Scaling-Up Exemplary Interventions

by Sarah-Kathryn McDonald, Venessa Ann Keesler, Nils J. Kauffman, and Barbara Schneider

Scale-up is the practice of introducing proven interventions into new settings with the goal of producing similarly positive effects in larger, more diverse populations. Scale-up *research* examines factors that influence the effectiveness of interventions as they are brought to scale across settings. This article has three objectives. First, it defines the goals of scale-up research with respect to broader efforts to enhance the quality of educational research and promote evidence-based education. Second, it clarifies the importance of context, conceptually and methodologically, in conducting scale-up research. Finally, it suggests practical guidelines that can assist researchers in developing designs that can be implemented in field settings to produce robust, generalizable findings.

In 2005, students across the United States were assessed in several achievement areas on the National Assessment of Educational Progress (NAEP). Reports from the periodic NAEP assessments, commonly known as "the nation's report card," indicate whether or not there have been significant increases or declines in achievement among 4th graders, 8th graders, and 12th graders over time. Whatever the results of the recent NAEP assessment, many teachers, district superintendents, colleges of education, researchers, politicians, and policymakers will be expected to explain what the results indicate about factors that impede or enhance learning among various groups of students in different types of schools. Attempting to answer such questions is indeed problematic: Some would argue that isolating the correlates and causes of differential rates of improvement across groups remains an inexact science. Yet determining what works best, for whom, and under what conditions are the central tasks that educational researchers are being asked to address. Teachers, principals, and district-level administrators look to educational research to help them identify curricula, pedagogy, and professional development activities that result in improvements in student learning. Policymakers and employers seek confirmation that tax-supported investments in education produce the desired outcomes. This increased attention to "what works" has challenged the educational research community to identify interventions with a reasonable prospect of producing sustainable improvements in school-based learning for broad sections of the student population.

Improving Student Achievement: The Challenges

For the past 8 years, federal education policy has actively supported a variety of initiatives focused on enhancing the quality of educational research.[1] These initiatives have been designed to ensure that the demands for improvement in education culminate in sound, systematic, and successful efforts to close achievement gaps. Federal efforts to promote evidence-based educational research are most visible in the No Child Left Behind Act of 2001 and the Education Sciences Reform Act of 2002, which encourage researchers to use random assignment field trials to assess interventions.[2] In such trials, subjects are randomly assigned to treatment conditions to minimize the impact of differences in participant characteristics between groups on the outcome of interest (National Research Council, 2004). One focus of concern has been the relatively small proportion of studies that employ random assignment to treatment and control groups, sometimes referred to as the gold standard for establishing internal validity (see Boruch, 2002; Cook, 2003; Shadish, Cook, & Campbell, 2002).

Random controlled field trials are seen as a necessary step toward the successful scale-up of interventions. One effort to increase the number of random field trials in education research is the Interagency Education Research Initiative (IERI), a collaborative venture of the National Science Foundation, the National Institute for Child Health and Human Development, and the U.S. Department of Education. IERI was developed in response to a 1997 report of the President's Committee of Advisors on Science and Technology Panel on Educational Technology. In that report, the panel "strongly urged that a significant federal research investment be undertaken to improve our understanding of how children learn and to support the development of best practice that supports optimal learning. . . ." It was envisioned that this investment would "support the development, testing, and implementation of scalable and sustainable interventions to improve teaching and learning, particularly through the use of technology" (U.S. Office of Science and Technology Policy, n.d.).

Although IERI emphasizes the importance of interdisciplinary research that is focused on scaling-up exemplary interventions to positively affect student learning and achievement in larger numbers of classrooms, the concept of "scale-up" is not fully operationalized or theorized in the IERI program solicitations.[3] It is often assumed that scale-up is solely about numbers. However, there has been considerable discussion suggesting that scale-up should be conceived multi-dimensionally. One of the more encompassing approaches is Coburn's analysis of what it takes to move "beyond numbers to deep and lasting change" (2003, p. 3). Coburn argues that scale-up requires "consequential change, endurance over time, and a shift such that knowledge and authority for the reform is transferred from external organization to teachers, schools and districts" (2003, p. 4). She proposes conceptualizing scale in four dimensions: "depth, sustainability, spread and shift in reform ownership" (2003, p. 4).

We agree with Coburn's assertion that scale is under-theorized in the educational literature (2003, p. 3), but we disagree with the

Educational Researcher, Vol. 35, No. 3, pp. 15–24

argument that scale-up should be redefined as about something *other than* "increasing the numbers of teachers, schools, or districts involved in a [successful] reform" (Coburn, 2003, p. 4). We view scale-up as inherently about size, numbers, "doing more"—about extending the reach of an exemplary intervention to produce similarly positive effects in different settings and to help a greater number of students. Interventions that are not implemented with larger numbers (of students, teachers) are not "scaled-up"—they are local interventions with promising results. Slavin (2002, cited in Foorman, Santi, & Berger, in press) provides an algorithm for scale-up that highlights this concept:

scale = number of students × time × impact

Individual scale-up studies investigate whether a given intervention leads to improvements in clearly defined outcomes among a particular population of students. Each produces an essentially dichotomous answer—either the intervention does or does not lead to improvement in a given set of circumstances. Conducting such studies is an important first stage in scale-up research.

To determine whether an intervention is scalable, it is necessary to conduct multiple studies in various settings and with various populations of students. The goal in conducting multiple studies is to generate a body of evidence that demonstrates that the intervention can produce similarly positive results across settings and populations. Accumulating results from multiple trials of an intervention's effectiveness is a second and essential stage of scale-up research. Individual studies indicate only that the intervention works in a particular setting.

There is arguably a third stage in scale-up research. It involves the ongoing evaluation of implementations to enrich understanding of the factors influencing efficacy and sustainability. It is at this stage that key questions raised by Coburn and others (e.g., regarding the depth of change and the locus of reform ownership) are particularly salient. Such concerns are not, however, of interest only in post-scale implementation studies. Scale-up research is iterative; implementation evaluation (Stage 3) also informs the identification and implementation of promising interventions (Stages 1 and 2). A good example of this is found in Foorman, Santi, and Berger's discussion of a 3-year randomized study of the impact of technological tools to facilitate assessment-driven instruction using the Texas Primary Reading Inventory (Foorman, Santi, & Berger, in press).

