Contract No.: ED-01-CO-0039 (0004)
MPR Reference No.: 8936-600

MATHEMATICA
Policy Research, Inc.

The Effectiveness Of Educational Technology: Issues and Recommendations for the National Study

Draft

May 9, 2003

Roberto Agodini
Mark Dynarski
Margaret Honey, Education Development Center
Douglas Levin, American Institutes for Research

Submitted to:

Institute of Education Sciences
U.S. Department of Education
80 F Street NW
Washington, DC 20208

Project Officer:
Audrey Pendleton

Submitted by:

Mathematica Policy Research, Inc.
P.O. Box 2393
Princeton, NJ 08543-2393
Telephone: (609) 799-3535
Facsimile: (609) 799-0005

Project Director:

CONTENTS *(continued)*

Mark Dynarski

TABLE OF CONTENTS

LIST OF FIGURES

EXECUTIVE SUMMARY

THE EFFECTIVENESS OF EDUCATIONAL TECHNOLOGY: ISSUES AND RECOMMENDATIONS FOR THE NATIONAL STUDY

The No Child Left Behind Act (P.L. 107-110) called for the U.S. Department of Education to carry out a national study of the effectiveness of educational technology. With computers becoming ubiquitous in American schools, and purchases of hardware and software now substantial expenses for school districts, whether funding is supporting effective uses of technology and whether spending can be more effective have become concerns. The legislation's mandate called for the study to use rigorous methods to provide evidence of effectiveness.

In October 2002, the U.S. Department of Education began working with Mathematica Policy Research, Inc. and its partners, the American Institutes for Research and the Education Development Center, to identify issues confronting a national study of technology effectiveness and to develop designs for the study. A key part of the design effort was to engage a panel of outside experts on educational technology, educational policy, and research methodology, to help identify important questions to be addressed in the study and to suggest possible approaches for answering them.

The design team worked with the advisory panel and with ED staff to arrive at nine recommendations for how the national study could focus its attention (see box, next page). The panel played an important role in suggesting issues and approaches, and in discussing the strengths and weaknesses of various approaches, but did not formally make recommendations. The key broad question to be addressed by the evaluation is "Is educational technology effective in improving student academic achievement?" The design team recognized that, stated in this way, no single study could answer the question. In effect, many questions are implied, related to alternative definitions of education technology, effectiveness, and improving student achievement. The team needed to define what is meant by "educational technology," "effective," and so on.

The design team's recommendations refine the study, so it can have the potential to contribute substantially to what is known about the effectiveness of educational technology. The recommendations focus attention on technology applications that support instruction in reading or math in low-income schools serving the K-12 grade levels. The study would use experimental designs (with random assignment of students, classrooms, or schools, depending on the type of technology application) to ensure that measured effects can be attributed to the technology applications. The key outcome would be scores from a commonly used standardized test, supported by other academic outcomes collected from extant data. The report provides rationales for the recommendations and discusses

Design Team Recommendations for a National Study of the Effectiveness of Educational Technology

Question: What is "educational technology?"

Recommendation 1: Examine technology applications designed to support teaching and learning.

Recommendation 2: Use a public submission process to identify technology applications to study.

Question: What is "effective?"

Recommendation 3: Use experimental designs to measure effects.

Recommendation 4: Study the effects of technology applications for schools or teachers that do not currently use the applications but are interested in using them.

Recommendation 5: Design the study to detect "moderate" to "large" effects of technology applications.

Question: What kinds of students?

Recommendation 6: Study the effects of technology applications for students in the primary and secondary grade levels (K-12).

Recommendation 7: Study the effects of technology applications for schools that receive Title I funds.

Question: What is "academic achievement?"

Recommendation 8: Study the effects of technology applications on student academic achievement as measured by commonly used standardized tests, and collect data on other academic indicators to provide a fuller picture.

Recommendation 9: Study the effects of technology applications that support instruction in reading and math.

conceptual frameworks and statistical issues related to measuring effects and determining sample sizes.

The legislative mandate for the study calls for going beyond the question of effectiveness and asking about the conditions and practices related to effectiveness. The design team and the advisory panel discussed conditions and practices that could be related to effectiveness. There was broad agreement that teacher training is potentially important, as prior research had noted its relationship to the effectiveness of technology applications. Other factors that could be studied included characteristics of students, their parents, their teachers, their classroom, their school, their district, and their neighborhood. The design team recommends using statistical modeling techniques to estimate a set of relationships between various conditions and practices, and the effectiveness of technology applications.

The design team also considered how to select technology applications for the national study and how to recruit school districts and schools to be part of the study. The team recommended a public submission process: developers of technology applications would provide information for a review panel to consider in identifying promising applications for the study. The information would include general characteristics of the application and its users, as well as prior evidence of its effectiveness. The design team considered it important that a wide range of technology applications was considered for the national study, which would be facilitated by a public process.

Recruiting school districts and schools will be an important challenge for the national study. The team recommended that the study seek out school districts and schools that are interested in using the selected technology applications but do not already do so. Selecting schools that want to use the applications would be one aspect of ensuring that the study examined strong implementations of the technology applications included in the study. Other aspects contributing to strong implementations would be to ensure that schools devote adequate resources to train teachers on how to use the applications appropriately and to provide ongoing support to teachers throughout the school year to respond to questions or problems. Issues of how best to support implementation need to be considered further as the national study unfolds.

CHAPTER I

STUDYING THE EFFECTIVENESS OF EDUCATIONAL TECHNOLOGY

Today nearly every school and a rapidly growing number of classrooms have computers and internet connections, and student-to-computer ratios are reaching levels that permit sustained instructional use of computers. Some research has identified a correlation between teachers' technology skills, their use of technology in classroom instruction, and higher academic achievement; but most researchers agree that a small minority of teachers use computers as part of their instruction in academic subjects with sufficient frequency or skill to improve student achievement.

Like other areas of education, more rigorous research using scientific methods is needed to determine the effectiveness of these tools in improving instruction and student achievement. After years of significant investments made in educational technology, policymakers and budget decision makers are demanding evidence that the investments are improving instruction and student achievement. The No Child Left Behind Act (NCLB), signed January 8, 2002, called for the U.S. Department of Education (ED) to carry out a national study of technology effectiveness and provided questions to be addressed by the study (see box, next page).

In October 2002, ED began working with Mathematica Policy Research, Inc., and its partners, the American Institutes for Research and the Education Development Center, to identify issues of interest for a national study of technology effectiveness and to develop designs for the study. A major part of the design effort was to engage a panel of outside experts on educational technology, educational policy, and research methodology to help identify important questions to be addressed in the study and to suggest possible approaches for answering them. The design team worked with the advisory panel and with ED staff to develop recommendations for how the national study could focus its attention.

From *No Child Left Behind (P.L. 107-110)*

SEC. 2421. NATIONAL ACTIVITIES.

(a) STUDY- Using funds made available under section 2404(b)(2), the Secretary —

(1) shall conduct an independent, long-term study, utilizing scientifically based research methods and control groups or control conditions —

(A) on the conditions and practices under which educational technology is effective in increasing student academic achievement; and

(B) on the conditions and practices that increase the ability of teachers to integrate technology effectively into curricula and instruction, that enhance the learning environment and opportunities, and that increase student academic achievement, including technology literacy;

(2) shall establish an independent review panel to advise the Secretary on methodological and other issues that arise in conducting the long-term study;

(3) shall consult with other interested Federal departments or agencies, State and local educational practitioners and policymakers (including teachers, principals, and superintendents), and experts in technology, regarding the study; and

(4) shall submit to Congress interim reports, when appropriate, and a final report, to be submitted no later than April 1, 2006, on the findings of the study.

Between November 2002 and February 2003, the advisory panel met three times at the Institute for Education Sciences offices in Washington, DC.[1]

The central framing question embodied in the legislation is: Is educational technology effective in improving student academic achievement? The legislation notes that the study should examine the "conditions and practices" under which technology is effective, but the question of whether technology is effective logically comes before questions of the conditions and practices under which it is effective. Consideration of the approaches for answering the question required the team and the advisory panel to narrow each aspect of the question: the meaning of "educational technology," the meaning of "effective," the types

[1] Appendix A lists the names of panel members. Transcripts of the meetings are available on request.

of "students" for whom effectiveness would be best studied, and the outcomes implied by "academic achievement." Narrowing the study is necessary because no one study—no matter its scale—can answer the central framing question in all its possible dimensions. Narrowing the study also is a useful step for considering approaches for studying conditions and practices and for considering operational aspects about how a study could be carried out in the complex settings of school districts and schools.

A. What Is "Educational Technology"?

Schools have long used tools as part of instruction and learning. For example, modern classrooms often contain textbooks, televisions and videocassette players, and computers, and many secondary schools have classrooms that contain equipment used in commercial and manufacturing settings. All of these tools could be considered broadly as forms of "educational technology."[2]

For the purposes of the study, however, the definition of technology is usefully narrowed to computers and, in particular, to technology applications that are intended to support teaching and learning. Under this definition, the study would focus on technology applications regardless of whether the applications are accessed through various hardware devices such as desktop computers, monitors connected to a server, hand-held devices, or smart keyboards. A focus on technology applications offers two advantages. First, application expenditures are a significant component of overall spending on educational technology, and a study focusing on applications would provide evidence related to the large share of spending on applications. Second, given the rapid improvements in hardware, a study of effects based on hardware could become obsolete in a few years. A study based on technology applications, however, would yield information with more lasting usefulness.

Recommendation 1: Examine technology applications designed to support teaching and learning.

