cognitively before

being entering first grade (Plummer & Graziano, 1987; Smith & Shepherd, 1988). This perspective suggests that kindergarten retention would have a positive effect on the learning outcomes of children who are retained because they would be given additional time to master concepts and skills that their classmates have already mastered. Children who are promoted may also benefit by being in classrooms with students who have similar levels of academic achievement instead of in classrooms that vary widely in achievement levels, assuming that retained students have substantially lower levels of achievement than those who are promoted (Byrnes, 1989; see also Smith & Shepard for a review). Since both retained and promoted students would potentially benefit from a policy of retention, the average learning growth of students in retention schools should be higher than that of students in nonretention schools.

Other developmental psychologists contend that having children repeat an unsuccessful learning experience is more likely to impede than enhance students' cognitive and social development (Morrison, Griffith, & Alberts, 1997). It has been argued that retention stigmatizes students, leading to lower parent, teacher, and self- expectations (Jackson, 1975; Shepard, 1989). Supporters of eliminating retention maintain that reforming instructional practices to correct children's learning difficulties may be more effective than retention in improving the learning outcomes of retained students (Karweit, 1992; Leinhardt, 1980; Reynolds, 1992; Tanner & Galis, 1997). At both the individual and school levels, retention is likely to have a negligible or negative effect on student learning.

Problems With Studies of Retention

Results of previous research on retention effects have been inconclusive. A large number of studies show a negative relationship between kindergarten retention and academic

achievement or personal/social development (see, e.g., Holmes, 1989; Nagaoka & Roderick, 2004). A similarly large number of studies show no statistically significant relationship between retention and these outcomes (see, e.g., Shepard, 1989; Jimerson, 2001). Such inconsistencies appear to be due in part to weaknesses in study design. One common problem with previous retention studies is that researchers have not considered whether the retained and promoted groups are comparable (i.e., at similar risk for retention). Failure to control for observed differences between groups may have led investigators to draw invalid inferences about retention effects.

Two primary strategies have been used in past retention research: (a) same-grade comparisons, and (b) same-age comparisons. Same-grade studies, which constitute the majority of retention studies, compare the outcomes of students who are repeating a grade with those of students who are completing that grade for the first time. In same-age studies, the outcomes of retained children are compared with those of children of the same age who were promoted to the next grade. Both strategies are problematic with regard to drawing valid causal inferences about the effects of retention on children's academic progress. In the case of same-grade studies, researchers are able to compare the academic standing of children who are retained with that of their classmates both before and after retention but are unable to make inferences about how retained children might have performed had they been promoted to the next grade. In same-age studies, the outcomes of retained students are often compared with those of all promoted students, including those who had virtually no chance of being retained. These low-risk students provide no information on which to base inferences about how retained students might have performed if promoted. Such studies also typically rely on statistical adjustments for a limited number of background variables to equate groups. But when the groups are barely comparable

with respect to their probability of being retained, this technique is unlikely to produce valid results.

Studies that restrict their comparisons of retained students to low-achieving students who have been promoted are more promising with respect to drawing causal inferences because these were students who were at risk for retention in the previous year and thus are more likely to be similar to the students who were actually retained. To adjust for any remaining differences between retained and promoted groups, however, researchers typically adjust for only a few background characteristics. Most do not adjust for prior learning growth rate, a variable that needs to be included if valid inferences are to be made about differences in the academic progress of the two groups. Most of these studies also assume that all background characteristics associated with retention have been included in the analysis, an assumption that typically is unwarranted.

Constructing a comparison group by matching students on background characteristics has the advantage of making differences between groups more readily apparent (e.g., the extent to which there is overlap between the groups with respect to the risk of being retained). However, most studies using this approach have been able to match on only a limited number of characteristics, raising questions about the initial equivalence of the matched groups. In addition, none of the studies have compared the academic achievement or social development of retained students with these outcomes for matched peers who were promoted.

Conducting a randomized controlled experiment designed to study retention effects would be problematic, as it is unlikely that parents would allow their child to be retained or promoted on a random basis (e.g., irrespective of their grades, ability, and social development). If it is not feasible to randomly assign students to treatment conditions (retention or promotion), how can a randomized experiment be approximated?

If the goal is to determine whether retained students would have performed better if they had been promoted, then the outcomes of retained students need to be compared to those of promoted students who have similar characteristics, including similar probabilities of being retained. Some mechanism is needed to construct comparison or treatment groups that are balanced with respect to background characteristics. Propensity score matching, the method used by Hong and Raudenbush, provides this mechanism.

Controlling for Selection Bias: Propensity Score Matching

Propensity score methods approximate randomized assignment to treatment conditions by ensuring that students have equivalent chances of being in the retained or in the promoted group. Because groups are comparable in terms of their pre-treatment characteristics, any differences in the learning outcomes of the two groups can be attributed to differences in treatment (retention versus promotion). As a result, the cognitive growth of promoted students can be interpreted as indicating how retained students might have performed if they had been promoted instead.

An advantage of propensity score methods is that they estimate each student's probability of being retained based on an aggregate of characteristics. Being able to summarize these characteristics in one composite measure (the propensity score) makes it possible for analysts to make a straightforward assessment of whether there is sufficient overlap between groups to justify comparison and to match students based on their propensity scores when there is sufficient overlap (Rosenbaum & Rubin, 1983; Rubin 1997). Propensity score matching is facilitated by using large-scale nationally representative datasets such as ECLS.⁴⁴

Propensity score matching adjusts for systematic differences in the characteristics of the observed groups in two

ways: (a) by eliminating students who have virtually no probability of being retained (high-achieving students, for example, may have almost no chance of being retained and offer no useful information on which to base estimates of the learning outcomes of retained and promoted students who are at similar risk for retention); and (b) by matching the remaining students in each group on the basis of characteristics known to be associated with retention.

To construct a propensity score model, in this case for kindergarten retention, the analyst first identifies variables that are systematically associated with retention using bivariate analysis. These variables are then included as predictors of kindergarten retention in multivariate regression models. Because many of the variables are associated with each other, only some of them will have a significant association with retention when all variables are included in the model, allowing the analyst to reduce the number of variables used in propensity models. These significant predictors are used to calculate propensity scores for students in the retained and promoted groups. By stratifying and matching students in each group on the basis of their propensity scores, analysts can identify students for whom there are no matches and exclude them from analysis. Students for whom matches are found will have similar characteristics.

Study Design, Data, and Approach

Hong and Raudenbush conduct three analyses. The first identifies the factors associated with student retention. The second estimates how retained students would have performed in reading and mathematics if they had instead been promoted. The third analysis estimates the effects of the school's retention policy (allowing or banning retention) on students' cognitive growth in reading and mathematics; this analysis was conducted at the level of the school rather than the student.

Thus school characteristics associated with the adoption of a retention policy were identified and used to calculate school propensity scores, using the same series of steps described for calculating student propensity scores.