The Importance of Context in Conducting Scale-Up Research

Increasing the number of students exposed to an intervention typically leads to increased variation in the contexts (classroom, school, district) in which the intervention occurs. It is the variability introduced by these contextual differences that creates uncertainty regarding the potential of an intervention to be brought to scale. Where contexts are fixed, scaling can be achieved through replication. However, significant modifications may be required to produce "identical" results when contexts differ. Many disciplines other than education have grappled with this issue.

The classic examples of scaling through replication come from chemical engineering, one of the physical sciences in which the concept of scale-up was first developed more than a century ago. Innovations are developed in laboratory settings and, if they prove useful, are often replicated on a larger scale, moving from the laboratory or bench test to "full scale" manufacture. In commercial settings, this typically translates into the task of transferring procedures developed in the laboratory into protocols that guide production in manufacturing plants (see Levin, 2001; Palakodaty, Walker, Townend, et al., 2000; York, 1999). When the heuristic does not hold in every dimension, the simple, linear "uncritical diffusion" model that many intuitively associate with scale-up clearly no longer applies. Strategies for scale-up that are developed under conditions of complete similarity may provide elegant (slightly simplified) models to guide practice, but the working protocols that are developed from them are suboptimal and may, by definition, be incapable of producing the original (desired) results (see Zlokarnik, 2002, p. 37; Royal Society of Chemistry, 1999, p. 2). To solve this problem, intermediate research is conducted to track small changes in the precise nature of interactions as one moves to successively larger scales of operation—information that is used to develop tailored procedures that yield the same outcome at other scales—a stages and test-beds approach to scale-up (see, e.g., Zlokarnik, 2002).

This simple notion—that it may be necessary to tailor an idea, product, process, or solution that "works" in order to achieve consistently reliable results—is the hallmark of approaches to scale-up in marketing, manufacturing, strategic management, and medicine. Successful commercialization strategies typically employ a mixture of customer-driven designs and customer-conscious adaptations and iterations, a tailored model approach to scale-up (see Conley & Wolcott, in press). In strategic management, scale-up requires that general precepts be applied appropriately; they are used to develop and guide the implementation of flexible, adaptable strategies for achieving organizational objectives in dynamic, occasionally volatile, conditions (see, e.g., Flamholtz & Randle, in press). Managers are trained to be acute observers of the internal and external environments in which they operate, diagnosing situations and prescribing courses of action with regular check-ups to determine their continued relevance and efficacy, a contingent, "scaling-to" approach to scale-up (see McDonald & Schneider, 2004). Similarly, in medicine, clinicians combine information on efficacy, potential adverse reactions, and contraindications with case-specific information in deciding how, if at all, to administer "proven" remedies to individual patients.

The more recent focus on scale-up in education also underscores the importance of understanding the context in which interventions are implemented and student learning occurs (see, e.g., Blumenfield, Fishman, Krajcik, et al., 2000; Bodilly, Keltner, Purnell, et al., 1998; Corcoran, n.d.; Elmore, 1996; Hassel & Steiner, 2000; McDermott, 2000). Skilled teachers, like their counterparts in medicine, management, marketing, and chemical engineering, do not expect that cookie-cutter solutions will be sufficient to adequately address the challenges posed by various, dynamic environments with unique and changing target populations. To inform their decisions, teachers need more than scientifically based evidence that a given intervention will yield consistently similar results when implemented with fidelity. They need information on the consistency and predictability of results when interventions are adapted, when they are *not* implemented with fidelity, and when the characteristics of the target population differ.

A context-focused approach to scale-up combines a commitment to establishing an evidence base on the effectiveness of interventions, with the recognition that powerful environmental influences mean that "proven" interventions must be implemented with a combination of fidelity and appropriate flexibility. Identifying the key contextual variables that must first be controlled and later varied to ascertain an intervention's ability to consistently produce the desired impact on student learning is an important step in designing educational scale-up research. One approach is to build on the literature from sociology of education and use statistical models to determine the extent to which these contextual factors co-vary with student outcomes, an approach that is discussed below. The scale-up researcher then faces the challenge of designing research that will accurately measure the effect of the intervention on a certain sample population, in order to produce valid evidence that the intervention can be used in larger, more varied populations. In other words, the scale-up researcher must design adequately powered, practical, and internally and externally valid studies.

The Impact of Context on Student Learning: Key Lessons From the Sociology of Education

Classroom-based learning is heavily contextualized. Powerful influences—which teachers are often powerless to control or manipulate—impede, constrain, support, and promote student learning. These include educational and occupational aspirations (influenced by family characteristics and individual abilities and interests), social contexts, community influences, norms, values, cultural capital, the social organization of schooling, social relationships in school communities, and the roles and beliefs of members of the teaching profession (see, e.g., Hallinan, 2000).[4] Each student is affected by multiple factors widely understood to account for variation in student achievement. For example, we expect major national comparisons to report on key subnational differences in order to compare across similar contexts.

However, individual student characteristics alone cannot be used to explain the success or failure of an intervention. Sociologists have provided compelling empirical evidence of school- and classroom-level influences that promote and constrain student learning. Important sources of variation operating at the school level include the beliefs, commitments, education, experience, roles, professionalization, and autonomy of teachers; the formal and informal organizational properties of schools; the occupational structure of the teaching profession; teachers' interactions with their colleagues and their students; teachers' knowledge of the subject-matter focus of the classrooms to which they are assigned; curriculum differentiation and the allocation of social and material resources for instruction, including the basis on which teachers organize classes and group students for discussion; the academic climate of schools; and, in high schools, programmatic and course-sequencing issues.

These variables are likely to influence not only achievement but also the interventions designed to improve achievement. In scale-up research it is essential to vary contexts in order to identify possible context–intervention interactions. Given both individual- and school-level effects on learning, a multilevel model is needed to capture the influences of these contextual variables both on the intervention and on student achievement.

Fortunately, advanced statistical techniques (Bryk & Raudenbush, 1992; Goldstein, 1995) enable us to develop empirical models that simultaneously capture individual- and school-level influences on student achievement. These techniques make it possible to quantify the influence of a particular combination of resource and structural factors on student learning, controlling for student background characteristics and thereby isolating school effects. Such analyses help to address questions of whether it is reasonable to expect that student achievement gains attributable to the school can be sustained or further improved, given the changing nature of student and teacher populations and school-level resources (see, e.g., Raudenbush & Willms, 1995).