To identify technology applications for the study, the design team recommends the use of a public submission process, in which technology application developers provide information about the particular application, including any prior evidence of its effectiveness. A panel of qualified reviewers then would assess the submissions and make recommendations about technology applications for inclusion in the national study. A public submissions process would help ensure that the study has access to information about a wide range of technology applications and that application developers and interested members of the public

Recommendation 2: Use a public submission process to identify technology applications to study.

[2] School districts also use computer "technology" for management and administrative functions such as recordkeeping and accounting. The design team considered such uses of technology to be outside the purview of the study.

understand the criteria for selection. The public aspect of the process also ensures that the process of being considered for the national study is open to any developer.

B. What Is "Effective"?

The mandate of the national study calls for the use of scientifically based methods and a "control group or control condition" to study technology's effectiveness. The use of these terms is consistent with the application of experimental designs to study whether technology applications are effective, with random assignment used to create a "treatment" group that has access to the technology application and a "control" group that does not have such access. When experimental designs are used, differences in outcomes between the two groups can be interpreted as causally related to the technology application. Designs with this causal property sometimes are said to have strong "internal validity," meaning that the differences in outcomes are caused only by the program or intervention under study and not by other factors.

Recommendation 3: Use experimental designs to measure effects.

An issue considered by the design team and the advisory panel was the settings in which experimental designs should be used. At one end of the spectrum, technology applications could be studied, for example, in a tightly controlled laboratory setting, in which students are exposed to the application and their outcomes measured. However, the design team recognized that evidence about effectiveness in such a setting might bear little relationship to the actual uses of the technology application in districts and schools. At the other end of the spectrum, technology applications could be studied in loosely controlled school settings, in which the applications already were being used (though possibly not according to designer intentions). In that circumstance, the design team recognized that the study would need to create a control group that would be denied access to a technology application that was already being used, creating a situation that many schools would find unacceptable.

The spectrum relates to the extent to which the national study should focus on either the efficacy of technology applications or the effectiveness of the applications. Studying the applications in laboratory-like settings is asking whether technology applications *can* improve student outcomes (efficacy). Studying the technology applications as they actually are used is asking whether they *do* improve student outcomes (effectiveness). The distinction is common in medical research, in which the early stages of studying a new medical treatment or intervention involve a determination of whether the treatment can have effects, and later, after efficacy is established, whether it can be effective under conditions that correspond to actual practice. However, actual practice for technology applications could include studying situations in which teachers did not use the application in instructionally useful ways. Anecdotes about computers sitting in closets or never being turned on suggest that technology can be purchased but not necessarily used well.

After consideration, the team recommends an intermediate strategy: the study should focus on the effects of technology applications for schools or teachers that do not currently use the applications but are interested in doing so. The crucial issue is that the schools and teachers should be interested in using the technology application. The team concluded that it would make sense to ask schools and teachers to test a technology application that *may be* superior to what they currently use but about which more evidence is needed.

Recommendation 4: Study the effects of technology applications for schools or teachers who do not currently use the applications but are interested in doing so.

The strategy combines elements of efficacy and effectiveness. It includes elements of efficacy by including schools and teachers interested in using the technology application. It includes elements of effectiveness in that it would study technology applications in the real settings of districts and schools rather than in laboratory settings.

Another aspect of efficacy the national study will need to consider is that schools and teachers should use the technology applications of interest in a manner that is consistent with the designers' intention. Of course, deviations from designer intention arise naturally in the implementation of programs or approaches in the complex settings of real schools. However, the design team believes that the national study should work to provide schools and teachers with adequate resources and support, to ensure that technology applications are implemented in a manner their designers would regard as constituting a fair test of the application in use. Chapter II considers how to assess whether technology is adequately implemented.

It is important to recognize one shortcoming of the strategy of recruiting schools and districts that are interested in using a technology application, as compared with sampling a representative set of schools and districts. Sampling a range of schools and districts has certain desirable properties because a study's findings based on a sample of schools would be straightforward to generalize or, in the language used by research methodologists, it would exhibit greater "external validity." For the national study, however, the likelihood of achieving external validity is questionable. The need to use experimental design techniques means that the national study would be negotiating with schools and districts to implement random assignment techniques to measure the effects of technology applications. External validity would be compromised if any schools and districts declined to participate in the study, which is a near certainty.

Designing an experimental study also requires some consideration of the magnitude of effectiveness that the study would be able to detect. Studies that include only a few students or classrooms are suitable only if there are prior reasons to believe that technology applications will lead to very large effects. Larger studies with more students or classrooms are necessary if technology applications are thought to have smaller effects. Considering recent evidence of the effectiveness of various technology applications, the advisory panel and the design team believe that the study should be designed to detect effects that are considered "moderate" to "large" by research methodologists, meaning that the effects are on the order of

Recommendation 5: Design the study to detect "moderate" to "large" effects of technology applications.

one-quarter to one-third as large or larger relative to the standard deviation of the outcome under study. Chapter II considers in greater detail issues of the target effect size.

The design team and advisory panel recognized that the national study could serve as a template for future studies of educational technology. In particular, the national study's reliance on experimental design techniques, it is hoped, will lead other studies to use experimental techniques, which would contribute to a more substantial base of evidence about effective educational practices.

C. What Types of Students Should Be Studied?

Given that a broad range of students use technology applications, it is useful to narrow the range of student users and thus focus the study. Today, students may be exposed to technology applications in preschools, may use assistive technologies to overcome disabilities, and may use distance education technologies to gain access to educational opportunities not provided by their local schools.

The legislated mandate for the study suggests a focus on elementary and secondary students and low-income schools. The study's mandate is in NCLB, and the major use of federal funds for educational technology is to support K–12 education in low-income schools. The design team recommends that the study focus on the effectiveness of technology applications for students in low-income primary and secondary schools. A school's receipt of Title I funds could be used as an indicator that the institution is a low-income school. The latter recommendation could be modified to include schools with higher income levels if a determination were made that what would be learned in these schools also would apply to Title I schools. The objective is that the findings of the study should benefit students who are at risk of falling behind academically. The design team recommends that the study not include assistive technologies designed to support students with particular education needs, nor include technology applications designed to support instruction in English as a Second Language. Both types of technologies would require consideration of many other issues and may merit their own studies.

Recommendation 6:
Study the effects of technology applications on students in primary and secondary schools (K–12).

Recommendation 7:
Study the effects of technology applications in schools that receive Title I funds.

Recommendation 8:
Study the effects of technology applications on student academic achievement as measured by commonly used standardized tests, and collect data on other academic indicators to provide a fuller picture.

D. What Academic Achievement Outcomes Should Be Studied?

Student academic outcomes could include a range of indicators. One might use course grades, standardized test scores, proficiency test scores, or outcomes associated with academic achievement, such as attendance or placement in remedial education. Academic achievement also could be defined by subject area, such as reading, mathematics, science, social studies, or foreign languages, all of which may use technology applications.

For federal policy, the use of standardized test scores is appealing because of the scores' consistency and ease of interpretation. The disadvantages of standardized scores include the tests' infrequent administration (no more than annually), the possible insensitivity of the tests to skills learned through the use of technology applications, and the need to administer the same tests across districts that normally use different tests. Notwithstanding these disadvantages, the design team recommends that the study consider effectiveness in terms of increases in standardized test scores. The team also recommends that data on other academic indicators be collected to provide a fuller picture of achievement. These other data often are found in school records and could be obtained at a reasonable cost.

Recommendation 9:
Study the effects of technology applications that support instruction in reading and mathematics.

The study could focus on the full range of subject areas. However, recent federal efforts in NCLB have focused on ways to improve mathematics and reading instruction (which will be English in secondary schools and may include instruction in language arts in elementary schools). The design team recommends that the national study focus on mathematics and reading and, if resources permit, consider including other subject areas, especially if the other areas include reading and mathematics. For example, physics instruction may include components of mathematics instruction, and social science instruction may include components of English instruction. However, studying more than the two subjects—reading and math—may put pressure on the study's sample size requirements. The design team believes that the study's higher priority should be to examine the effectiveness of technology in the areas of reading and math, rather than its effectiveness for a larger number of subjects.

E. What Are the Conditions and Practices that Influence Effectiveness?

The legislation's mandate goes beyond asking the key framing question—Is educational technology effective in improving student academic achievement?—and inquires about the conditions and practices related to effectiveness. The design team and advisory panel discussed a number of conditions and practices that could be related to effectiveness.

Two questions emerged as particularly important for the national study. The first is the relationship between teacher training and the effectiveness of technology applications. Developers of technology applications and researchers studying their effects often indicate that teachers need to receive adequate training to use the applications appropriately. However, some schools or districts may provide training above the specified or minimum amount. The national study could intentionally vary the extent of training and thus examine whether technology applications are more effective if more training is provided. If training is not varied intentionally, the national study will need to consider the extent to which training varied across schools and teachers and consider how its variation may be related to effectiveness—for instance, by using the statistical modeling techniques described in Chapter II.

In addition to training, the effectiveness of technology applications is likely to depend on other factors that shape the learning context. These factors include characteristics of students, their parents, teachers, classrooms, schools, districts, and neighborhoods. Unlike teacher training, these factors are not amenable to intentional variation. Chapter II considers approaches for assessing the relationship between effectiveness and these conditions and practices, and their interaction with the use of technology applications.

The rest of this report describes approaches for designing and conducting the national study. Chapter II considers how to conceptualize the logical pathways through which a technology application may have effects. It also discusses issues involving statistical power and the number of schools and districts that may be needed for the national study to ensure that the effects of technology applications are detected by conventional methods for estimating effects. Chapter III considers how technology applications could be selected for the national study and looks at approaches for identifying schools and districts for the study. These operational considerations are an important complement to the conceptual issues identified in Chapter II.