Of the more than 20,000 first-time kindergartners included in the ECLS study, there are 13,520 for whom retention/promotion information is available. Information on kindergarten retention policies is also available for 1,221 schools in the study. Due to missing data on kindergarten retention policies, 1,667 students were excluded from the analyses.⁴⁵ After exclusions, the analysis sample consisted of 471 retained kindergartners and 10,255 promoted students in 1,080 retention schools, and 1,117 promoted students in 141 nonretention schools. For most students, mathematics and reading assessment data were obtained during the fall and spring of the kindergarten year and the spring of the following year.⁴⁶ The ECLS dataset also contains extensive information on the background characteristics of students obtained through parent, teacher, and school administrator surveys.

Analyses and Results

Using bivariate analysis, Hong and Raudenbush initially identified 207 student characteristics that were associated with retention based on prior research. When these variables were included in multivariate regression analyses, 39 of them were found to be significant predictors of retention. For example, children from single-parent families with several siblings, those whose parents had a lower commitment to parenting, and those who had lower scores on kindergarten assessments had a greater likelihood of being retained. Teacher perceptions were also found to be significant predictors of retention. Students who were retained were more likely to be placed in the lowest reading group in kindergarten (based on the teacher's

perception of the child's reading ability) and were rarely able to move into a higher reading group.⁴⁷

This set of predictors was used to calculate a propensity score for each student in the observed retention and promotion groups (i.e., each student's probability of being retained based on this combined set of characteristics). Propensity scores of students in the two groups were then examined to ensure that the distributions overlapped. No matches were found for 3,087 students in retention schools who had virtually no chance of being retained, and these students were excluded from the analyses. The remaining students were stratified and matched on the basis of their propensity scores. This process resulted in retention and promotion groups that were balanced with respect to students' background characteristics and their probabilities of being retained, thus approximating the random assignment of students to treatment conditions.

Because students' potential learning outcomes are likely to depend on treatment setting (the school), hierarchical linear modeling (HLM) was used to estimate the effects of retention on students' reading and mathematics achievement. HLM adjusts for similarities among students who attend the same school and allows the analyst to examine variation in retention effects at both the student and school levels. Results of this analysis indicated that, on average, retained students had significantly lower growth trajectories in reading and mathematics than promoted students who were at similar risk for retention. If a retained student had instead been promoted, his or her expected achievement would be approximately 9 points higher in reading and 6 points higher in math at the end of the treatment year. The magnitude of these estimated effects was about two-thirds of a standard deviation of the outcome in both reading and mathematics, a difference equivalent to approximately a half-year's learning growth in each subject area.⁴⁸

The HLM results also indicated that retention effects varied significantly across schools. The difference in the reading growth trajectories of retained and promoted students at similar risk of retention was greatest in schools with higher average reading achievement. In contrast, the difference in mathematics growth trajectories was greatest in schools with lower average mathematics achievement. The investigators suggest that curricular and instructional differences between kindergarten and first grade may account for these differences. In high-achieving schools, reading instruction typically occurs at a relatively fast pace. Students who are promoted may therefore learn at a faster rate than those who are retained. In low-achieving schools, the kindergarten curriculum often includes little mathematics content, providing retained students with fewer opportunities for learning growth relative to promoted students.⁴⁹

In contrast to randomized assignment, which tends to create groups that are balanced with respect to observed and unobserved characteristics, propensity score matching takes into account only the observed characteristics of group members. The investigators therefore conducted an additional analysis to check for the sensitivity of their results to the inclusion of adjustments for unmeasured characteristics.⁵⁰ They found that these adjustments did not significantly affect their estimates of retention effects, suggesting that their results were not biased due to the exclusion of unobserved characteristics.

Taking their analysis a step further, the investigators estimated the overall impact of a school's retention policy (allowing or banning retention) on the average learning outcomes of students; they also estimated its impact on students who were likely to be promoted if retention were adopted. Propensity score methods and hierarchical linear modeling were again used to address these questions.⁵¹ Results of these analyses indicated that adopting a kindergarten retention policy had no significant effect on students' average learning growth, nor

did it have an effect on the learning growth of students who were likely to be promoted under the policy.

Implications for Estimating Causal Effects

The use of sophisticated statistical techniques, together with a comprehensive dataset based on a large, nationally representative sample, allows the investigators to draw causal inferences about the effects of kindergarten retention policies on student cognitive growth with a relatively high degree of confidence. They are able to make these causal inferences because propensity score matching effectively makes treatment group assignment independent of students' pretreatment characteristics, including their probability of being retained. Previous studies of retention effects that relied on conventional statistical methods often made unwarranted assumptions about the extent of overlap between comparison groups and typically adjusted for only a few background characteristics. In contrast, propensity score matching uses straightforward procedures for determining whether there is sufficient overlap between groups (e.g., with respect to the risk of being retained) and makes simultaneous adjustments for a large number of background characteristics.

A limitation of propensity score methods is that they can adjust only for observed differences in the background characteristics of group members. It is possible that a relevant variable, such as the onset of a serious illness, may have been omitted from the analysis. However, there are statistical techniques that can be used to test for the possible effects of omitted variables (Lin, Psaty, & Kronmal, 1998; Rosenbaum & Rubin, 1983; Rosenbaum, 1986, 2002).

Approximating a randomized controlled experiment with observational data allows the investigators to draw on the strengths of both experimental and observational designs. Because their analyses are based on data from a nationally

representative sample of kindergartners and schools, they are able to describe the characteristics of students who were retained as well as the characteristics of schools that adopted a kindergarten retention policy; thus they can identify students and schools that are most likely to be affected by retention policies.

In addition to estimating the average effect of retention on the cognitive growth of retained students, the investigators are able to demonstrate variation in these effects across schools. Such school-to-school variation suggests that school characteristics, such as approaches to curriculum and instruction, may moderate the negative effects of retention; closer examination of such variation in future studies could prove useful in identifying the causal mechanisms through which retention and promotion affect students' cognitive growth.

Conducting a randomized controlled experiment to study retention effects is likely to be operationally difficult; it could also be argued that since the consequences of retention are inconclusive, subjecting students to an untested condition is unethical. Since a randomized experiment is not feasible in this instance, the investigators creatively use a large-scale national dataset to approximate an experimental design. Their study demonstrates that quasi-experiments can be very powerful if the datasets are well designed and comprehensive and contain reliable and valid measures of the variables of interest. More studies of this type are needed in investigating the effects of educational interventions, particularly in situations where randomized controlled experiments are not feasible.

Case III. Structural Modeling: *Why Schools Matter:* *A Cross-National Comparison of Curriculum and Learning*

Why Schools Matter (Schmidt et al., 2001) investigates the relationship between curriculum (content standards, textbooks, teacher coverage, and teacher instructional time) and learning

using data from the Third International Mathematics and Science Study (TIMSS), a cross-national study of mathematics and science achievement. The authors define student learning as the gains in subject-specific competencies and knowledge over a 1-year period. Their focus on learning—systematic gains over time not due to maturation—leads them to explore achievement gains (the change in achievement from one time point to another) rather than achievement status (a measure of cumulative achievement up to a particular time point). Although the study is not experimental in design, the investigators' use of sophisticated statistical techniques allows them to generate causal hypotheses concerning specific aspects of the curriculum on student learning.