This information is also critically important to research design. Multilevel models informed by sociological studies demonstrate the importance of sample designs that capture variation at the individual and school levels. Secondary analyses of major national datasets provide a more finely grained understanding of the relative variation in student learning outcomes between and within schools, and between and within classrooms. Specifically, it is possible to calculate the proportion of the variability in key outcomes accounted for by variations in individual, family, and school contexts. Such secondary data analyses will not, on their own, yield a definitive model of the sources of variation in student learning outcomes. They can, however, enhance our grasp of the issues and provide more practical guidelines for research designs. To illustrate, we consider variations in mathematics and science achievement status using data from the National Educational Longitudinal Study (NELS).[5] These data are used to explore the proportion of variance in 12th-grade students' mathematics and science achievement scores that is explained by factors operating at the individual as opposed to the school level. We include this example to demonstrate the influence of context on student achievement, not to present a definitive model of mathematics or science learning.

Context and Variation in Achievement: Examples From NELS

NELS was designed to "inform decision makers, education practitioners, and parents about the changes in the operation of the educational system over time" and to "study the relationships of various elements of the system with the educational and vocational development of the individuals who go through it" (U.S. Department of Education, 1996, p. 2).[6] We use these data to explore the impact of school and teacher characteristics on mathematics and science achievement for the subsample of 12th-grade students surveyed in the 1992 NELS second follow-up.

The NELS second follow-up surveyed students, parents, teachers, and school administrators and administered achievement tests to students participating in the study. Parent and student responses provide data on student characteristics that are likely to correlate with math and science achievement. We focus here on gender, race/ethnicity, family composition, number of siblings, family income, parent educational attainment, students' reports of the educational expectations that their parents had for them, and the mathematics and science course sequences that students followed. Responses from principals provided data on school context characteristics that previous research indicates are likely to be related with mathematics and science achievement; we focus here on the percentage of a school's graduating class who went on to a 4-year college (an indicator of academic press);[7] the number of colleges sending a representative to the school to talk with college-bound students during the school year; the composition of the student body; the percentage of a school's student body participating in a free or reduced-price school lunch program; and the extent to which absenteeism was a problem in the school (see Figure 1).[8]

Analyses and Findings Regarding Sources of Variation in Mathematics and Science Achievement

In ordinary least squares (OLS) regression analysis, the independence of observations is assumed. This assumption is violated by the hierarchical data structure of schools. As a result, analyses of variance based on OLS fail to take into account the clustering that occurs within schools and therefore the extent to which students in those clusters (entire schools or individual classrooms) are subject to common influences (see Bryk & Raudenbush, 1992). It is therefore inappropriate to use OLS to characterize the relationship between these explanatory variables and student achievement in mathematics and science. Instead, we report the results of hierarchical linear modeling (HLM) analyses using procedures specifically developed to deal with the clustering inherent in hierarchical data—that is, data pertaining to both students and schools (see Bryk & Raudenbush, 1992).

The full HLM model is presented in the appendix; we focus here on the within-school and between-school variance components presented in Table 1. Overall, the analyses confirm that at Level 1 (the individual level, controlling for other variables), both individual and family characteristics tend to have significant effects on student performance in mathematics and science at Grade 12. With respect to mathematics achievement, the net impact of the student and family characteristics described above is to reduce the total variance at Level 1 by 58%. These same student and family characteristics account for a smaller proportion (31%) of the total variance in science achievement at Level 1. Specifically, White males with well-educated parents whom the students report as having high educational expectations for them tend to have higher test scores than other students in both mathematics and science. Family income is significantly related to mathematics ($r = 0.31$) and science ($r = 0.28$) test scores; but the effect of income is not significant when other variables are included in the model, suggesting that family income has an indirect effect on academic performance. Important but not surprising, these results suggest that students who commit themselves to undertaking challenging academic work tend to perform better on the NELS mathematics and science achievement tests; previous enrollment in challenging (high-level) mathematics and science course sequences is strongly and positively related to mathematics and science achievement.

These results underscore the powerful impact that both individual- and school-level factors have on student achievement in mathematics and science. Across the two subjects, two of the most powerful effects are those associated with the composition of the student body that the school serves. Schools with larger percentages (more than 35%) of Black students enrolled in Grade 12 tend to have lower test scores in mathematics and science

INDIVIDUAL-LEVEL VARIABLES

Student Characteristics

- *Gender*
  Student is female
- *Race/ethnicity*[a]
  Student is Asian or Pacific Islander
  Student is Black
  Student is Hispanic, non-White
  Student is White (the omitted category)
- *Math course sequence* (level at Grade 12, from student transcript)
- *Science course sequence* (level at Grade 12, from student transcript)

Family Characteristics

- *Family composition* (at Grade 8)
- *Number of siblings* (at Grade 8)
- *Family income* (parent report at Grade 8)
- *Parent educational attainment* (highest level of parent education attained at Grade 10)
- *Parent educational expectations* (student report of parent educational expectation for their adolescents at Grade 10)

SCHOOL-LEVEL VARIABLES

- Percentage of *graduates going on to 4-year college*
- Number of *colleges sending a representative* to talk to college-bound students
- More than 35% of current Grade 12 students are Black, not of Hispanic origin
- Percentage of total student body participating in *free or reduced-price school lunch program*
- Extent to which *absenteeism* is a problem with students in this school

[a] Native Americans are excluded from analyses because of their small numbers in the sample.

FIGURE 1. *Individual- and school-level characteristics.*

Table 1
HLM Model for Math and Science Achievement, Grade 12: Variance Components

Variance Components	Math	Science
Unconditional		
Within school	109.88	21.51
Between school	53.24	10.55
Conditional		
Within school	46.40	14.84
Between school	22.72	5.28
Percentage of variance explained		
Within school	57.77	31.01
Between school	57.33	49.95

Note. Data are weighted to produce results generalizable to the population of U.S. high school students. The panel weight applies to sample members who were participants in 1988, 1990, and 1992. Sample size: Math (Level 1 = 7,138; Level 2 = 914); Science (Level 1 = 7,065; Level 2 = 909); both student-level and school-level variables are centered on grand-mean. Data are from the National Education Longitudinal Study of 1988–1994, National Center for Education Statistics, U.S. Department of Education.

than schools serving lower proportions of Black students in the 12th grade. Schools with larger percentages of students participating in free or reduced-price school lunch programs also tend to have lower mathematics and science test scores, a finding consistent with the literature documenting the important impact of school socioeconomic resources on the academic performance of the students they serve. Overall, approximately one third of the variation in these 12th-grade students' mathematics (32.9%) and science (35.6%) achievement scores are explained by factors operating at the school level (Level 2).