CHAPTER II

CONCEPTUAL AND STATISTICAL ISSUES

The mandate of the national study provides general questions and emphasizes one measurement approach over others. Nonetheless, many issues still need to be considered to refine the study further and move it closer to a concrete design.

This chapter sets out a conceptual framework that can guide thinking about the links between technology and learning. The chapter then explores issues related to both structuring experimental designs to measure the effects of technology on learning and the statistical power of various approaches. It concludes with a discussion of approaches for studying the relationships between various conditions and practices and the effectiveness of technology applications.[3]

The discussion presented below suggests that the available resources for studying technology applications that support reading instruction in the elementary grades and math instruction in the middle or high school grades are adequate.[4] At this time, the design team considers classroom-level random assignment within schools to be a desirable approach for studying many technology applications. There would be a need to randomly assign whole schools to study technology applications that can be implemented only for entire schools or for entire grade levels within schools; however, the large number of schools necessary to achieve adequate statistical power would put pressure on resources. Randomly assigning individual students would yield more statistical power and put less pressure on resources, but may be difficult for schools to accommodate. The ultimate choice of approach requires more information about the technology applications being considered and the districts and schools wanting to implement the applications.

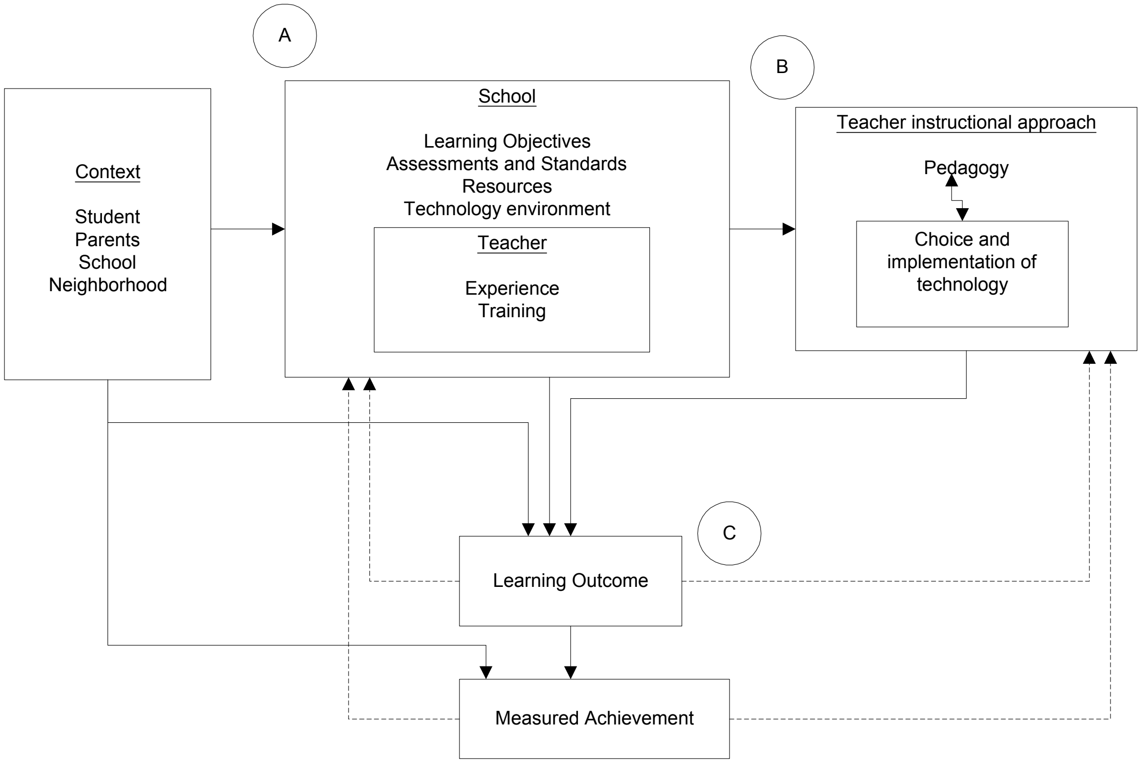
[3] The feasibility of various approaches noted in the chapter have not been assessed by talking directly with schools and teachers; indeed, doing so would not be possible without knowing which particular technology applications were being studied.

[4] The legislation set aside up to $15 million for the study.

A. Conceptual Framework Linking Technology and Achievement

A useful starting point in considering possible approaches for studying the effectiveness of educational technology is to conceptualize the links that connect technology and achievement. Figure II.1 shows a conceptual framework for a technology application that a teacher might use to support instruction in reading or math. In the framework, a student is assumed to have individual, family, and neighborhood characteristics, and the school is assumed to have particular learning objectives, an assessment context (for example, students may take a state or local assessment test), and nontechnology resources.

Figure II.1
Conceptual Framework Linking Technology Application and Learning



Within the school, the student has a teacher with particular experiences and training, including those related to using technology. The school and teacher context and the other factors contribute to a teacher's choice of a pedagogic approach that uses technology or not. In turn, the pedagogic approach contributes to learning outcomes, which contribute to measured achievement. Feedback loops in the framework indicate that the school and the teacher may modify their approaches over time based on outcomes and test scores. Links also are shown between achievement and student, family, and school characteristics, which exert their own influence on learning.

Three aspects of the framework, highlighted with circles, are particularly important for understanding how technology affects student achievement. Probably the most complex set of factors in the framework is the school and teacher nexus (circle A). The many forces that influence a school's teaching and instructional character need to be understood, so that the effectiveness of technology can be separated from the effects of other factors. Another set of factors operates within the classroom (circle B). Teachers' actual use of technology and the factors contributing to their decisions affect the context and the technology's effectiveness.

Finally, learning outcomes can differ from what is being measured by standardized tests or state assessments, and this may lead to different views regarding the effectiveness of a technology application (circle C). For example, teachers may see high scores on in-class tests and credit the intervention, while a standardized test measuring different competencies may not reveal gains.

Different technology applications may call for modified or more detailed conceptual frameworks. For example, the pathways by which a computer-based application supports students learning to decode words may differ from the pathways by which an Internet-based application connecting students in different countries supports their learning to write. Technology applications that diagnose or assess student skills—and that thereby enable teachers to aim instruction and activities at particular skills—may have different pathways than applications focusing on student instruction. The national study can develop more specific conceptual frameworks when the specific technology applications are known.

B. Random Assignment and Sample Size Considerations

Experimental methods were mandated in NCLB, and have been supported by IES and the advisory panel to estimate the effects of technology applications. Generally, an experimental design creates a treatment group (the group of students who are able to use a technology application) and a control group (the group of students who are not able to use that technology application) by using a randomizing device equivalent to a coin toss. The use of the randomizing device ensures that the two groups differ *only* in that one group can use the technology application.[5] Any differences in outcomes between the two groups can then be attributed to the technology application. The ability to make causal statements ("the outcome difference is caused by the technology application") is a powerful argument for using random assignment. Other approaches for creating comparison groups of students would necessarily mingle differences between groups and the effects of the technology applications, and make it difficult to attribute any measured outcome differences to the technology application.

[5] The two groups can differ due to chance variation, especially when the groups are small. However, random assignment assures that differences between groups diminish as sample sizes grow. Another desirable property of random assignment is that well-known statistical formulas can be used to test whether observed differences between the two groups could have arisen due to chance.

Random assignment is a flexible tool. In principle, the treatment and control groups could consist of schools, classrooms, or students. The choice of unit always involves tradeoffs. For example, randomly assigning 600 students who attend two schools into a treatment group and a control group is quite different from randomly assigning two schools that serve 300 students each into a treatment group and a control group (containing one school each). Using schools as a unit might be more convenient, but the schools might have different types of students and the differences could affect student outcomes. Therefore, the estimated effect of the technology would depend on which school was assigned to the treatment group. Changes in the estimated effect related to which students are assigned to the treatment or control groups are likely to be far smaller in size.

The following sections describe the strengths and weaknesses of various options for random assignment. The appropriate choice would depend on the implementation of the technology application and schools' ability to support random assignment of students to classrooms.

Approach A: Randomly assign students to use the technology application

Identified students (for example, students attending a school that agreed to participate in the national study), possibly at a single grade level, could be assigned randomly to either use or not use the technology application. This approach is statistically powerful, but may be infeasible in most schools. Treatment and control group students would be together in classrooms, and mixing creates at least a theoretical possibility that the treatment group could affect the control group. Treatment group students would be able to use the technology application only if the class split apart during a period. For example, treatment group students might go to a computer lab, while control group students remained in the classroom. Many schools may not be able to accommodate the scheduling demands created by this approach to random assignment.

Approach B: Randomly assign classrooms to use the technology application

The grouping of students within classrooms points toward a natural approach for random assignment: whole classrooms could be assigned to the treatment or control group. For example, a school with four first-grade classrooms (four teachers) could have two classrooms assigned to use the technology applications and two assigned not to use it. If the number of classrooms was not even, unbalanced assignment (such as two classrooms to the treatment group and one to the control group) could be accommodated. Assigning classrooms randomly would be best suited to situations in which teachers volunteer to use the technology. Otherwise, simply assigning all classrooms at a grade level could result in teachers who do not want to use the technology being assigned to use it, which could lead to weak implementations.

Approach B has two interesting variants. The first randomly assigns students to classrooms, and randomly assigns teachers to use the technology application or not. This approach is essentially equivalent to approach A. In fact, it would be better, because treatment and control group students would not be mixed together. The purpose of randomly assigning teachers to use the technology application would be to break the relationship between teacher choices and the effects of applications. For example, younger teachers, who are probably more comfortable with using technology, would be more likely to choose to use it, and the effect of the technology would then be mingled both with the greater facility and the lower experience level of the younger teachers.