Research Questions and Theoretical Frame

Through an examination of cross-national variation in topic coverage in mathematics and science, Schmidt and his colleagues investigate the relationship between curriculum coverage and student learning gains in these subjects. Variation in topic coverage within individual countries and its relationship to student learning are also examined. The investigators' primary research questions are as follows: (a) To what extent do countries vary in their coverage of particular mathematics and science topics? (b) What is the relationship between topic coverage and student learning gains? and (c) Within individual countries, how does variation in topic coverage across schools and classrooms relate to differences in student learning outcomes?

Research on curriculum and instruction indicates that what gets taught in school and how much time teachers devote to instruction affects student learning and achievement. Students' opportunities to learn are structured both by the content and organization of the curriculum and by the time teachers devote to specific topics of instruction. Research drawing on

both the opportunity-to-learn paradigm (Sorenson, 1970, 1987) and organizational approaches to schooling (see, e.g., Bidwell, 1965, 2000; Bidwell, Frank, & Quiroz, 1997; Firestone, 1985; Ingersoll, 1993; Kilgore, 1991; Kilgore & Pendleton, 1993) has shown that curriculum and instruction are important factors in the stratification of student learning (see, e.g., Dreeben & Gamoran, 1986; Gamoran & Berends, 1988; Gamoran, 1989; Lee, Smith, & Croninger, 1997).

Recognizing the importance of the curriculum for student learning outcomes, policymakers have moved toward developing curricular content standards under the assumption that such standards ultimately will influence what is actually taught in schools. It is at this level that Schmidt and his colleagues enter the debate on curriculum and instruction and approach it as a problem of importance not only to the United States but also internationally. In conceptualizing the relationship between curriculum and student learning outcomes, they move beyond standard definitions of curriculum (teacher content coverage and instructional time) to include national content standards. They also expand their definition of content coverage to include textbook coverage, under the assumption that the textbooks that teachers use for instruction will influence their coverage of particular topics. Four aspects of the curriculum are thus identified in their conceptualization: content standards, textbook coverage, teacher coverage, and time devoted to instruction.

Problems With Research on Curriculum and Learning

As Schmidt et al. note, most studies of curriculum have been qualitative, and studies that have attempted to examine the relationship between curriculum and learning quantitatively have relied on teacher assessments of students' opportunities to learn material tapped by items on achievement tests. Such assessments are potentially biased and unreliable indicators of

topic coverage. In devising measures of specific aspects of curriculum and in using multiple indicators of curriculum coverage, Schmidt and his colleagues are able to provide much stronger evidence linking curriculum and learning. They are also able to quantitatively assess and model the relationships among aspects of the curriculum and the relation of each to student achievement gains. Their use of curriculum-sensitive measurement within major content categories in turn increases the likelihood of finding curriculum effects and of finding effects that vary by topic.

By focusing on specific sets of topics and individual countries, the investigators are able to determine how emphasis on a particular topic varies across countries; they can also determine which topics constitute the core mathematics and science curriculum (e.g., in eighth grade) for the majority of countries participating in TIMSS. Within individual countries, they are able to determine the relative emphasis given to particular topics across classrooms and schools. The investigators observe that in “countries such as the U.S., where local control or even school control of the curriculum is the rule,” one might expect large variations in curriculum coverage (and opportunities to learn) across schools, and thus large variations in student achievement.

Modeling the Potential Causal Effects of Curriculum on Student Learning

The investigators develop and test a structural model of the relationships among specific aspects of the curriculum (content standards, textbook coverage, teacher coverage, and instructional time) and between each of these curricular aspects and student achievement.⁵² As specified in the model, content standards are assumed to influence textbook coverage of specific topics and the selection of textbooks for use in classrooms. Similarly, content standards may affect teacher

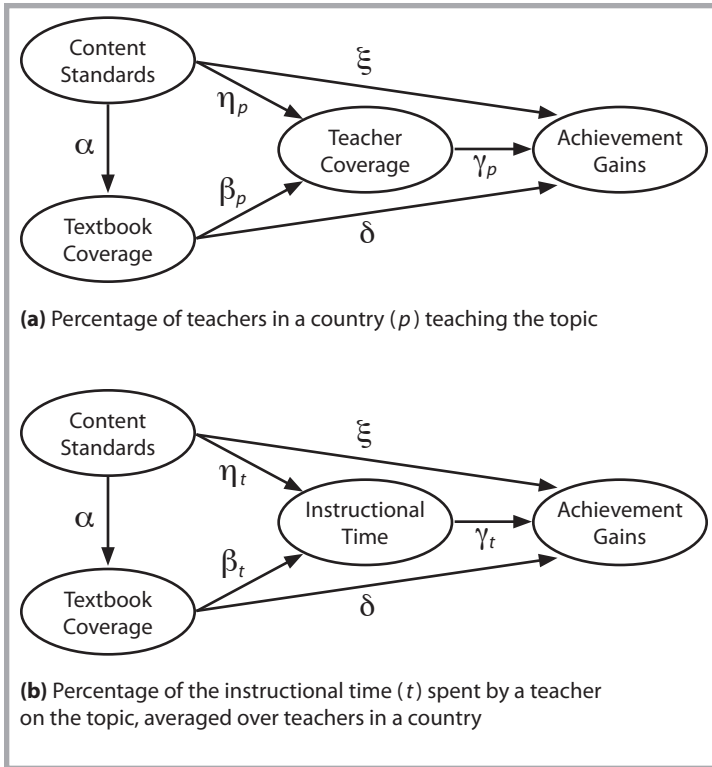


Figure 1. A structural model of relationships among curricular aspects and student learning. Adapted from Schmidt et al., 2001, p. 31, with permission from the author.

coverage or the amount of instructional time devoted to particular topics through their role in teacher preparation and professional development. Textbook coverage of particular topics is also likely to affect teacher coverage of those topics (see Figure 1).

Each of these aspects of the curriculum (content standards, textbook coverage, teacher coverage, and instructional time) may relate to student learning either directly or indirectly. Direct relationships (often referred to as “direct effects”)

can be thought of as the simple path from content standards to learning, while the indirect relationships can be thought of as the compound path from content standards to textbook coverage to teacher coverage and instructional time to student learning. For example, the quality of textbooks as reflected in the content coverage of particular topics may directly affect student achievement gains for those topics, as well as indirectly affecting such gains through their effect on teacher coverage.

The investigators use this structural model to isolate the relationships among specific aspects of the curriculum on student learning outcomes. They do not attempt to draw causal inferences about these effects. Rather, they conceptually model and statistically evaluate the potential causal effects of specific aspects of the curriculum on student learning. Developing conceptual models and using statistical analyses to identify associations among elements of the model is an important and necessary precursor to designing randomized controlled experiments to test the effects of specific curricular interventions on student learning outcomes. The study is methodologically innovative in its use of sophisticated statistical techniques to estimate the relative effects of different aspects of the curriculum on student learning outcomes and to compare these effects cross-nationally. The sophistication of the analysis is one reason that we chose to include this study as an example of recent work on causal modeling. In isolating the effects of different aspects of the curriculum on learning, the study suggests where curricular interventions might be most effective.