Clearly, dissecting the contributions of individual- and school-level characteristics is critical to understanding the variance in academic achievement. The analyses presented above do not demonstrate that these characteristics will necessarily have an impact on the generalizability of treatment effects. It is possible that an intervention is not context dependent and that its efficacy is not determined or influenced by context. However, the analyses do show that these variance components are relevant in planning studies of generalizability. If we expect an intervention to work equally well for all students, it is critical to examine its impacts in all those contexts which theory suggests and such secondary analyses confirm are typically associated with variation in academic achievement. If we suspect efficacy is context dependent, then repeated trials must assess impacts across contexts in order to document the extent to which context and intervention interact to affect student learning outcomes. It is with this understanding of the importance of context that we now turn our attention to designing effective scale-up research.

Designing Scale-Up Research: Practical Guidelines

Educational scale-up research is methodologically sophisticated and analytically complex. A key component of scale-up research is developing designs that could realistically be executed in field settings. There are three main concerns facing scale-up researchers: internal and external validity; statistical power and sample size; and methodological tools that individual researchers can use to produce robust, generalizable findings.

Internal and External Validity

The central objective of scale-up research in education is to provide evidence that the impacts of interventions documented in demonstration studies can be realized in dynamic and variable "real-world" contexts across student and teacher populations. The first task we face in forging the chain of evidence prescribed by the logic of scientific inquiry is to establish a cause-and-effect relationship between the intervention and the student learning outcome of interest. Importantly, we need not articulate the full, "true" cause of the changes in student achievement that we hope to bring about. That task may be philosophically as well as methodologically and analytically beyond our capabilities. However, to warrant the adoption of the intervention, we must provide convincing evidence that there is a sufficiently high probability it will yield the desired outcome. To do so, we must design and implement studies that are internally valid—that is, studies that rule out factors other than the intervention that could plausibly account for variations in the outcome variable.[9]

The major threats to internal validity are well documented in the literature, as are the elements of research design that can reduce or eliminate them.[10] One threat that has been a recent focus of concern in educational research is selection bias. Selection bias threatens internal validity by raising doubts about whether observable changes in outcomes of interest are attributable to differences in the subjects rather than to the treatment condition. Properly implemented, random assignment to treatment conditions removes this threat. The appropriate design for such studies is the randomized controlled trial (RCT).

The benefits of conducting RCTs and the challenges associated with them are discussed at length elsewhere (see, e.g., Campbell & Stanley, 1963; Shadish, Cook, & Campbell, 2002; Cook, 2001; Dignam, in press; Hempel, 1966; Myers & Dynarski, 2003; National Research Council, 2004). In the present context it is useful to recall two points. First are the concerns which have been noted regarding the resources required to conduct RCTs in social settings. Researchers and school administrators often find it necessary to devote considerable resources to overcome the practical obstacles to achieving random assignment of students and/or teachers to treatment conditions (e.g., within schools or classrooms). These resources are often scarce; yet the relatively high price of random assignment is typically justified on the basis of the public benefit of higher-quality evidence from efficacy and effectiveness trials, with no real expectation that these costs will be offset by short-term returns on the research-and-development investment (see Boruch, 2002, p. 39). Stated another way, the economy of public education (reinforced by the current accountability culture) encourages researchers to be parsimonious in planning the expenditure of funds that are made available to provide evidence of "what works."

Second, important design considerations need to be addressed if RCTs are to provide scalable evidence. Of special concern are the extent to which conditions in classrooms under study are representative of those in other schools, and assumptions regarding the likelihood that interventions will be enacted with similar fidelity in "real world" contexts. These are the types of challenges that must be addressed to ensure that research findings can be extrapolated to other contexts, that is, to assure a study's external validity.

Internal and external validity are complementary. Scale-up research, by definition, must yield both internally and externally valid findings. If scale-up research is not internally valid, then the researcher has failed to determine whether the intervention had the predicted effect. Similarly, if scale-up research is not externally valid, the researcher is unable to predict results for larger, more diverse populations, making it difficult to know whether the intervention should be taken to scale. Unfortunately, the challenge of designing studies that meet standards for both internal and external validity is frequently perceived as pitting the two against each other in a competition for scarce resources.

Statistical Power and Sample Size

Sample design is critically important in scale-up research, as it is the place where the optimal allocation of resources is decided for any given project. Methodologically, it is the stage at which we decide on the appropriate balance between the number of units to be studied at Level 2 (i.e., schools), and the number of elements within each unit to be studied at Level 1 (i.e., students). Operationally, it is the stage at which we must grapple with the higher costs typically associated with including more units at Level 2 in the study. So, for example, a budget that will support surveying, observing, and collecting assignments from 100 students in each of 10 schools is unlikely to support the same data collection strategy for 50 students in each of 20 schools. Researchers are likely to incur substantial human and financial costs in including additional schools in their studies, so that the actual cost equivalent may be between 1,000 students (100 in each of 10 schools) and 800 students (40 in each of 20 schools).

Clearly, data on the appropriate cost parameters must be factored into the sample design. We do not address those here, other than to underscore their significance. For a given set of resources we must decide what will be the optimal combination of units and elements nested within them to identify intervention effects. Statistical power calculations included in the design process can help scale-up researchers to balance the cost with the information content of a particular sample.

Underpowering is a common design problem in many fields (see Cuijpers, 2003; Dignam, in press; Halpern, Karlawish, & Berlin, 2002a, 2002b; Hughes, 2002; Lilford, 2002). Underpowered designs produce questionable findings because they fail to ensure an acceptable likelihood that differences in outcomes attributable to the treatment in question will be detected when they exist. Researchers who collect data from participants in studies with inadequate power run the risk that they will be unable to reject the null hypothesis (i.e., that the intervention does *not* yield the desired changes in student achievement) even when the null hypothesis is, in fact, false.