The second variant discussed with the advisory panel would assign some classrooms to a technology application—for example, to improve math skills—and assign other classrooms to another technology application with a distinct objective—for example, to improve computer keyboarding skills. This approach may be attractive to schools, because all students would receive some form of treatment and no student would be denied access to a technology application. However, the approach needs careful consideration, because, in principle, effects are measured correctly only if the less-intensive treatment has no relationship with the main outcome of the more-intensive treatment (for example, better keyboarding skills would be assumed to have no effect on math ability).

Approach C: Randomly assign teachers to use a technology application for one of their class sections

The study also could assign teachers to use a technology application for some of their class sections. The approach requires that teachers have at least two sections of the same subject, a requirement that may limit the feasibility of the approach. For example, a teacher who taught geometry to two sections could be assigned randomly to use a technology application in one of the two sections. This "within-teacher" design would be further enhanced if students were assigned randomly to the class sections. (The complexity of class scheduling would be an issue.) Also, the within-teacher design would need to consider possible spillover between a teacher's technology application and non-technology-application periods. For example, if teachers adopted more effective instructional approaches in their non-technology-application class section because of what they learned using the technology application, the design would underestimate the effect of the technology application.

Approach D: Randomly assign schools to use the technology application

From the group of schools interested in using a technology application, some would be assigned to use the application, and others would not. The approach is more powerful if all or most teachers within the school use the technology application (for example, all first-grade teachers use the application to support reading instruction). Otherwise, as noted above, the effects of the technology application can be confounded with the characteristics of the teachers who choose to use it. This approach is more powerful when all the schools have similar characteristics.

C. Statistical Power Analysis

Statistical formulas can be used to compare the ability of studies using different approaches to random assignment to detect an effect of a particular size. Considering issues of statistical power helps in understanding the desirability of various study structures and the concomitant demands on resources. Two parameters are needed to conduct a power analysis: the size of the effect that the study wants to be able to detect, and the degree to which students are clustered in classrooms and schools.

An effect size of 0.35 emerged from advisory panel discussions as a reasonable target, for two reasons.[6] First, a review of small-scale studies had suggested that individual technology applications can have effects of that amount or more (Murphy et al. 2002). Second, a smaller effect size would do little to close achievement gaps between various segments of the student population, a key objective of the NCLB legislation. However, it may be reasonable to consider a more conservative strategy. The research literature is oriented more toward small-scale studies, and literature reviews generally present findings only from published or released studies. This may lead to an overstatement of effectiveness, under the assumption that studies are more likely to be published or released if their findings are larger and favor effectiveness. Also, large-scale studies such as the Tennessee STAR experiment, which measured the effects of reducing class size, have been considered successful with a 0.20 effect size. For these reasons, the calculations presented below use a lower bound of 0.25 and an upper bound of 0.35 for a target effect size.

The second parameter is the degree to which students are "clustered" in classrooms and schools. Generally, students in the same school or classroom are clustered in the sense that their outcomes are related. As will be shown below, the sample size needed to detect a target effect size increases as clustering increases.[7] As described in Appendix B, the design team used data from a longitudinal study to estimate classroom-level clustering and school-level clustering for reading and math, both in levels and in growth over time. These estimates are based on an analysis of reading and math test scores received by third, fourth, and fifth graders attending schools similar to the kinds of schools recommended for the national study.

Figures II.2, II.3, and II.4 present the relationship between minimum detectable effect size and the number of schools, classrooms, and students. They show that having more clusters in the sample generally improves power.[8]

[6] By definition, effect sizes are a percentage of a standard deviation. The use in the text of an effect size of 0.35 means the effect is 35 percent as large as the standard deviation of the outcome being considered.

[7] Raudenbush (1997) showed that the sample size needed to detect a particular effect size and the study costs implied by that sample size are smaller if within- and between-cluster variance can be reduced using a baseline covariate. In a related study, Bloom et al. (1999) showed that these components of variance are reduced substantially, if test scores are the outcomes being analyzed and a baseline test score is available. The calculations in the text assume that a baseline test score will be available and that the baseline score reduces between- and within-cluster variance by 20 percent. The reduction is less than what Bloom et al. (1999) found, so the sample sizes noted in the text are conservative estimates.

[8] The calculations use a fixed classroom size of 20 students. The calculations could be refined further by allowing the number of sample students per classroom to vary. This would acknowledge that sampling fewer students per classroom can free up resources to sample more classrooms. Trading off fewer students for more

- ***Student random assignment***

  Achieving the target effect size of 0.35 would require 10 classrooms with 20 students in each—a total sample size of 200 (see Figure II.2). Achieving the effect size of 0.25 would require 20 classrooms with 20 students in each—a total sample size of 400.

- ***Classroom random assignment***

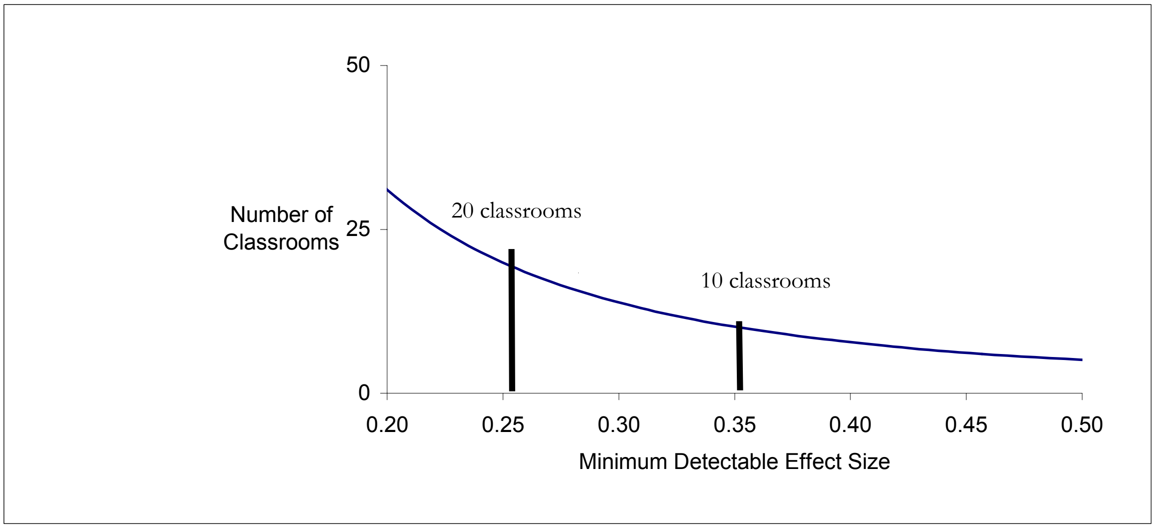
  Achieving the target effect size of 0.35 for an application focused on reading would require 30 classrooms with 20 students in each—a total sample size of 600 students (see Figure II.3). Achieving the target effect size for an application focused on math would require 40 classrooms with 20 students in each—a total sample size of 800 students. Sample sizes are larger for math, because classroom clustering is greater for math, as shown in Appendix B. For a target effect size of 0.25, the analogous sample sizes are 58 classrooms for a reading application and 76 classrooms for a math application.

- ***School random assignment***

  Achieving the target effect size of 0.35 would require 29 schools with 40 students in each—a total sample size of 1,160 (see Figure II.4). Achieving the target effect size of 0.25 would require 57 schools with 40 students in each, a total sample size of 2,280.

Figure II.2

Classrooms Needed to Achieve Minimum Detectable Effect Sizes
(Student Random Assignment to Classrooms)



(continued)
classrooms can reduce sampling variance under specific assumptions about student-level and classroom level

Figure II.3

Classrooms Needed to Achieve Minimum Detectable Effect Sizes
(Classroom Random Assignment)