Study Design, Data, and Approach

TIMSS is an international comparative study of mathematics and science achievement involving nearly 50 countries.⁵³ TIMSS focused on three populations of students: (a) two adjacent grades consisting of the majority of 9-year-olds in each

country; (b) two adjacent grades consisting of the majority of 13-year-olds in each country; and (c) all students in the last year of secondary school, with subpopulations focusing on those studying advanced mathematics, physics, or both. For each population, mathematics and science assessments were administered toward the end of the school year.⁵⁴ In addition, students completed surveys concerning their interests, study habits, motivations, and classroom experiences; surveys were also completed by teachers and school administrators.

Curriculum measures. To identify and measure curriculum standards and textbooks for students participating in TIMSS, the investigators systematically collected the official content standards (e.g., curriculum frameworks, guides, national curricula) and a representative sample of student textbooks from each participating country.⁵⁵ Documents were first divided into specific segments or units. For content standards, units specifying content and objectives were the most prevalent. For textbooks, lesson units (the amount of material likely to be covered in 1 to 3 days of instruction) were the most prevalent. Units were further divided into homogeneous blocks for purposes of coding.⁵⁶

To measure teacher implementation (teacher coverage and instructional time), the investigators used responses to questions in the TIMSS Teacher Questionnaire, which was administered to teachers of students in the study. Teachers were asked the number of lessons devoted to specific topics of instruction. Topics were taken from the TIMSS mathematics and science frameworks. Listed topics covered all content areas in the frameworks and were tightly related to specific content topics. For each listed topic, teachers were also asked to indicate how much instructional time was devoted to the topic.⁵⁷

Achievement measures. Scores on the TIMSS achievement tests were used to measure achievement gains in mathematics and science at specific grades. Because test items were

based on content categories specified in the TIMSS mathematics and science frameworks, it was possible to identify the specific content being measured by particular test items.⁵⁸ Achievement gains ideally are measured at two time points for each individual student, so that learning growth can be assessed over time for each student. Due to the design of the TIMSS study, this was not possible. However, because the same tests were administered to two adjacent grades for 9- and 13-year-olds, it was possible to measure achievement at the end of both grades and to construct national estimates of the gains from one grade to the next.

Analyses and Results

To test their model of the relationships among aspects of the curriculum and student learning, the investigators estimated a structural model (one of several analytic approaches taken) that takes into account both the direct and indirect relationships of each of these aspects to student achievement gains, controlling for the other aspects of the curriculum. In cross-national comparisons that controlled for national wealth and other country-level variables, results for the estimated structural model of the effects of the four aspects of curriculum on student achievement gains provide general support for the investigators' conceptual model. For example, in mathematics, content standards were found to be related to teacher implementation both directly and indirectly through textbook coverage; teacher implementation was, in turn, related to achievement gains. In general, the more coverage of topics by a country (whether in content standards, textbooks, teacher coverage, or instructional time), the greater were student achievement gains for that country. These relationships, however, were not uniform across countries as countries varied in content coverage.

In science, a more complex pattern of relationships emerged. Content standards were directly related to student gains, as was teacher implementation (in terms of both instructional time and the percentage of teachers covering a topic). For these two measures of curriculum, the conclusions are essentially the same as for mathematics. The greater the priority a country assigned to a given topic in its content standards, the greater the achievement gains for that topic at eighth grade. The strength of this relationship depended on the country involved. In science, however, the relationship between textbook coverage and achievement gains was negative, indicating that the greater a country's textbook coverage of a topic (relative to other topics), the lower were the achievement gains for that topic.⁵⁹ In contrast to mathematics, there was no direct relationship between content standards and textbook coverage, nor was there a direct relationship between content standards and teacher instructional time. Overall, these results suggest that content standards were unrelated to textbook coverage in science.

Country-specific analyses were also conducted. For the United States, the investigators found that, for both mathematics and science, textbook coverage had a strong direct relationship to achievement gains, as well as a strong indirect relationship through instructional time allotted to particular topics. There was little variation in content standards within the United States because content standards covered virtually every mathematics and science topic included in the TIMSS assessments.

Hierarchical linear modeling was used to analyze the relationship between classroom instructional time and mathematics achievement across classrooms and schools within the United States; indicators of student SES and aggregate classroom SES were included in the model. The relationship of student SES to both achievement and opportunities to learn, as measured by instructional time and teacher coverage, has

been documented in numerous studies in the United States and with multiple datasets (see, e.g., Anderson, Hollinger, & Conaty, 1993; Burstein, 1993; McKnight et al., 1987; Schmidt, McKnight, Cogan, Jakwerth, & Houang, 1999; Raudenbush, Fotiu, & Cheong, 1998; Stevenson, Schiller, & Schneider, 1994). Thus, in examining the relationship between achievement gains and aspects of the curriculum, it was necessary to adjust for student and classroom SES.⁶⁰

Results of this analysis indicated that differences in achievement gains across eighth-grade classrooms were related to the amount of instructional time teachers allocated to particular topics, even when adjusting for differences across classrooms in SES and prior achievement. In general, the greater the instructional time devoted to a particular topic (measured as a percentage of total instructional time), the greater were achievement gains for that topic. The relationship between instructional time and mathematics achievement gains was positive and significant for all subtest areas in geometry, algebra, and proportionality. However, the relationship between achievement gains and instructional time devoted to whole numbers and fractions was negative and significant. These are topics typically taught in earlier grades, so instructional time devoted to these topics in eighth grade may do little to improve students' mastery of the topics.

Demanding performance expectations were also positively related to achievement gains for several advanced topics, including polygons and circles, three-dimensional geometry, and functions. These results suggest that engaging children in activities that go beyond routine drill and practice has an effect beyond the amount of time devoted to instruction. For U.S. eighth graders, the *quality* as well as the quantity of instruction appears to be important to achievement gains in mathematics, at least for these topics. Additional analyses examined predicted increases in achievement associated with increases in instructional time devoted to particular topics. The largest

predicted increases were for geometry-related areas, proportionality problems, and equations; these were topics in which the United States generally did not provide much topic coverage. These results suggest that even a modest increase in the instructional time devoted to these topics could substantially increase student learning gains.

Implications for Estimating Causal Effects

The use of structural modeling allows the investigators to model the potential causal relationships between curriculum and learning. They are able to estimate both direct and indirect relationships between specific aspects of the curriculum and student achievement gains. For example, in mathematics, content standards appear to influence learning directly as well as indirectly through textbook coverage and teacher implementation. The investigators thus are able to identify variables through which learning may occur, information that is important in designing intervention studies.

Although this study does not provide evidence of a causal relationship between curriculum and learning, it does provide evidence of a strong association between them based on several different measures. In many cases, the aspects of the curriculum related to achievement gains differed for different topics and countries. However, some significant relationship between curriculum and achievement gains was found for all but five countries, even when controlling for national wealth and other country-level variables.

The investigators note several limitations of the study. They observe that “measures used in analyses are not perfect indicators of [curriculum] emphasis, neither do they function as perfect statistical indicators” (p. 359). In turn, measures of achievement gain are based on comparisons of the assessment scores of students from adjoining grades rather than on differences in the scores of the same students at two different time

points (a longitudinal design). The investigators also acknowledge that both longitudinal studies and experiments with randomization are needed to refine their analysis of curricular effects.