Methodological Tools for Generalizable Findings

The steps required to perform a power analysis are well documented for many frequently used education research designs, and as Hedberg, Santana, and Hedges note (2004, p. 2), "several investigators have developed software to assist in these computations (Borenstein, 2000; Elashoff, 2000; O'Brien, 1998)." Tools for addressing the special challenges of computing statistical power for multilevel studies of educational interventions are less widely available, although important strides are being made in this area. For example, the William T. Grant Foundation recently announced the establishment of a free consulting service created in collaboration with Stephen W. Raudenbush and his colleagues at the University of Michigan to provide advice on sample size determination, statistical power, and minimum detectable effect sizes to researchers and funders interested in the design of certain types of group-randomized intervention studies (William T. Grant Foundation, 2005).[11] Our understanding of the importance of contextual factors operating at the school level cautions us, however, against assuming that schools produce only random effects in multi-site (multi-school) randomized experiments. Consequently, power analyses for multi-site trials require some estimate of the extent to which variation in achievement differs across sites on factors (e.g., SES, urbanicity) that characterize the design in question. At its 2004 Annual Meeting, the American Educational Research Association's Division D (Measurement and Research Methodology) devoted an entire quantitative methods and statistical theory symposium to the importance of computing statistical power in multilevel studies of educational interventions. Of particular interest is the consideration of the implications for sample designs of these multilevel studies, notably the gains in statistical power that can be achieved by sample designs that increase the number of units at higher levels as opposed to those that increase the number of elements at lower levels (see, e.g., Konstantopoulos, 2004).

These advances notwithstanding, education researchers remain severely hampered in efforts to construct adequately powered designs, most notably the multilevel designs necessary to monitor the effects of scale-up. A primary problem in constructing such designs is the lack of "definitive prior information" concerning the variance structure of academic achievement (Hedberg, Santana, & Hedges, 2004, p. 2). The necessity of designing research that distinguishes the impact of interventions from the possible interactions between the interventions and the social context in which they occur is clear. Conceptually it is not difficult to see how information on the proportion of variance attributable to Level-1 (individual) and Level-2 (school) factors should be applied to the design of scale-up research. Studies should be designed that maximize variance on the independent variables of interest, controlling for the key contextual variables and/or replicating studies in various contexts in ways that allow evidence from multiple studies to be accumulated over time. The design implications are also relatively straightforward. Indeed, it is reasonable to expect that a graduate student armed with a copy of Campbell and Stanley (1963), Galtung (1969), Shadish, Cook, and Campbell (2002), or any of a number of other widely available methods texts would be able to sketch out a study designed to control and manipulate key variables appropriately.

Often, however, there are large differences between what it is possible to do "in principle" (e.g., by using well-developed and widely accepted principles of good design, sampling, and analysis) to devise studies capable of yielding robust, generalizable findings and what it is possible for individual researchers to accomplish in practice. Many scale-up researchers do not have information on the precise nature of within- and between-school variation—the intraclass correlations required to calculate the statistical power associated with optimal multilevel designs.[12] Hedges and his colleagues are currently engaged in a program of IERI-funded research to address this need. Specifically, they are reanalyzing data obtained from major national probability sample surveys, a process that involves computing "between-school and within-school variance components . . . for the nation as a whole separately for mathematics and reading achievement, and separately (by subject matter) for different grade levels, regions of the country and urbanicity of settings" (Hedberg, Santana, & Hedges, 2004, pp. 3, 5).[13] The results echo the findings presented earlier in the NELS analysis regarding the significance of Level-2 contextual variables on student achievement. They clearly demonstrate that a significant proportion of the variation in student achievement occurs at the school level. Moreover, they indicate that the amount of variation differs significantly based on context (e.g., by region of the country, urbanicity, and students' stage in their own life-course). Ultimately, Hedges and his colleagues plan to compile these coefficients and make them available in the form of a variance almanac through IERI's research and technical center, the Data Research and Development Center (*http://drdc.uchicago.edu*).

When the variance almanac is published, researchers designing studies will be able to use it to ascertain optimal design parameters. Researchers want to target their populations as accurately as possible, not spend time and money investigating between-school effects when the changes are really at the individual level or vice-versa. This is one solution to the cost dilemma—using the intraclass correlations to design studies that reach the target population most accurately, thereby generating robust results.

Scale-Up in Practice

A rigorous approach to scale-up research is critical in creating the evidence base needed to improve student achievement through proven interventions. Such research involves the use of randomized controlled trials with sufficient power to provide robust results. Analyses of the complex and multilevel nature of individual, family, and school effects on student achievement require advanced statistical tools and a thorough understanding of the sociological research on schooling. In addition, such analyses require close collaboration with practitioners while researchers test interventions in diverse school settings. The results should help educators not only predict the likely benefit of an intervention but also provide guidance regarding the possible modifications required in other contexts.

The aim of scale-up research is not to prescribe a course of action for all schools. Scale-up is not a euphemism for the uncritical diffusion of interventions shown to have a positive impact on student learning outcomes in one setting to different teacher and student populations in diverse and dynamic circumstances. Nor is scale-up a veiled argument for the deterministic use of research findings: assigning education researchers the role of determining what works with the expectation that teachers and principals will then replicate those interventions. To the contrary, scale-up research is doomed to fail if practitioners and policymakers expect it to generate absolute solutions to the nested, multifaceted, and often mutually reinforcing sets of social problems that contribute to low achievement scores. A context-based approach to scale-up research provides the evidence that educators need to select the interventions that are most likely to work in specific settings.

NOTES

This article is based upon work supported by the National Science Foundation under Grant No. 0129365. The opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors thank Fengbin Cheng for his assistance in the preparation of the manuscript. We also thank Larry Hedges, Lisa Hoogstra, and two anonymous reviewers for their helpful comments on early drafts.

[1] Jointly and separately, policymakers and researchers have funded studies, convened panels, and issued reports that examine the nature of scientific inquiry in education, work to advance understanding of how to address education problems scientifically and ensure that the resulting knowledge accumulates, consider the role of the federal government in fostering and supporting such research, and provide guidance to practitioners seeking to identify and implement educational practices supported by evidence from rigorous, scientifically based research (see, e.g., National Research Council, 2002, 2005; Coalition for Evidence-Based Policy, 2003; Learning Point Associates, 2004; What Works Clearinghouse, 2004).