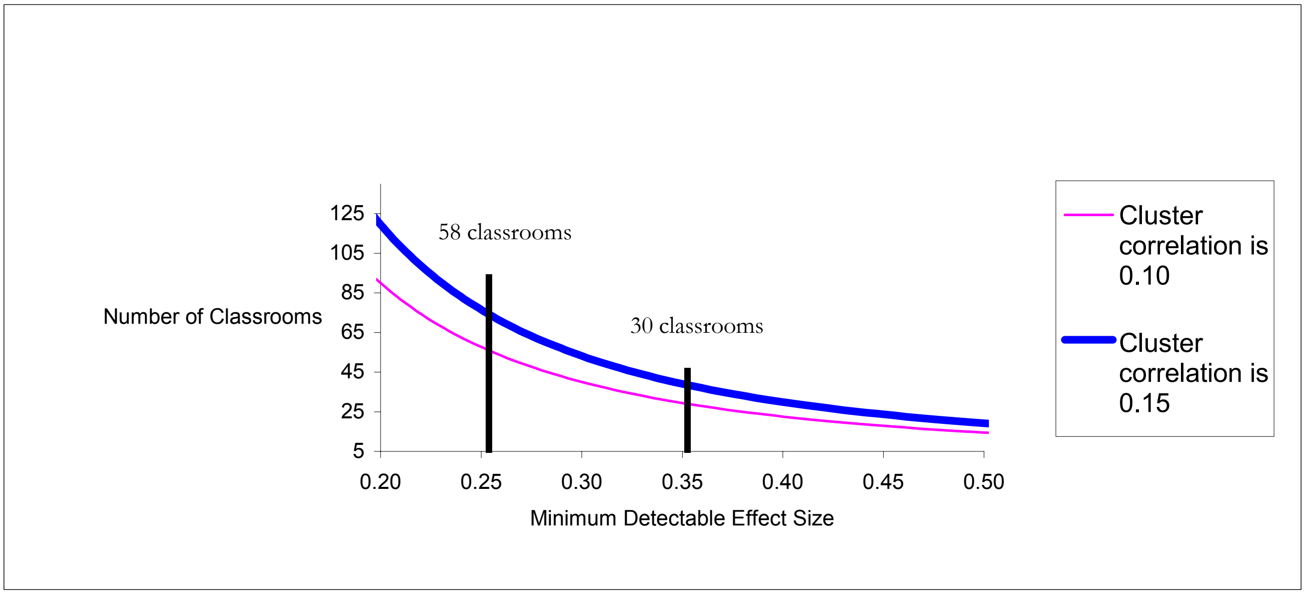
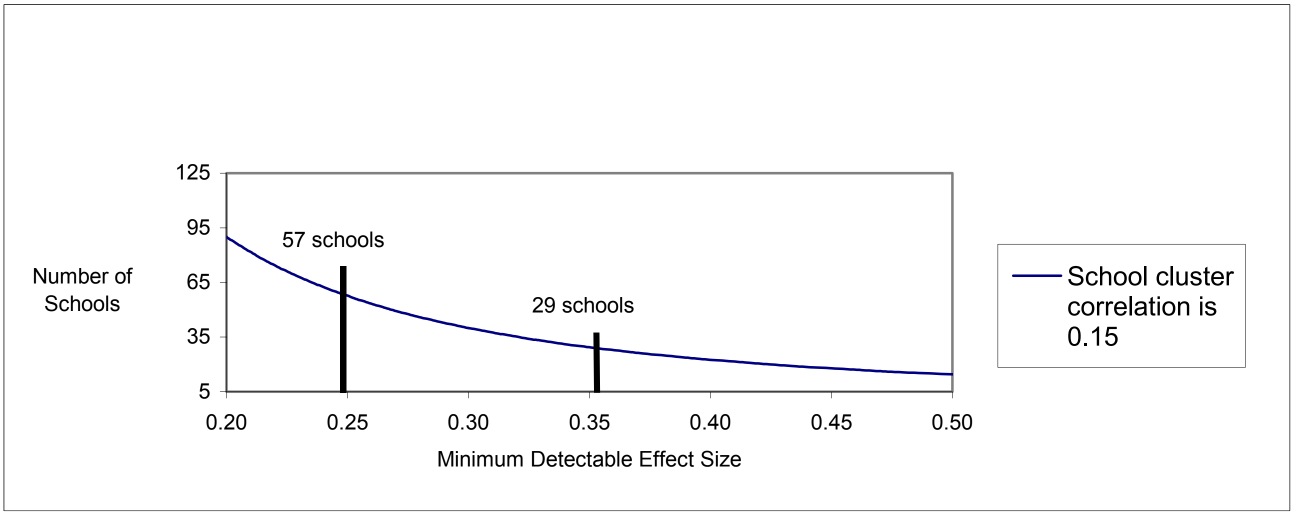


Figure II.4

Schools Needed to Achieve Minimum Detectable Effect Sizes
(School Random Assignment)



(continued)
variance and the costs of collecting data from students and classrooms. The calculations also assume that

(continued) classrooms are evenly divided into treatment and control groups.

The detectable effect estimates provide guidelines for assessing the approximate number of schools that the study would need. If an elementary school, for example, has four classrooms at a grade level, all four participate in the study of a reading application, and classrooms are randomly assigned, 10 to 15 such schools would be needed for the study (depending on whether the target effect size is 0.35 or 0.25). Even medium-size urban school districts have 10 to 15 elementary schools, so a single school district could be the host of a study of one technology application and provide a suitable level of statistical power. For reasons noted below, however, spreading the schools across different districts would enable the study to provide more information about the factors influencing effects. School random assignment would require more schools. Achieving the target effect size requires 30 to 60 schools, roughly.

Assessing the overall size of the study would require applying estimates of student, classroom, and school data collection costs to the sample sizes. However, if the study were to focus on two areas of applications, such as applications for elementary reading and middle school math, and if classroom random assignment were used, 20 to 30 schools would be needed for the study. This could require as many as 30 school districts, if each district had only one school participating in the study; but it would be reasonable to assume that a smaller number of districts, between 5 and 15, would need to participate to reach the target number of schools. If student random assignment were possible, fewer schools would be required, and if any school random assignment were required, more schools would be needed.

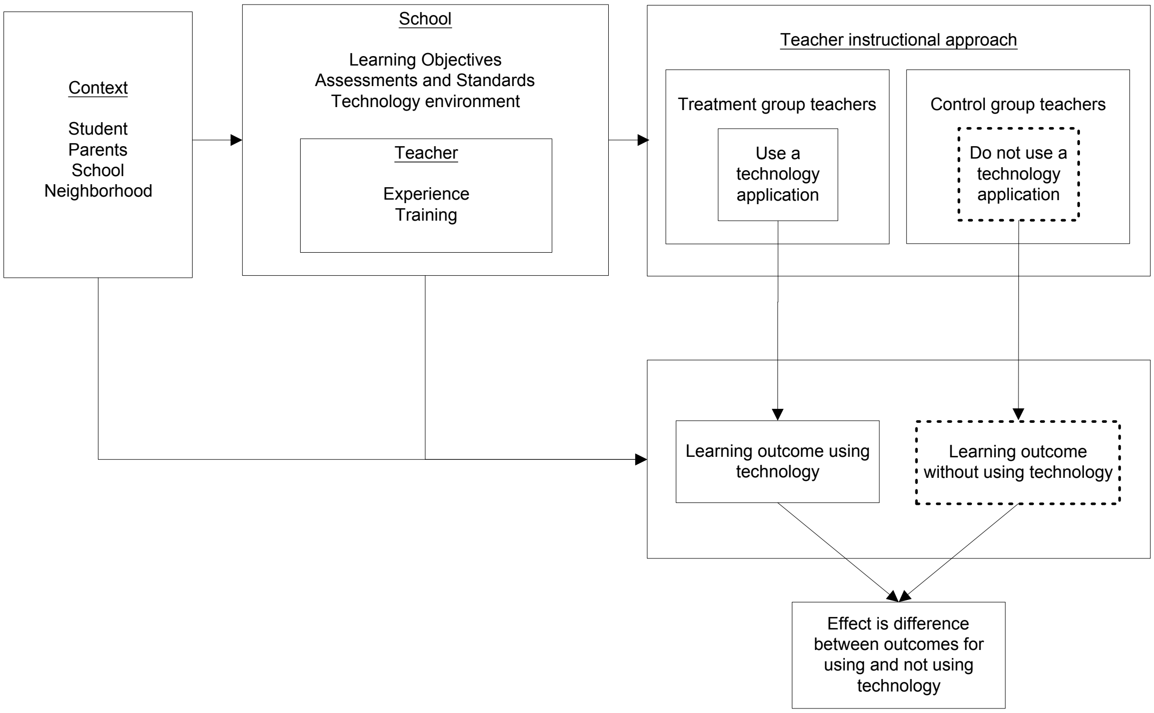
D. Estimating Impacts

Experimental designs yield a simple and elegant estimator of the effects of an intervention: the difference between the average outcomes of treatment and control groups at followup. The simple estimator can be enhanced by placing it within a regression model that can yield increases in statistical power by using available information about the sample members.

The effects of technology applications may be understood better by estimating effects for various groups of students of special interest. The groups could include common, demographically defined groups such as those defined by age, sex, and race, as well as students whose characteristics may provide information about how better to target technology use—for example, students with low and high achievement at baseline, students with or without access to computers in their homes, and students from lower- and higher-income households. Generally, technology effects can be estimated for subgroups of students by segmenting the sample according to the variable that defines the subgroups and estimating an effect for students within the subgroups. Statistical tests then can be conducted to assess whether differences in the estimated effects for the subgroups may be due to chance variation or whether they reflect actual differences in effects.

An important goal of the study is to understand the relationship between technology's effectiveness and the conditions and practices that shape the learning context. Figure II.5 modifies the conceptual framework in Figure II.1 to illustrate how conditions and practices can be viewed as contributing to the effects of technology. In the modified framework,

Figure II.5
Conceptual Framework Illustrating the Effect of a Technology Application


Context
Student
Parents
School
Neighborhood
School
Learning Objectives
Assessments and Standards
Technology environment
Teacher
Experience
Training
Teacher instructional approach
Treatment group teachers
Use a technology application
Control group teachers
Do not use a technology application
Learning outcome using technology
Learning outcome without using technology
Effect is difference between outcomes for using and not using technology

students and teachers are randomly assigned to either an instructional approach that uses a technology application (shown with the solid line in the "teacher instructional approach" box) or one that does not (shown with a dashed line). The two instructional approaches would generate two learning outcome levels, such as an average test score when a technology application was used and not used. If the experimental design is set up appropriately, the only difference between the two groups will be technology use, enabling the difference in outcomes to be interpreted as being caused by the difference in technology use.

The conceptual framework illustrates how conditions and practices may influence the technology effect. For example, local and school contextual factors may be related to outcome levels and, thereby, to the measured effect. Similarly, differences in teacher background, experience, and training may be related to outcome levels and, thereby, to the measured effect. To gain a better understanding of the relationships between effects and contextual factors, the study would need to be implemented in a variety of different contexts, so as to create the potential of observing how effects vary as the contexts ("conditions and practices") vary.