Despite its limitations, this study moves us much closer to being able to construct and implement intervention studies designed to assess the causal relationship between curriculum and learning. The investigators' analysis of a large-scale observational dataset provides guidance on where to focus our efforts with respect to providing adequate coverage of topics as well as creating greater coherence across topics and aspects of the curriculum. For example, textbook coverage in the United States was found to have strong relationships to achievement gains in science and mathematics (both directly and indirectly through teacher instructional time), suggesting that increased textbook coverage of particular topics would result in achievement gains for those topics. The analysis of the relationship between instructional time and mathematics achievement among U.S. eighth graders also suggests that additional time devoted to particular topics in geometry and algebra would result in fairly large achievement gains for these topics. In addition, for certain advanced topics, the positive relationship between mathematics achievement and performance demands, as measured by the complexity of instructional activities students engaged in, points to the importance of the quality as well as the quantity of instruction for student learning. These findings suggest a potential focus for intervention studies based on experimental designs with randomization.

Case IV. A Standard Analytic Approach: “The Role of Gender and Friendship in Advanced Course-Taking”

In “The Role of Gender and Friendship in Advanced Course-Taking,” Riegle-Crumb et al. (2006) examine the role of friendship groups in male and female high school students' advanced

course-taking using data from the National Longitudinal Study of Adolescent Health. Given the continuing gender gap in science and mathematics achievement, the investigators are particularly interested in whether friends positively influence girls' advanced high school course-taking in these subjects. Focusing on the importance of same-sex friends as role models and sources of support, they examine whether girls who have high-achieving same-sex friends early in high school are more likely to enroll in advanced mathematics and science courses in their junior and senior years.

Research Questions and Theoretical Frame

Three research questions are addressed in the study: (a) Is same-sex friends' academic achievement positively associated with the advanced course-taking of male and female students? (b) Is same-sex friends' academic achievement more strongly associated with girls' advanced course-taking in mathematics and science (stereotypically male subject areas) than with their advanced course-taking in English (a stereotypically female subject area)? and (c) Do gender and friendship group composition interact such that there is a stronger relationship between same-sex friends' academic performance and girls' advanced course-taking when girls' friendship groups are predominantly female?

Theories of adolescent development suggest that peers become increasingly important during adolescence (see, e.g., Coleman, 1961; Erikson, 1968). By the time adolescents enter high school, they spend less time with their families and more time with friends. Depending on the nature of the relationship, friends may positively or negatively influence behavior. The social-psychological literature on adolescent development suggests that friendships may function differently for males and females. Girls' friendships with each other tend to be more supportive and encouraging than friendships between boys

(Beutel & Marini, 1995; Felmlee, 1999; Giordano, 2003; McCarthy, Felmlee, & Haga, 2004; South & Haynie, 2004). Boys' friendships tend to be more competitive and activity-based, whereas girls' friendships are more cooperative and centered on discussion (Beutel & Marini, 1995). Although friendships with the opposite sex emerge in adolescence, same-sex friends continue to be important companions and role models and may have a greater influence on academic outcomes (Schneider & Stevenson, 1999).

Limitations of Previous Research on Peer Influences

Previous research on adolescent friendships has focused primarily on their potential to negatively influence behavior through encouraging drinking, drug use, or other problem behaviors (see, e.g., Granic & Dishion, 2003; Matsueda & Anderson, 1998; Warr, 1993; Weermand & Smeenk, 2005). The potential for friendships to positively influence adolescent behavior and development has received less attention. The few studies that have been done suggest that friends can play an important role in encouraging academic achievement (Crosnoe, Cavanaugh, & Elder, 2003; Epstein, 1983; Hallinan & Williams, 1990). Given the potential importance of friendships to educational outcomes, Riegle-Crumb et al. (2006) focus on whether having high-achieving friends of the same sex is associated with advanced course-taking, particularly in mathematics and science.

Although women have begun to enter mathematics and science occupations in greater numbers, and gender differences in mathematics and science test scores have declined over the past few decades, girls are still less likely to express interest in mathematics and science in high school or to see themselves as competent in these subjects, even when they perform at similar levels (Benbow & Minor, 1986; Correll, 2001; Xie & Shauman, 2003). Building on previous research indicating the

importance of adult mentors and same-sex schools and classrooms in promoting girls' interest and advanced course-taking in mathematics and science, Riegle-Crumb et al. (2006) suggest that girls whose same-sex friends are high achieving may be more likely to take advanced coursework in mathematics and science.⁶¹ Such friends may help to establish norms about doing well in these subjects, function as role models, and serve as sources of emotional or psychological support. Riegle-Crumb and her colleagues, in turn, suggest that this association will be stronger in the context of a friendship group that is predominantly female.

Identifying Potential Causal Effects Using Conventional Statistical Techniques

Friends are not selected at random. It is therefore not possible to design a randomized experiment to investigate the potential effects of friendships on an outcome such as advanced course-taking. Approximating a randomized experiment through an approach such as propensity score matching is also not feasible. Propensity score matching assumes that there is a mechanism by which individuals are assigned to groups when assignment is nonrandom, as in the case of retention and promotion. Friends are not assigned, however; they are chosen, and many factors other than academic performance may influence that choice for any given individual. To examine the relationship of friendship characteristics to academic outcomes, analysts must rely on non-experimental designs and use statistical techniques to adjust for selection bias.

As Riegle-Crumb et al. (2006) indicate, selection bias is a primary concern in studies attempting to model friendship influences. Because individuals may select friends who have similar characteristics, it can be difficult to determine whether friends have an independent (socializing) effect on an individual's behavior. Although the investigators argue that

friendship effects on course-taking are likely the result of both selection and socialization, they take several steps to reduce the likelihood that any given association is due only to selection. For example, in attempting to isolate the relationship between same-sex friends' grades and advanced course-taking in a given subject, they controlled for the influence of respondents' grades in the same year and course level in that subject. They also checked to determine whether students who were high achievers were more likely than low achievers to benefit from having high-achieving friends; no significant differences between high and low achievers were found in these analyses. In correlating respondents' grades with their same-sex friends' average grades for a given subject, they found only a moderate association ($r = .4$), suggesting that students selected their friends based on a number of factors other than academic achievement.

Measuring advanced course-taking at a later time point (11th and 12th grades) than friends' academic achievement also reduced potential selection bias. By examining multiple outcomes (male and female advanced course-taking in math, science, and English), the investigators were also able to determine whether relationships between friends' academic achievement and students' advanced course-taking varied by gender and subject. If associations vary across outcomes, then the likelihood that these results are due to selection bias rather than socialization is reduced.

Although there are statistical procedures for reducing bias, there is always the possibility that observed or unobserved characteristics associated with the outcome have been omitted from analytic models. It is therefore not possible to draw causal inferences from these analyses; they can only demonstrate associations between particular characteristics and outcomes of interest, and analysts should be careful not to use causal language in describing their results. However, well-designed observational analyses can provide insights into

relationships that cannot be studied experimentally, provide evidence that confirms the results of previous studies, and suggest where interventions that can be studied experimentally might be most effective.