[2] These reforms have many parallels, both in other fields and in education in other countries. See, e.g., the Cochrane Collaboration (*http://www.cochrane.org/index0.htm*); the Campbell Collaboration (*http://www.campbellcollaboration.org/*); the UK Economic and Social Research Council (ESRC) Centre for Evidence Based Policy and Practice (*http://www.evidencenetwork.org*); the Evidence for Policy and Practice Information Coordinating Centre (EPPI-Centre, *http://eppi.ioe.ac.uk*); the Research Evidence in Education Library (REEL), "home for the Centre for Evidence-Informed Policy and Practice in Education" (*http://eppi.ioe.ac.uk/EPPIWeb/home.aspx?page=/reel/intro.htm*); Oxford University's Centre for Evidence-Based Medicine (*http://www.cebm.net/*); the New York Academy of Medicine and Evidence-based Medicine Committee of the American College of Physicians, New York Chapter's Evidence-Based Medicine Resource Center (*http://www.ebmny.org/*); the series of conferences entitled Evidence-Based Policies and Indicator Systems, organized by the Curriculum, Evaluation, and Management Centre at the University of Durham, England (*http://www.cemcentre.org/ebeuk/*); and "www.nettingtheevidence.org.uk: A ScHARR Introduction to Evidence Based Practice on the Internet" (*http://www.shef.ac.uk/scharr/ir/netting/*).

[3] See National Science Foundation 1999, 2000, 2001, 2002, 2004; U.S. Department of Education 2003, 2004a, 2004b, 2004c.

[4] See also Schneider (2003), Dreeben (1994), Bidwell and Friedkin (1988), and Levinson, Cookson, and Sadovnik (2001), for detailed reviews of the history and key findings of the field of sociology of education from the turn of the last century.

[5] It is important to underscore that we consider here achievement status as measured in NELS 12th-grade math and science achievement tests, not achievement gains; for a discussion of the benefits of using such status measures (e.g., they reflect cumulative effects arising from key independent variables), see Nye, Konstantopoulos, and Hedges (2004).

[6] Additional information on the National Educational Longitudinal Study of 1988, including questionnaires, publications, and products, is available online at *http://nces.ed.gov/surveys/nels88/.*

[7] Schools in which teachers and students share a normative commitment to academic achievement are described as pressing for academic success (i.e., exhibiting higher levels of academic press than those that lack the same commitment); see, e.g., McDill, Natriello, and Pallas (1986); Shouse (1996, 1997); Lee and Smith (1999).

[8] A full description of variables used in the analysis, including means and standard deviations, is available from the authors on request.

[9] The context-dependent nature of causality was underscored by Campbell, who proposed relabeling internal validity to emphasize this point, among others (see Campbell, 1986). As described in Shadish, Cook, and Campbell (2002, p. 54), Campbell included the word *local*

in the new nomenclature ("local molar causal validity") in order to emphasize "that causal conclusions are limited to the context of the particular treatments, outcomes, times, settings, and persons studied."

[10] Threats to internal validity include ambiguous temporal precedence, selection, history, maturation, regression, attrition, testing, instrumentation, and the additive and interactive effects of threats to internal validity. For descriptions of each threat and recommendations regarding research designs that mitigate or eliminate such effects, see Shadish, Cook, and Campbell (2002).

[11] This consultancy service is provided as part of a larger special initiative of the foundation entitled Capacity Building for Group-Randomized Studies, co-directed by Raudenbush and Bloom. Other resources available from the foundation that are relevant to the conduct of group-randomized trials include the od.exe software (developed by Raudenbush, Liu, and Congdon), which "enables researchers to determine sample size, power, and optimal allocation of resources for group-randomized studies." For additional information, see William T. Grant Foundation (2005).

[12] Intraclass correlations tell us how much within-school variation exists and how much between-school variation exists. If there is considerable within-school variation, this indicates that individual-level factors are more important influences on student learning. If there is substantial between-school variation and relatively little within-school variation, this indicates that school-level factors have the predominant impact.

[13] Major data sets being analyzed include the Early Childhood Longitudinal Study (ECLS), Prospects, the National Educational Longitudinal Study (NELS), and the National Assessment of Educational Progress (NAEP).

REFERENCES

Bidwell, C. E., & Friedkin, N. E. (1988). Sociology of education. In N. J. Smelser (Ed.), *Handbook of sociology* (pp. 449–471). Newbury Park, CA: Sage.

Blumenfield, P., Fishman, B. J., Krajcik, J., & Marx, R. W. (2000). Creating usable innovations in systemic reform: Scaling up technology-embedded project-based science in urban schools. *Educational Psychologist, 35*(3), 149–164.

Bodilly, S., Keltner, B., Purnell, S., Reichardt, R., & Schuyler, G. (1998). *Lessons from New American Schools' scale-up phase.* Santa Monica, CA: RAND.

Borenstein, M. (2000). *Power and precision 2.0.* Englewood, NJ: Biostatistical Programming Associates.

Boruch, R. (2002). The virtues of randomness. *Education Next, 2*(3), 37–41.

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

Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. In W. M. K. Trochim (Ed.), *Advances in quasi-experimental design and analysis* (pp. 67–77). San Francisco, CA: Jossey-Bass.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Coalition for Evidence-Based Policy. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user-friendly guide* (NCEE 2004-3000). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Coburn, C. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational Researcher, 32*(6), 3–12.

Conley, J. G., & Wolcott, R. C. (in press). Scaling from prototype to production: A managed process perspective from industrial engineering and management. In B. Schneider & S.-K. McDonald (Eds.), *Scale-up education: Vol. 1. Ideas in principle.* Lanham, MD: Rowman & Littlefield.

Cook, T. D. (2001). A critical appraisal of the case against using experiments to assess school (or community) effects. *Education Next.* Retrieved February 11, 2005, from *http://www.educationnext.org/unabridged/20013/cook.html*

Cook, T. D. (2003). Why have educational evaluators chosen not to do randomized experiments? *Annals of American Academy of Political and Social Science, 589,* 114–149.

Corcoran., T. (n.d.). Scaling-up best practices in education: Building a more responsible profession. In *For our schools, for our children: Excellence and opportunity* (pp. 5–15). College Park, MD: James MacGregor Burns Academy of Leadership, University of Maryland. Retrieved July 12, 2005, from *http://www.academy.umd.edu/publications/leadership_publicpolicy/Education/Educ-FINAL.pdf*

Cuijpers, P. (2003). Examining the effects of prevention programs on the incidence of new cases of mental disorders: The lack of statistical power. *American Journal of Psychiatry, 160*(8), 1385–1391.