The national study provides an opportunity to consider experimental evidence about the effectiveness of a particular condition or practice of interest. For example, the national study could consider assigning teachers randomly to receive a standard amount of training (for example, the amount the developer of the technology application considers appropriate for its effective use) or an enhanced amount of training (which some developers may provide as a customized level for districts and schools that pay additional costs). By comparing the effects of technology for teachers receiving the low and high levels of training, the study would be providing evidence of the value of additional training.[9]

Whether the study could provide experimental evidence of the effects of conditions and practices is partly an issue of cost and partly one of feasibility. The cost issue arises because the study would need to include enough teachers at the two levels of training to have a reasonable probability of distinguishing the estimated effects for the high and low levels of training. However, using resources to study the effects of different conditions and practices would come at the expense of using the resources to study more technology applications themselves. The feasibility issue arises because using experimental designs to study two different levels of a condition or practice, such as training, would require that a large number of schools participate, or schools with a large number of classrooms at the same grade level, so that the random assignment can be carried out to the high or low level of the condition.

Another approach to understanding the importance of conditions and practices is to block the sample into observed levels of the condition or practice. For example, schools could be blocked into high-resource and low-resource schools, or high technology use and low technology use schools. Measured effects then can be compared across the various blocks to estimate the effects of the blocking variable. The blocking approach does not yield an experimental estimator (schools are not assigned randomly to the blocks), but it does generate useful information that can lead to further investigation and analysis.

A formal approach that incorporates information about blocking variables is to model the effects of technology within a hierarchical framework. This approach is referred to as "hierarchical linear modeling" (HLM). The approach examines the relationship between effects and characteristics, such as the relationship between classroom effects and the extent of professional development received by teachers of those classrooms. An advantage of the HLM approach is that it controls for other characteristics when estimating a particular relationship.

The HLM considered here can be illustrated as having two stages. In the first stage, effects will be estimated at a particular level of aggregation. For example, if classrooms within schools are randomly assigned to use the technology application, in the first stage an effect can be estimated for each school. In the second stage, the effect estimated in the first stage can be estimated as a function of school characteristics, such as its level of academic press or the level of poverty in the local area.[10]

[9] Even if the study could not experimentally vary training levels, it would still be desirable to ensure that all teachers in classrooms using the technology applications are appropriately trained in its use, to ensure that the study focuses on well-delivered applications.

[10] A useful feature of the HLM approach in this context is that its first-stage equation would be based on an experimental design, which strengthens the second-stage equation, since it would be based on valid estimates of effects. If the first stage were not based on experimental designs, the effects being modeled in the second stage would include elements of bias created by nonexperimental designs, and the bias may then be imparted to the second-stage estimates.

More formally, suppose that each school has one treatment classroom and one control classroom. A two-level model of academic achievement for student *i* in classroom *j* and school *k*: would be:

Level 1

$$y_{ijk} = \alpha_k + \beta_k \mathrm{T}_{ijk} + e_{ijk}$$

Level 2

$$\beta_k = \gamma + \delta(\mathrm{R}_{Tk} - \mathrm{R}_{Ck}) + \mu_k$$

where,

- $i = 1,2,\ldots,\mathrm{I}$
- $j = \mathrm{T}$(treatment),C(control)
- $k = 1,2,\ldots\mathrm{K}$
- y_{ijk} = post-random-assignment test score of student *i* in classroom *j* and school *k*
- α_k = average test score of the control classroom in school *k*
- T_{ijk} = 1 if student *i* in school *k* is in the treatment classroom, and zero otherwise
- β_k = difference in the average test score between the treatment and control classroom in school *k*—that is, the impact for school *k*
- γ= average of the *k* school-level impacts
- $(\mathrm{R}_{Tk} - \mathrm{R}_{Ck})$ = difference in the amount of professional development received by the treatment and control teacher in school *k*
- δ= effect of professional development
- e_{ijk}, μ_k = random error assumed to be uncorrelated

The element of particular interest in the Level-1 equation is β_k—the (k x 1) vector of school-level impacts.[11] The element of particular interest in the Level-2 equation is the single parameter δ—the relationship between school-level impacts and teacher professional development.

To increase the precision of the school-level impacts, other student characteristics would be included in the Level-1 equation, such as the baseline test scores of students. Similarly, school-level characteristics other than teacher professional development would be included in the Level-2 equation. However, in the latter case, other school-level characteristics would be included, both to increase the precision of the relationship between school-level impacts and teacher professional development, and to provide an unbiased estimate of that relationship.

Using the HLM approach will require that the study have several features. First, schools recruited for the study need to reflect combinations of the conditions and practices of interest because estimating the second-stage equation requires variability in conditions and practices (if two school characteristics always varied together, the second-stage equation could include only one of them). Second, including several schools for each combination of conditions and practices of interest would make it possible to assess the robustness of condition-and-practice results. Third, data on a variety of conditions and practices would be needed for the model to be estimated correctly, with some practices and conditions being amenable to change through policy, and some not. Leaving one variable out of the equation would lead to incorrect estimates of other variables. For example, whether a school is located in an urban area (which is not amenable to policy) may be related to effects as much or more so than the level of teacher professional development, which is amenable to policy. However, if only teacher professional development were included in the equation, the coefficient estimate for professional development would include some of the effect of being in an urban area, which would misstate the relationship between technology effects and teacher professional development.

The HLM approach makes it possible to study conditions and practices, but it is clear that the number of conditions and practices the study can focus on is limited. By conventional standards, the sample size for the second-stage equation would be small. In the example noted here, the sample size for the second-stage equation is the number of schools included in the study for that technology application, rather than the number of classrooms or students. Even combining schools for two technology applications would mean having only 30 or so schools, and a small number of variables—more than 5 or so—would reduce the degrees of freedom and statistical significance.[12]

[11] The model specification assumes that the Level-1 equation is estimated separately for each of the *k* schools.

[12] Note that with school random assignment, the second-stage of the HLM approach would be based on the number of districts in the sample, which may be too small to support estimation. This is another reason to prefer classroom or student random assignment.

E. Collecting Data to Estimate Effects

The study approaches described here would need to obtain data on student achievement, other characteristics of students, teachers, and schools, as well as data assessing the presence of various conditions and practices. Some of the data items can be collected through quantitative methods, survey instruments, and school records. Other data items may need to be collected through qualitative methods, such as site visits.

The issue of how to assess achievement outcomes arose frequently during advisory panel discussions. The possible shortcomings of using standardized tests to measure student achievement were noted. Tests are administered infrequently. They may not be sensitive to some skills or competencies that could be enhanced by technology. A mismatch might exist between the rate at which technology may increase scores and the timeframe for the study's data collection, which would likely involve only two years of score data being collected. However, the panel and the design team considered the policy relevance of test scores to outweigh concerns about their use as the primary outcome for the study and recommended that other data about achievement be collected, in addition to test scores, to provide a fuller picture of achievement. Other data could include student grades, retention in grade, attendance, and placement into special education or remedial reading programs. These data are commonly available in school records and would be relatively inexpensive to obtain.

The specific standardized test the study would use to measure student achievement also is an issue. The study would likely be conducted in more than one school district, and tests administered by districts themselves are likely to differ, leading to inconsistent measures of the effects of technology applications. Using one test, however, would mean that some students would need to be tested both by their school, for district purposes, and by the national study, for its purposes. The design team recommends that the properties of various commercially available, standardized tests be explored further and, in particular, that consideration be given to standardized tests that have shorter versions, which would enable the study to measure student skills in one area, such as reading, without having to administer a half-day-long test, as is common for full versions of tests.

The advisory panel also discussed the issue of assessing fidelity of implementation. Because the study would focus on implementing particular technology applications with districts, schools, and teachers that have not yet used them, some implementations may not succeed within the study's timeframe. The design team recommends that the national study develop metrics for assessing the fidelity of implementation, possibly based on developer guidelines and other sources, so that measures of implementation success can be applied in the effort to understand measured effects. The national study would provide useful information to the field if it found technology applications that were able to produce effects even when their implementations were difficult or appeared to fall short of developer guidelines, and technology applications that did not produce effects even when their implementations were close to ideal.

In addition, the opportunity to assess fidelity of implementation could be used to collect data on the various conditions and practices of interest. Trained field researchers could visit schools and classrooms participating in the study to observe implementation and to gather information about conditions and practices through qualitative techniques such as interviews with teachers and administrators. The researchers could code the information into variables that support the HLM approach to measuring the relationship between the conditions and practices (indeed, fidelity of implementation can be viewed as a practice). Field researchers also could visit schools or classrooms in the control groups, thus enabling the study to better understand counterfactual conditions. A fuller protocol for field research needs to be considered, together with quantitative instruments, so that the types of information that are collected mutually support each other.

Chapter III

Approaches for Selecting Technology Applications and Schools

Two major challenges for the national study will be to select the technology applications that will be examined and select the schools that should be included in the study. These tasks are challenging because both the technology applications and the schools need to satisfy a number of criteria that support the recommendation to study well-implemented applications in a real-world setting. In particular, the technology applications either need to have been shown to be effective or to use promising approaches. The schools included in the study need to have teachers that are interested in using technology, since this should increase a school's likelihood of actually using an application.

Technology applications and schools also need to satisfy other criteria that support policy goals. For example, the applications need to target reading or math skills, and the schools need to receive or be eligible to receive Title I funds. The focus on reading and math applications is consistent with one of the main educational objectives to improve these skills, and the focus on Title I schools is consistent with the goal to improve academic achievement of these students. This chapter describes two processes—one for selecting technology applications, and another for selecting schools. Several issues revolving around both processes will need to be resolved after the actual selection of technology applications and schools begins. The unresolved issues are mentioned in the chapter, along with possible solutions. Therefore, at this point, the processes should be viewed as a starting point in finding ways to select the technology applications and schools to include in the study.