Study Design, Data, and Approach

To investigate the role of friends in students' advanced course-taking, the investigators used data from the National Longitudinal Study of Adolescent Health (Add Health) and the study's high school transcript data component. The Add Health Study included an In-School Survey, administered in the fall of 1994 to almost all students in Grades 7–12 in a nationally representative sample of schools, and three waves of In-Home Survey data collected from a representative sample of students in each school in 1995 (Wave I), 1996 (Wave II), and 2000–2001 (Wave III). In 2002–2003, high school transcript data were collected from the high schools attended by Wave III respondents.⁶² For purposes of analyses, the investigators selected only students who were 9th and 10th graders in 1994–1995, had completed the In-School Survey and the Wave I In-Home Survey, and for whom high school transcript data were available. This selection process resulted in a subsample of approximately 2,500 students. The subsample is generalizable to U.S. 9th- and 10th-grade students with at least some friends who attended the same high school; statistical procedures were employed (weights) that make groups of individuals (categorized by race and ethnicity) in the subsample proportional to their numbers in the U.S. population.⁶³

Measure of advanced course-taking. Advanced course-taking in science, mathematics, and English, the outcome measures in this analysis, were measured based on students' enrollment in their junior or senior year in high school in the following subjects: physics (science), pre-calculus or calculus

(math), and advanced placement (AP) English or honors English IV (English), as recorded in the high school transcript file. These are the most advanced courses offered in each of these subject areas and typically are taken in the junior or senior year because of the prerequisites for course entry. Course information was coded using a classification scheme developed by the National Center for Education Statistics.

Friends' characteristics. The measures of friends' characteristics used to predict these outcomes were taken from students' responses to the In-School Survey. On the survey, students were asked to identify their five closest female friends and five closest male friends.⁶⁴ Measures of friends' academic performance include grades earned in science, mathematics, and English courses at the beginning of high school. Because prior research indicated that the academic performance of same-sex friends was more likely to influence girls' rather than boys' course-taking decisions, measures of the average grades of same-sex friends in each of these subjects were separately constructed for males and females. A dichotomous measure of the gender composition of students' friendship groups was also created.⁶⁵ To determine whether girls with predominantly high-achieving female friends were more likely to take advanced coursework, an interaction term was created by combining the average grades of same-sex friends in a given subject with the composition of the friendship group (i.e., predominantly female versus gender-equal or predominantly male).⁶⁶

To gauge students' levels of involvement with friends, measures of activities that students engaged in with friends were constructed. On surveys, students indicated for each friend listed whether they had, in the past week, visited the friend's house, spent time together after school, talked on the phone, and/or spent time together over the weekend. Responses were summed for each friend and averaged across friends.⁶⁷

Analyses and Results

The authors used logistic regression analysis to estimate the probability that students would enroll in physics, pre-calculus or calculus, and AP English or honors English IV in their junior or senior year of high school.⁶⁸ Separate analyses for males and females were conducted for each of these subjects.

For girls, having a predominantly female friendship group was not, by itself, associated with taking physics in the junior or senior year of high school. Grades of same-sex friends, however, were associated with an increased likelihood of taking physics. For example, as the science grades of a girl's female friends increased, the odds that the girl would take physics by the end of high school increased by a factor of approximately 1.5. There was also a significant positive effect for the interaction of friendship group composition (predominantly female) and grades of same-sex friends, indicating that the effect of same-sex friends' science grades was even stronger when girls' friendship groups were predominantly female. For girls, high parent expectations for college were also associated with an increased probability of taking physics.

In the model estimating the likelihood that girls would take calculus or pre-calculus by the end of high school, there was a significant overall association between friendship group composition and advanced course-taking (termed the main effect), such that girls with predominantly female friends were 1.7 times more likely to take calculus or pre-calculus as juniors or seniors. While there was a positive association between same-sex friends' mathematics grades and advanced course-taking, there was also a significant association for the interaction of friendship group composition with same-sex friends' average grades in mathematics early in high school. While all girls appeared to benefit from having girlfriends with higher math grades early in high school, those whose friendship

group was predominantly female were the most likely to take advanced coursework in math by the end of high school.

In contrast to the results for the mathematics and science models, friendship group composition was not significantly associated with girls' advanced course-taking in English. Regardless of the gender composition of their friendship group, girls whose same-sex friends' earned higher grades in English at the beginning of high school (controlling on their own freshman-year English grades) were more likely to take AP or honors English by the end of high school. For example, as the English grades of a girl's female friends increased, the odds that the girl would take AP/honors English by the end of high school increased by a factor of approximately 1.9.

For male students, the grades of same-sex friends had no association with the likelihood of taking advanced coursework in any of the subjects considered. With respect to the gender composition of the friendship group, males whose friendship groups were predominantly male were actually less likely to take physics than those whose friendship groups were gender-equal or predominantly female.

Overall, these results indicate that the grades of same-sex friends are positively associated with the advanced course-taking of female students in all three subjects, but have no association with the advanced course-taking of males. In addition, the association of same-sex friends' grades and girls' advanced course-taking in mathematics and science is stronger when their friendship groups are predominantly female. To estimate the size of these effects, the investigators predicted the probabilities of taking physics, calculus/pre-calculus, or AP/honors English for a White female who did not have a predominantly female friendship group or had friends with a B average in each subject; the investigators then examined the change in probability when these conditions were altered. For physics, the investigators found that the probability of taking this course almost doubled when a girl's friends were mostly

female and earned mostly A's in science. For calculus/pre-calculus, having predominantly female friends who earned mostly A's increased a girl's probability of taking the course from approximately .4 to .7.

Implications for Estimating Causal Effects

This is not a study that documents a causal relationship between friendship groups and student course-taking. It does show, through a longitudinal analysis of a large-scale, nationally representative dataset, some factors that are likely to increase the probability of taking advanced courses. The type of analysis used is representative in many ways of the majority of quantitative work being conducted in the field of education, but the study also has several strengths that make it a good example of this type of research. The investigators identify a research problem that cannot be studied experimentally. Relationships with peers and friends cannot readily be manipulated and studied in the context of an experiment. The investigators do not assume that they have discovered causal effects. At the same time, they are very aware of issues of selection bias and take several steps to reduce it as much as possible within the limits of their research question and dataset. Their work helps to identify one of the potential factors contributing to the gender gap in mathematics and science achievement. These results, combined with those of other studies, provide converging evidence on gender differences in course-taking patterns. Such results are persuasive in the absence of experimental designs.

The investigators also estimate the magnitude of the relationships of friends' grades and friendship group composition to advanced course-taking. In all cases, they find that same-sex friends' grades early in high school had a substantial association with girls' advanced course-taking. Such a finding is important in determining whether to develop an intervention

that builds on the effects of friendship groups on advanced course-taking. An association between friends' grades and advanced course-taking could be statistically significant but still be negligible in terms of magnitude. In this instance, it would make no sense to design an intervention that attempted to promote friendships that were supportive of advanced course-taking. If these effects are large, however, it may be worthwhile to develop and evaluate an intervention designed to promote the formation of peer relationships that are supportive of girls' advanced course-taking in mathematics and science. Schneider and Stevenson (1999), for example, argue that student activity groups such as school-sponsored clubs are particularly important contexts for promoting the development of stable peer relationships around shared interests and activities. An activity-based group designed to encourage girls' interest in mathematics and science might be one context in which to promote peer relationships around such interests. If a goal is to develop an intervention intended to reduce the gender gap in these subjects, it is critical to identify a context in which such an intervention might be introduced.