Dignam, J. (in press). From efficacy to effectiveness: Translating randomized controlled trial findings into treatment standards in the health professions. In B. Schneider & S.-K. McDonald (Eds.), *Scale-up in education: Vol. 1. Ideas in principle.* Lanham, MD: Rowman & Littlefield.

Dreeben, R. (1994). The sociology of education: Its development in the United States. In A. M. Pallas (Ed.), *Research in sociology of education and socialization* (Vol. 10, pp. 7–52). Greenwich, CT: JAI Press.

Education Sciences Reform Act of 2002, Pub. L. No. 107-279, 116 Stat. 1940 (2002).

Elashoff, J. D. (2000). *nQuery Advisor 4.0.* Saugus, MA: Statistical Solutions.

Elmore, R. (1996). Getting to scale with good educational practice. *Harvard Educational Review, 66*(1), 1–26.

Flamholtz, E. G., & Randle, Y. (in press). Measuring and managing successful organizational scale-up. In B. Schneider & S.-K. McDonald (Eds.), *Scale-up in education: Vol. 1. Ideas in principle.* Lanham, MD: Rowman & Littlefield.

Foorman, B. R., Santi, K. L., & Berger, L. (in press). Scaling assessment-driven instruction using the Internet and handheld computers. In B. Schneider & S.-K. McDonald (Eds.), *Scale-up in education: Vol. 2. Issues in practice.* Lanham, MD: Rowman & Littlefield.

Galtung, J. (1969). *Theory and methods of social research* (rev. ed.). New York: Columbia University Press.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Edward Arnold.

Hallinan, M. T. (Ed.). (2000). *Handbook of the sociology of education.* New York: Kluwer Academic/Plenum.

Halpern, S. D., Karlawish, J. H., & Berlin, J. A. (2002a). The continuing unethical conduct of underpowered clinical trials. *Journal of the American Medical Association, 288,* 358–362.

Halpern, S. D., Karlawish, J. H., & Berlin, J. A. (2002b). The ethics of underpowered clinical trials: Letters to the editor—In reply. *Journal of the American Medical Association, 288,* 2118–2119.

Hassel, B., & Steiner, L. (2000). *Strategies for scale: Learning from two educational innovations* (Occasional Paper 1-00). Cambridge, MA: Harvard University, John F. Kennedy School of Government.

Hedberg, E. C., Santana, R., & Hedges, L. V. (2004, April). *The variance structure of academic achievement in America.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Hempel, C. G. (1966). *Philosophy of natural science.* Englewood Cliffs, NJ: Prentice-Hall.

Hughes, J. R. (2002). The ethics of underpowered clinical trials: Letter to the editor. *Journal of the American Medical Association, 288,* 2118.

Konstantopoulos, S. (2004, April). *Planning sample sizes for multilevel studies of educational interventions.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Learning Point Associates. (2004). *Quick key No. 7: Understanding the No Child Left Behind Act of 2001: Scientifically based research.* Naperville, IL: Learning Point Associates. Retrieved March 21, 2006, from *http://www.ncrel.org/csri/tools/qkey7/science.htm*

Lee, V. E., & Smith, J. B. (1999). Social support and achievement for young adolescents in Chicago: The role of school academic press. *American Educational Research Journal, 36,* 907–945.

Levin, M. (Ed.). (2001). *Pharmaceutical process scale-up.* Marcel-Dekker.

Levinson, D., Cookson, P. W. Jr., & Sadovnik, A. R. (Eds.). (2001). *Education and sociology: An encyclopedia.* New York: Routledge Falmer.

Lilford, R. J. (2002). The ethics of underpowered clinical trials: Letter to the editor. *Journal of the American Medical Association, 288,* 2118.

McDermott, K. (2000). Barriers to large-scale success of models for urban school reform. *Educational Evaluation and Policy Analysis, 22*(1), 83–89.

McDill, E. L., Natriello, G., & Pallas, A. (1986). A population at risk: Potential consequences of tougher school standards for student dropouts. *American Journal of Education, 94,* 135–181.

McDonald, S.-K., & Schneider, B. (2004, April). *Conceptualizing scale-up: Multidisciplinary perspectives.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Myers, D., & Dynarski, M. (2003). *Random assignment in program evaluation and intervention research: Questions and answers.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

National Research Council. (2002). *Scientific research in education.* Washington, DC: National Academy Press.

National Research Council. (2004). *Implementing randomized field trials in education: Report of a workshop.* Washington, DC: National Academy Press.

National Research Council. (2005). *Advancing scientific research in education.* Washington, DC: The National Academies Press.

National Science Foundation. (1999, February 16). *Interagency Education Research Initiative (IERI), program announcement, NSF 99-84.* Retrieved December 31, 2004, from *http://www.nsf.gov/pubsys/ods/getpub.cfm?nsf9984*

National Science Foundation. (2000, March 7). *Interagency Education Research Initiative (IERI), program solicitation, NSF 00-74.* Retrieved December 31, 2004, from *http://www.nsf.gov/pubsys/ods/getpub.cfm?nsf0074*

National Science Foundation. (2001, April 20). *Interagency Education Research Initiative—revised version (IERI), program solicitation, NSF 01-92.* Retrieved December 31, 2004, from *http://www.nsf.gov/pubs/2001/nsf0192/nsf0192.htm*

National Science Foundation, Directorate for Education and Human Resources, Division of Research, Evaluation and Communication. (2002, February 14). *Interagency Education Research Initiative (FY2002) (IERI), program solicitation, NSF-02-062.* Retrieved December 31, 2004, from *http://www.nsf.gov/pubsys/ods/getpub.cfm?nsf02062*

National Science Foundation, Directorate for Education and Human Resources, Division of Research, Evaluation and Communication. (2004, April 1). *Interagency Education Research Initiative (IERI) program solicitation,* NSF 04-553. Retrieved December 29, 2004, from *http://www.nsf.gov/pubs/2004/nsf04553/nsf04553.htm*

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002). Retrieved December 29, 2004, from *http://www.ed.gov/policy/elsec/leg/esea02/107-110.pdf*

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

O'Brien, R. G. (1998). *UnifyPow: A SAS macro for sample-size analysis.* Cleveland, OH: Author. Retrieved March 21, 2006, from *http://www2.sas.com/proceedings/sugi22/STATS/PAPER287.PDF*

Palakodaty, S., Walker, S., Townend, G., York, P., & Humphreys, G. (2000, August). Scale-up and GMP Plant design—Pharmaceutical particle engineering by the SEDS process. *European Pharmaceutical Contractor.* Retrieved July 12, 2005, from *http://www.nektar.com/pdf/gmp_plant_design.pdf*

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20*(4), 307–335.