A. Selecting Technology Applications

Consistent with the legislation mandating the national study, the recommended focus would be on examining the effects of technology applications that have been designed to improve student academic achievement. While this focus reduces the number of technology applications the national study needs to consider, many more still exist than can be included in the study, which means the study will need to select a subset of applications to examine. This section describes a feasible process for selecting technology applications.

1. Specific Applications or Types?

An issue the advisory panel discussed at length is whether the national study should focus on examining the effect of several *specific* applications, or several *types* of applications, where each type contains several specific applications. The design team recommends studying types of applications because that will more likely provide useful and durable information. Consider, for example, the difference between studying one application that is designed to develop a student's skill in manipulating fractions (that is, a specific application), as opposed to several applications, each of which is designed to develop these skills (that is, several applications of the same type). The evidence produced by the former study would indicate whether the specific application is effective; it would not indicate whether applications of this type are effective. The opposite would be true of the evidence produced by the latter study. The latter type of study seems more useful because studying types would provide useful information about attributes of technology applications that may contribute to their effectiveness, which in turn contributes to more effective designs of future technology applications.

The first step in determining the types of applications that could be studied is to classify applications along the dimensions that distinguish them. These dimensions include: (1) skills that applications target, (2) instructional approach, (3) intensity of the applications, and (4) recommended amount and type of professional development.

- The ***skills that applications target*** can vary dramatically from application to application. For example, the five components of reading instruction that the National Reading Panel (2000) studied were phonemic awareness, phonics, fluency, vocabulary, and comprehension. Applications designed to target the development of early reading skills may focus on one or more of these essentials. Some applications may place greater emphasis on teaching phonics, whereas others may balance the learning of phonemic awareness, phonics, fluency, and vocabulary acquisition skills. Similarly, applications designed to teach elementary mathematics may focus on basic addition, subtraction, multiplication, and division skills. But the relative emphasis on any one of those skills may vary from application to application.

- Applications may differ in terms of their ***instructional approach***. Applications may have different expectations regarding how students work; some applications are intended for individual students, while others actively encourage students to work in pairs or groups. While many applications, particularly those designed to remediate basic skills, may rely on direct instruction or drill-and-practice techniques, several other applications may encourage more self-directed and open-ended learning by students. Some applications encourage students to use multiple methods in solving problems, while also providing ongoing assessment feedback to the student in the form of suggested strategies. This approach differs from applications that use management systems to provide students with feedback in the form of differentiated problem sets.

- The ***intensity of applications*** varies. Some applications are designed to be supplement instruction, whereas others are designed to deliver a significant portion of the instruction students receive. Additionally, the amount of time the application suggests that students spend with it, in order to achieve the expected learning outcomes, varies. Some early literacy applications, for example, recommend that students work with the application for 15 minutes a day. Other early literacy programs may recommend intermittent use (for example, every other day for 30 minutes), or they may leave the choice up to the classroom teacher.

- The ***amount and type of professional development*** may vary. Some applications require little professional development, while others are explicit about the need for sustained professional development over time. The amount of additional instruction teachers receive is likely to play a significant role in shaping learning outcomes. This becomes all the more true in cases where professional development goes beyond the "how to" of using the technology and focuses on content and instructional issues.

In theory, the types of applications that could be studied would include all combinations of the dimensions above; however, some combinations may not exist among current applications. For example, some math applications may focus on developing the skills needed to manipulate fractions, whereas others may focus on developing the skills needed to solve word problems. Further, some math applications may develop skills using drill-and-practice instructional techniques, while others may use a more open-ended or self-directed technique. It may be the case, however, that math applications that focus on developing skills in manipulating fractions do not use certain instructional approaches, such as an open-ended/self-directed technique.

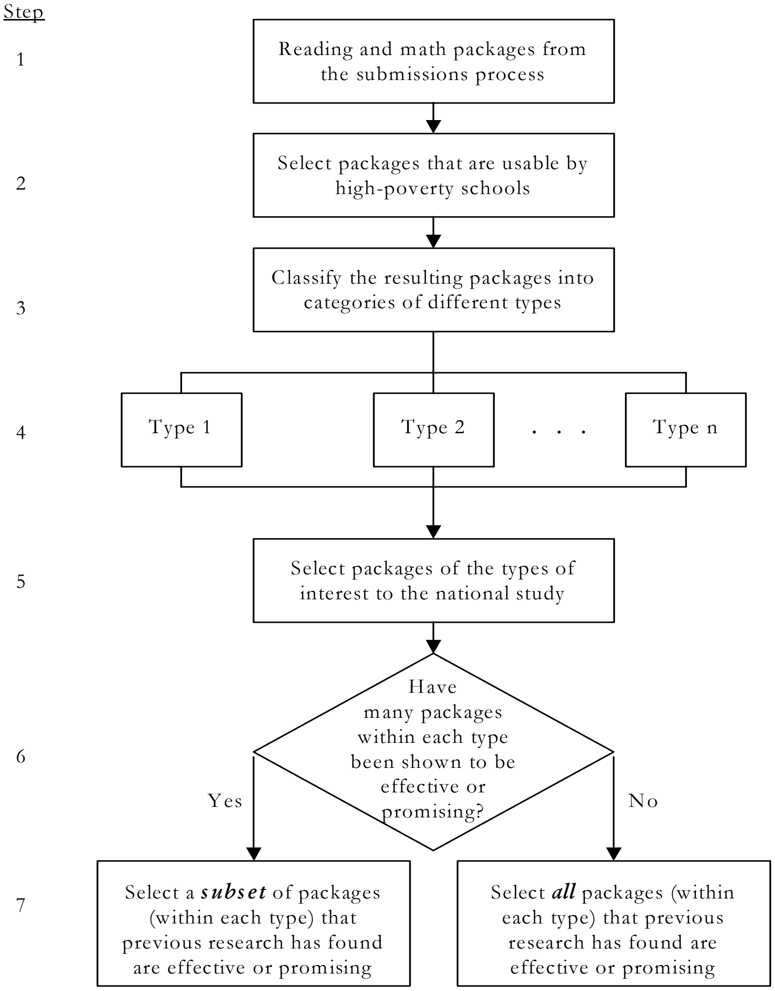
The national study should work with sources in the industry—such as the Software and Information Industry Association (SIIA)—to understand: (1) the important dimensions that distinguish applications, and (2) the most common combinations of these dimensions among the applications that currently exist. More information on the types of applications will be obtained through other commissioned papers currently being prepared by outside experts. Knowing which combinations are and are not common would be useful for the national study when deciding which types of applications to examine.

2. A Process for Selecting Technology Applications

Figure III.1 presents a process for selecting technology applications to include in the study. The process uses information from a Web-based submission process that is currently being developed by the design team.[13] This submission process asks developers to provide

FIGURE III.1

PROCESS FOR SELECTING TECHNOLOGY PACKAGES


Step
1
Reading and math packages from the submissions process
2
Select packages that are usable by high-poverty schools
3
Classify the resulting packages into categories of different types
4
Type 1
Type 2
. . .
Type n
5
Select packages of the types of interest to the national study
6
Have many packages within each type been shown to be effective or promising?
Yes
No
7
Select a *subset* of packages (within each type) that previous research has found are effective or promising
Select *all* packages (within each type) that previous research has found are effective or promising

[13] Developers who will submit their technology applications for consideration will, in a sense, be applying for inclusion in the study. To avoid confusion when referring to technology "applications" and a developer "applying" for inclusion in the study, we will refer to technology applications as technology packages from this point forward.

information about their technology packages, including: (1) demographics about students and schools using their technology, (2) prevalence rate, (3) intensity of the instruction provided, (4) implementation costs, and (5) documents that present evidence of the technology's effectiveness in improving student academic achievement. The process for selecting technology packages presented in the figure also would use input from a panel of experts the national study should convene. This panel would be responsible for providing guidance in selecting both the types of technologies to study and the specific packages within each type. This section provides details of this process.

a. Identifying the Universe of Eligible Technology Packages

A useful starting point for selecting technology packages is a list of all the technology packages that meet the inclusion criteria for the national study. This would ensure that all eligible technology packages are considered.

The advisory panel made two recommendations about the technology packages that should be studied. First, the technology packages should be limited to those that have been designed to enhance reading and math skills, since such a focus is consistent with one of the main educational objectives. Second, the technology packages should be limited to those previous research has found effective, or those with promising approaches but for which evidence of effectiveness has not yet been produced.

Based on developers who submit their products, the submission process will generate a list of technology packages designed to enhance reading and math skills. Therefore, the list will already be limited only to reading and math packages (step 1 in the figure). For each technology package, the submission process will generate key pieces of information that can be used to further reduce the list. For example, the advisory panel recommended that the study should focus on the effect of technology in high-poverty schools. Information collected as part of the Web-based submission process can be used to determine the types of technology that high-poverty schools currently use and thus may be easier for high-poverty schools—which currently do not use technology but which are interested in doing so—to implement (step 2 in the figure).

The next step in the selection process uses information collected as part of the submission process to classify technology packages into different types (steps 3 and 4 in the figure). For example, each submission will indicate the grade levels that predominantly use the technology package, the intensity of the package, and the qualifications teachers should have in order to use the package. This information can be used to classify technologies into those used by different grade levels, those intended to be used different amounts of time, and those requiring different qualifications.

b. Developing a Short List

The design team recommends that a panel of experts in technology packages be convened to provide guidance for deciding which types of packages to study (step 5 in the figure). At least two issues should be considered when making this decision. First, the panel may want to recommend studying the types of technology packages that fit within reasonable bounds of costs. Implementing some technology packages requires a significant amount of computer hardware and particular qualifications. While these packages may be effective in increasing student academic achievement, they may be too expensive or too difficult to implement in high-poverty schools that were recommended as the focus of the study.