Chapter 4 Notes

- 33** Secondary analyses are analyses of existing datasets.
- 34** Details of the Tennessee Class Size Experiment and results of the study have been presented in a number of publications, including Achilles, Finn, and Bain (1997), Finn and Achilles (1990, 1999), Krueger (1999), and Nye et al. (2000).
- 35** This subset of schools ranged from 71% (in Grade 1) to 78% (in Grades 2 and 3) of the schools in the complete sample. A comparison of demographic characteristics of the schools in the analytic sample with those in the complete sample revealed that these characteristics were similar for the two samples.
- 36** Reading and mathematics test scores from the Stanford Achievement Test (SAT), administered at the end of each school year in kindergarten through third grade, served as the measures of student achievement. In analyses of achievement gains, student achievement scores from the prior year were included in the models. Within-classroom variables included student gender, SES (coded as 1 if a student received a free or reduced-price lunch, otherwise coded as 0), and minority group status (coded as 1 if a student was Black, Hispanic, or Asian, and coded as 0 if the student was White). Between-classroom variables included class size and presence of an instructional aide, teacher experience, and teacher education.
- 37** Previous studies also confirmed that these deviations did not bias results (see Krueger, 1999; Nye et al., 2000).
- 38** However, students and teachers were not randomly assigned to schools. As Nye et al. (2000) note, "It is clear [from observational studies] that teachers are not randomly allocated to schools. Research on teacher allocation to schools has documented that schools with high proportions of low-income and minority students often have difficulty recruiting and retaining high-quality teachers" (p. 249). See results reported by Darling-Hammond (1995), Krei (1998), and Langford, Loeb, and Wyckoff (2002). By including school characteristics in their model, Nye et al. were able to investigate whether teacher effects varied systematically across schools.
- 39** In order to decompose the variation between students, classrooms, and schools, a three-level model was used (students, classrooms, and schools).

- 40 Teacher effects on reading and mathematics achievement were approximately twice as large as school effects at Grade 2 and approximately three times as large as school effects at Grade 3.
- 41 In Grades 1–3, teacher effects on mathematics achievement were nearly twice as large as teacher effects on reading achievement.
- 42 Between-school variation in achievement status (each year and at third grade) was larger than the between-school variation in achievement gains, suggesting that teacher effects are closer in magnitude to school effects for achievement status. In other words, teachers have a greater impact than schools on student achievement gains from one year to the next. With respect to students' overall achievement (each year and at the end of third grade), however, teacher and school effects are similar in magnitude. Note that school effects are associational, not causal, because students and teachers were not randomly assigned to schools in the experiment.
- 43 Across grades, the proportion of the total variance in reading achievement accounted for by teacher effects was 1.4 to 1.7 times higher in low- versus high-SES schools; the proportion of the total variance in mathematics achievement accounted for by teacher effects was 1.6 to 3.7 times higher in low- versus high-SES schools.
- 44 Because large-scale datasets generally include comprehensive background information on large numbers of students (e.g., gender, race/ethnicity, family structure, SES, prior academic achievement, and a host of other variables), a large number of characteristics can be taken into account in computing propensity scores. Such large-scale longitudinal datasets also include data on students' performance on standardized tests over time, making it possible to examine differences in cognitive growth of retained and promoted students who are at similar risk of retention.
- 45 Hong and Raudenbush compared the 11,843 students in their analytic sample with the full ECLS sample to determine whether the analytic sample was a representative subsample. They did find some differences. The analytic sample had a lower percentage of poor and minority children (a 2–3% difference) and were less likely to come from non-English-speaking families (a 4% difference).
- 46 For a random sample of 4,024 students, assessment data were also obtained in the fall of the treatment year.
- 47 Classroom and school characteristics were also significant predictors of retention. Students who were in kindergarten classes with higher

proportions of boys, higher proportions of younger children (e.g., 4-year-olds), and higher proportions of children who were repeating kindergarten were more likely to be retained. Teachers of such kindergarten classes also reported more behavioral problems at the beginning of the year and tended to spend less time in reading and literacy instruction and to cover lower-level content in reading and mathematics. Children were also more likely to be retained if they attended schools that were smaller in size, nonpublic, had inadequate instructional resources and facilities, lower teacher salaries, and fewer classroom teachers and ESL teachers.

- 48** For both reading and math, the investigators found that the *observed* achievement gap between retained and promoted students doubled in width between the fall and spring of the treatment year. On the basis of estimated growth rates, if retained students had instead been promoted, their growth rates would have been comparable to those of promoted students, substantially reducing the achievement gap in both reading and math.
- 49** Results of supplementary analyses also suggested that there was a diminishing effect of retention for students who had a greater probability of being retained. In other words, if these high-risk students had instead been promoted, their growth trajectories would have remained low. However, the authors indicate that “even for those who tended to be diagnosed as in a relatively higher need of repeating a grade, there was no evidence that they received any immediate benefit from the retention treatment. In general, kindergarten retention seemed to have constrained the learning potential of all but the highest-risk children” (Hong & Raudenbush, 2005, p. 220).
- 50** It was assumed that there might be unmeasured student- and school-level characteristics that were comparable to the most important student- and school-level variables in their models for each subject area. Adjustments for the inclusion of these hypothetical variables were made, and retention effects were re-estimated (Lin, Psaty, & Kronmal, 1998; Rosenbaum & Rubin, 1983; Rosenbaum, 1986, 2002).
- 51** In their examination of the characteristics of retention and nonretention schools, the investigators identified 238 school-level variables that were associated with retention. They found that nonpublic schools, suburban schools, and schools with lower percentages of minority students and teachers were more likely to adopt a kindergarten retention policy. In general, retention schools tended to have

smaller class sizes, greater parent involvement, and fewer disciplinary problems than nonretention schools. In the pretreatment year, kindergartners in retention schools also had higher average reading scores than those in nonretention schools. However, when propensity score methods were used to create groups that were balanced with respect to these school characteristics, no significant differences in the learning outcomes of students were found between groups.

- 52** Structural modeling (also referred to as structural equation modeling, or SEM) is an extension of regression analysis that offers several advantages. In contrast to multivariate regression analysis—where associations between multiple predictor variables and one outcome variable are modeled, and associations among predictor variables are adjusted for—structural modeling allows for the inclusion of more than one outcome variable. Whereas in multivariate regression analysis a variable can be *either* a predictor variable or an outcome variable but not both, in structural modeling a given variable may be an outcome variable with respect to some variables and a predictor of other variables. For example, teacher instructional practices may be an outcome of content standards and textbook coverage but may also be a predictor, along with content standards and textbook coverage, of student achievement; structural modeling allows the analyst to model this complex relationship.

Although structural modeling is sometimes referred to as causal modeling, it does not allow the analyst to make causal inferences. As Norman and Streiner (2004) observe, “Cause and effect can be established only through the proper research design [e.g., a randomized controlled experiment]” (p. 159). Structural modeling is a model-testing procedure. A conceptual model that specifies relationships among a set of variables is tested by means of appropriate statistical procedures.