Royal Society of Chemistry. Environment, Health and Safety Committee. (1999). *Safety issues in the scale up of chemical reactions, version 1/3/99.* London: Royal Society of Chemistry. Retrieved July 12, 2005, from *http://www.rsc.org/pdf/ehsc/scaleup.pdf*

Schneider, B. (2003). Sociology of education: An overview of the field at the turn of the twenty-first century. In M. T. Hallinan, A. Gamoran, W. Kubitschek, & T. Loveless (Eds.), *Stability and change in American education: Structure, process, and outcomes* (pp. 193–226). Clinton Corners, NY: Eliot Werner Publications.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Shouse, R. (1996). Academic press and sense of community: Conflict and congruence in American high schools. In A. M. Pallas (Ed.), *Research in sociology of education and socialization* (pp. 173–202). Greenwich, CT: JAI Press.

Shouse, R. (1997). Academic press, sense of community, and student achievement. In J. S. Coleman, B. Schneider, S. Plan, K. Schiller, R. Shouse, & H. Wang (Eds.), *Redesigning American education* (pp. 60–86). Boulder, CO: Westview Press.

Slavin, R. E. (2002, November). Remarks offered as a member of the Panel on Scaling at the Interagency Education Research Initiative (IERI) Annual Principal Investigators Meeting. Washington, DC.

U.S. Department of Education, Institute of Education Sciences. (2003). *Interagency Education Research Initiative, request for applications, NCER-03-04.* Washington, DC: U.S. Department of Education.

U.S. Department of Education, Institute of Education Sciences. (2004a). *Teacher quality research grants, request for applications, NCER-04-02.* Washington, DC: U.S. Department of Education.

U.S. Department of Education, Institute of Education Sciences. (2004b). *Mathematics and science education research grants, request for applications, NCER-04-03.* Washington, DC: U.S. Department of Education.

U.S. Department of Education, Institute of Education Sciences. (2004c). *Reading comprehension and reading scale-up research grants, request for applications, NCER-04-04.* Washington, DC: U.S. Department of Education.

U.S. Department of Education, National Center for Education Statistics. (1996). *National Education Longitudinal Study of 1988 (NELS:88) research framework and issues* (Working Paper No. 96-03). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

U.S. Office of Science and Technology Policy. (n.d.). *Interagency education research initiative.* Retrieved December, 31, 2004, from *http://clinton2.nara.gov/WH/EOP/OSTP/Science/html/ieri.html*

What Works Clearinghouse. (2004). *What Works Clearinghouse: A trusted source of scientific evidence of what works in education.* Retrieved December 29, 2004, from *http://www.whatworks.ed.gov/*

William T. Grant Foundation. (2005). *Consultation on design of group-randomized studies.* Retrieved February 9, 2005, from *http://www.wtgrantfoundation.org/info-url_nocat3040/info-url_nocat_show.htm?doc_id=227403&attrib_id=9485*

York, P. (1999). Strategies for particle design using supercritical fluid technologies. *PSTT, 2*(11), 430–440. Retrieved July 12, 2005, from *http://www.nektar.com/pdf/scf_strategies.pdf*

Zlokarnik, M. (2002). *Scale-up in chemical engineering.* Weinheim, Germany: Wiley-VCH Verlag GmbH & Co.

AUTHORS

SARAH-KATHRYN McDONALD is the Executive Director of the Data Research and Development Center and a Senior Research Scientist with the National Opinion Research Center at the University of Chicago, 1155 East 60th Street, Room 277, Chicago, IL 60637; *mcdonald-sarah@norc.uchicago.edu.* Her areas of special interest include evidence-based decision making and the diffusion of innovation.

VENESSA ANN KEESLER is a doctoral student in the Measurement and Qualitative Methods Program at Michigan State University, 516C Erickson Hall, East Lansing, MI 48824; *rosevene@msu.edu.* Her areas of special interest include the sociology of education, work and family studies, and quantitative methodology.

NILS J. KAUFFMAN is a doctoral student at Michigan State University, 222 North Francis Avenue, Lansing, MI 48912; *kauffm35@msu.edu.* His areas of special interest include issues of policy and practice in education, focusing on the changes taking place in Eastern Europe.

BARBARA SCHNEIDER is the John A. Hannah Chair University Distinguished Professor in the College of Education and the Sociology Department at Michigan State University, 516 Erickson Hall, East Lansing, MI 48824; *bschneid@msu.edu.* Her areas of special interest include the social organization of schooling and knowledge accumulation.

Manuscript received March 1, 2005
Revision received October 13, 2005
Accepted January 3, 2006

APPENDIX

HLM Model for Math and Science Achievement, Grade 12

Characteristics	Math	Science
Individual level (Level 1)		
Intercept	51.547***	24.489***
Female	−1.790***	−1.832***
Asian	−1.432*	−1.209***
Hispanic	−3.672***	−2.070***
Black	−4.861***	−4.126***
Nontraditional family	−.079	−.016
Number of siblings	.242*	.045
Family income (natural logarithm)	.252	.126
Parent educational attainment	.783***	.630***
Parent educational expectations	1.153***	.817***
Math course sequence	3.561***	
Science course sequence		1.136***
School level (Level 2)		
Percentage of graduates going to 4-year college	−.003	−.002
Number of colleges sending representative to school	.003	.003
More than 35% of students are Black	−2.823***	−1.032**
Percentage of total students in free lunch program	−.025*	−.015**
Absenteeism is a problem in school	−.109	−.247
Variance components		
Unconditional		
Within school	109.88	21.51
Between school	53.24	10.55
Conditional		
Within school	46.40	14.84
Between school	22.72	5.28
Percentage of variance explained		
Within school	57.77	31.01
Between school	57.33	49.95

Note. Data are weighted to produce results generalizable to the population of U.S. high school students. The weight applies to 1992 sample members for whom transcript data are available. Sample size: Math (Level 1 = 7,138; Level 2 = 914); Science (Level 1 = 7,065; Level 2 = 909). Both student-level and school-level variables are centered on grand-mean. Data are from the National Education Longitudinal Study of 1988–1994, National Center for Education Statistics, U.S. Department of Education.

* $p < .05$; ** $p < .01$; *** $p < .001$ (two-tailed tests).