- 53** When this study was conducted, the name was as it appears in the text. The name has now been changed to Trends in International Mathematics and Science Study.
- 54** The items included in the TIMSS assessments are based on a categorization of topics that describe possible contents of the mathematics and science curricula in participating countries and the performance that might be expected of students with respect to these content areas. As the authors note, these category systems—the TIMSS mathematics and science frameworks—“were developed to provide a

common language for describing and examining what students in many different countries study in their schools. Although the frameworks were developed and published in English, they needed to be sufficiently broad to include any topic found in any of the participating countries' curricula, yet sufficiently precise as to provide accurate portraits that could be compared and analyzed" (p. 21). Because a major focus of TIMSS was on 9- and 13-year-olds, the curriculum frameworks were developed with elementary and middle school students in mind. In addition to specifying subject matter topics, the TIMSS frameworks also specify performance expectations ("what students were expected to do with particular subject matter topics") and perspectives ("any overarching orientation to the subject matter and its place in the disciplines and in the everyday world") (p. 363).

- 55** The authors note that "the selection criteria for the documents that would be coded—standards and textbooks—required that a national sample include sufficient documents that pertained to at least 50 percent of the students in the TIMSS focal grades. In addition a country's document sample was required to cover all major regions and all types of schools and educational tracks (e.g., public, private, vocational, technical, and academic)" (p. 24).
- 56** "Each block [from the TIMSS mathematics and science frameworks] . . . was coded by assigning as many content categories, performance expectations, and perspectives to it as were needed to characterize the content" (p. 24). The measures of standards and textbooks were for a specific year of schooling (i.e., fourth grade or eighth grade). The measures of classroom instruction were based on teachers' responses regarding topic coverage in a particular class during the year in which the TIMSS achievement tests were administered.
- 57** Instructional time was coded as follows: 1–5 periods/lessons; 6–10 periods/lessons; 11–15 periods/lessons; or more than 15 periods/lessons.
- 58** The validity of the TIMSS test items was assessed by several panels of U.S. mathematicians and scientists. Panel members concluded that the items included in the TIMSS assessments adequately represented and measured the specific topics covered. Measures of reliability (an indication of the amount of measurement error in the tests) indicated that measurement error was relatively low. The median reliability estimates for the TIMSS eighth-grade mathematics and sciences tests were .78 and .89, respectively (coefficients range from 0 to 1.00; higher coefficients indicate greater reliability/lower measurement error).

- 59** Additional analyses were conducted to determine why there was a negative relationship between textbook coverage and achievement gains in science. The analyses revealed that three content areas primarily accounted for this relationship: energy and physical processes, chemical changes, and structure of matter. Both physical processes and chemical changes constituted a large proportion of textbook coverage across countries relative to other topics, but student gains for these topics were only average. In contrast, the topic of structure of matter constituted a small proportion of textbook coverage but showed the largest achievement gains. When these topics were dropped from the analysis, the relationship between textbook coverage and student achievement gains was much less negative. One explanation for these anomalous findings may be that the TIMSS test produced floor and ceiling effects for these topics in seventh grade, which limited estimates of achievement gains between seventh and eighth grades.
- 60** Due to sampling limitations, it was not possible to relate classroom instructional time to student achievement gains in science. Many countries organized eighth-grade science instruction into separate courses, and sampling may have included teachers of different science courses within a given country, making it difficult to link student and science teacher data. This set of analyses therefore focuses only on mathematics.
- 61** See, e.g., Baker and Leary (1995), Dryler (1998), Eccles, Jacobs, and Harold (1990), Lee (2002), Shu and Marini (1998), Seymour and Hewitt (1997), and Stake and Nicken (2005) on the importance of adult mentors. See Burkham, Lee, and Smerdon (1997), Shapka and Keating (2003), and Lee and Bryk (1986) on student interest and course-taking in science and mathematics in predominantly female environments.
- 62** Overall, data were collected from six nationally representative cohorts of students based on their grade level (7th through 12th) in 1994–1995.
- 63** In addition to weights, the analysts used a statistical program to account for the clustering of students within a school when calculating standard errors, that is, an estimate of the deviation of the sample mean from the population mean.
- 64** Since almost all students within a given school were surveyed, the investigators were able to link the survey responses of identified friends with those of the respondent. Measures of friends' characteristics

were thus based on the friends' self-reports rather than on respondents' characterizations of their friends' qualities.

- 65 A friendship group that had a greater number of same-sex friends was coded as 1; a group that had a greater number of opposite-sex friends or was gender-equitable was coded as 0.
- 66 An interaction term is a variable that takes into account the associations between two measures in order to predict their relationship with the dependent variable, independent of the separate effects of each.
- 67 All models included the following variables: students' race and ethnicity; parents' education level; family income; family structure; a measure of students' self-perceived intelligence relative to peers; students' educational expectations; parental expectations; school engagement; school attachment; students' freshman year course placements in science, math, and English; and their corresponding grades in those subjects. Freshman-year math and science course placements were assigned a numerical coding based on the level of the course taken (e.g., for science, no science = 0; remedial science = 1; general/earth science = 2; biology I = 3). For English, a dichotomous measure was created indicating whether the student was enrolled in an honors English course as a freshman.
- 68 Several types of analysis are used to identify associations between the dependent and independent variables; the specific approach used depends on the nature of the outcome variable being analyzed and how it is measured. In its most easily interpretable form, this procedure is termed *linear regression* and is used to test whether there is a linear relationship between an outcome variable and a set of variables believed to be associated with that outcome; the strength of relationship between the dependent variable and each of the independent variables is also calculated, adjusting for any associations that might exist among the set of independent variables. When an outcome variable is a continuous variable such as grade point average (GPA, which can take on a range of values, usually between 0 and 4), the strength of this relationship can be described in terms of the change in the outcome variable (e.g., an increase or decrease in GPA) associated with the change in a particular independent variable (e.g., hours per week spent on homework). Say, for example, an increase of an hour per week in study time is found to be associated with an

increase of 0.25 in students' overall GPA (controlling for other variables included in the analysis).

When the outcome measure is dichotomous (coded 0 or 1), as is the case in this study, the outcome either occurs or does not (e.g., a student either takes physics or does not). The outcome variable does not take on a range of values, as in the example above, but has only two values. The relationship between the dependent variable and an independent variable (e.g., friends' grades in science early in high school) cannot be meaningfully expressed as an increase or decrease in physics course-taking associated with friends' grades; what is being measured is not the number of physics courses taken but only whether the student did or did not take physics. In this case, it is desirable to express this relationship as the probability that the outcome will occur (taking physics) when certain conditions apply (e.g., friends earn higher or lower grades in science early in high school). This is done using a form of regression analysis known as logistic regression—the statistical technique used in this study.

In addition to the statistical analyses reported here, Riegle-Crumb et al. also conducted analyses using HLM, which allowed them to examine the variability in advanced course-taking both within and across schools. These analyses yielded similar results with respect to the associations between friendship groups and advanced course-taking regardless of the size of the sample within each school.