Second, the panel may want to recommend studying technology packages that deliver a particular amount of instruction or deliver instruction in a particular way. As mentioned in the previous chapter, the advisory panel recommended that the study be designed to detect a moderate to large effect size. Compared to the effects of educational interventions studied previously, one that has a moderate effect size on academic achievement is a high standard of effectiveness. The panel that provides guidance in selecting the types of technology packages to study may want to select those types that appear to have been designed to have a large effect on academic achievement, or for which prior evidence suggests they have a large effect.

c. Further Investigating the Short List

The final step in the selection process is to choose the technology packages within each type that will be included in the national study. For some types of technology packages, this may be a straightforward task because the number of packages that have been shown to be effective may be small enough to include all of them in the study. For other types of technology packages, this task could be more difficult because there may be numerous packages that have been shown to be effective. In the latter case, the design team will continue to work with ED and the panel to develop criteria for selecting the specific packages within each type.

Whether the number of technology packages within each type is small or large, the national study will need to determine which packages within each type have been shown to be effective. Information collected as part of the submission process can be used for this purpose. Developers will be asked to provide evidence of their technology's effectiveness in improving student academic achievement. The evidence can be used to identify technology packages that previous research has found are effective.

An important issue is whether the evidence about the effectiveness of a particular technology package is reliable. Few technology packages have been evaluated using rigorous methods. A review of recent evidence on the effectiveness of technology packages indicates that, out of the 195 studies that were identified, only 31 were based on experimental or quasi-experimental methods (Murphy et al. 2002). The other studies were based on other methods, such as a pre-post design.

The design team will develop a rubric that can be used to determine whether a particular piece of evidence is reliable. Essentially, the rubric will rate the degree to which a study provides causal evidence of the effect of a technology package. A study that is rated as providing strong causal evidence is one that has addressed all the issues that arise when comparing the outcomes of individuals who used a technology package with the counterfactual—that is, the outcomes individuals who used a technology package would have experienced had they not used the intervention.

Several issues will be addressed when developing the rubric:

- ***Which aspects of a study are important to consider?*** The aspects should include the type of design used for the evaluation, quality of the data, and whether the evidence is internally and externally valid.
- ***How much weight should each aspect of a study receive?*** Are all aspects of a study equally important? Or, are certain aspects of a study—such as the design—more important? This is an important issue because different weights are likely to result in different scores for the same study.
- ***Should the rubric produce a single score for each study, or more than one score?*** Two studies with the same score could be quite different along aspects of the studies (such as their evaluation designs) that are used to computer their scores.

Two rubrics are currently being developed that may be useful when developing a rubric for this study. One—part of the activities of the What Works Clearinghouse—is called the Study Design and Implementation Assessment Device (hereafter, WWC Study DIAD). The WWC Study DIAD is particularly useful because it focuses on research about the causal effects of educational interventions. A final version is scheduled for release in spring 2003 (Valentine and Cooper 2003).

A committee of Division 16 of the American Psychological Association also is developing a rubric that may be useful (hereafter, APA coding criteria). Like the WWC Study DIAD, the APA coding criteria also focus on research about the causal effects of educational interventions. Several articles about the APA coding criteria can be found in the December 2002 issues of the *School Psychology Quarterly*.

B. Selecting Schools

1. Defining Eligible Schools

The schools included in the study need to support both the design team's recommendation to study well-implemented applications in a real-world setting, and policy goals such as improving student achievement in schools that receive Title I funds. The schools also need to support the study's legislative mandate that calls for the use of experimental methods to estimate effects. The specific recommendations and the selection criteria that can be used to support them include:

Recommendation: ***NCLB emphasizes improving academic achievement of students in Title I schools.***

Selection criterion: ***Schools that serve a high proportion of students who are eligible for free/reduced-price lunch.*** What constitutes a "high" proportion needs to be resolved. Setting the cut point too high may make it difficult to implement a technology in the poorest schools because those schools may not have the infrastructure to support it. Setting it too low would expand the number of eligible schools, but would also include more students who are not in poverty.

Recommendation: ***Study the effect of technology use.***

Selection criterion: ***Schools that have teachers with a demonstrated interest in using technology.*** The extent to which a school will use a technology package depends, in part, on teacher interest in using educational technology. Conducting the study in schools with teachers that are interested in using technology is likely to produce the desired evidence. The study needs to identify ways to assess teacher interest in using a technology package.

Have the infrastructure to support a technology package. This may include having a particular number of computers of a particular speed, a high-speed Internet connection, reliable technical support, willingness to train teachers to use a technology package, and so on. If ED or a vendor will supply computers and train teachers to use a technology package, the candidate schools only need to have adequate space and furniture to house the computers, security for the computers, electrical capacity, the possibility of installing a high-speed internet connection, resources to pay for reliable technical support, and so on.

Recommendation: ***Use experimental methods to estimate effects.***

Selection criterion: ***Schools that have a sufficient number of classes at a particular grade level to support random assignment.*** Randomly assigning classrooms requires that each school have more than one class at the grade level where the technology application will be implemented.

More schools are likely to meet these criteria than what is needed for the study. Therefore, the study will need to select a subset of eligible schools. The design team makes two recommendations for selecting the subset of schools, along with selection criteria that can be used to support them. One of the recommendations supports the study's legislative mandate that calls for understanding the conditions and practices related to technology's effectiveness; the other enhances the study's generalizability:

Recommendation: ***Study the conditions and practices related to technology's effectiveness.***

Selection criterion: ***Select several schools for each combination of conditions and practices that will be studied.*** As described in the previous chapter, understanding the importance of conditions and practices depends on having several schools for each of the condition and practice combinations that will be studied.

Recommendation: ***Enhance the study's generalizability.***

Selection criterion: ***Select schools from different parts of the country.*** The study's external validity (its generalizability) will be enhanced by including schools from a wide geographic area.

2. Identifying Candidate Schools

Two approaches could be used to identify candidate schools. One approach merges public-use data about schools with information technology developers are likely to maintain. Public-use data, such as the Common Core of Data, could be used to winnow out the universe of schools and reduce it to those that meet certain study inclusion criteria, such as schools that receive Title I funds to serve students from low-income households. However, public-use data cannot be used to further winnow the list of schools along other inclusion criteria, such as those that do not currently use technology, because that information is not available.

One way to help ensure that the schools included in the study actually use technology is to limit the candidate schools to those that have expressed an interest in using it. Technology developers may have information that could help further winnow down the list. Specifically, technology developers may maintain lists of the schools interested in purchasing their products, perhaps from databases that track requests for brochures. This information could be used to further reduce the list of schools obtained from the public-use data. The national study should work with sources in the industry to determine whether developers could provide any information that is useful for selecting schools.

The second approach that can be used to identify candidate schools involves contacting organizations that work with education agencies. For example, the Council of Chief State School Officers (CCSSO) assists states in improving their educational systems. The CCSSO tries to complete this mission in a number of ways, such as by creating partnerships that support excellence and equity in education, providing professional development, supporting efforts designed to increase student achievement, and collecting useful education data. In the course of carrying out these tasks, CCSSO staff may come into contact with individuals at the state level who are aware of districts or schools interested in using technology.

Conversations with districts and schools would center on the requirements of random assignment and other aspects of participating in the study. It is likely that incentives will have to be offered to schools in order to get them to participate in the study. Among the possible incentives is the donation of technology packages and licenses to participating schools for the duration of the study, including the provision of whatever standard professional development and technical support a particular vendor typically provides its customers. Another possible incentive is the provision of hardware and related infrastructure components to schools that enable them to operate the chosen technology package. The presumption is that this would be in the form of a donation to a school for at least the period of time they are participating in the study. A final incentive is providing to schools specific findings of the study pertinent to their interests and needs.

REFERENCES

Bloom, Howard S., Johannes M. Bos, and Suk-Won Lee. "Using Cluster Random Assignment to Measure Program Impacts." *Evaluation Review*, vol. 23, no. 4, 1999, pp. 445-469.

LeBlanc, Linda, and Dawn Thomas. "The Longitudinal Evaluation of School Change and Performance (LESCP) in Title I Schools: Report of Study Methods." Rockville, MD: Westat, October 2002.

Murphy, Robert F., William R. Penuel, Barbara Means, Christine Korbak, Alexis Whaley, and Jacob E. Allen. "E-DESK: A Review of Recent Evidence on the Effectiveness of Discrete Educational Software." Palo Alto, CA: SRI International, April 2002.

National Reading Panel, National Institute of Child Health and Human Development. "Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction (NIH Publication No. 00-4769)." Washington, DC: U.S. Government Printing Office, 2000.

Raudenbush, Stephen W. "Statistical Analysis and Optimal Design for Cluster Randomized Trials." *Psychological Methods*, vol. 2, no. 2, 1997, pp. 173-185.

Valentine, Jeff C., and Harris Cooper. *What Works Clearinghouse Study Design and Implementation Assessment Device (Version 0.6).* Washington, DC: U.S. Department of Education, March 2003.

APPENDIX A

TECHNICAL WORKING GROUP MEMBERS

Tim Best
Ohio Board of Regents
Rhodes Tower, 36th floor
30 East Broad Street
Columbus, OH 43215
phone (614) 387-1213
fax (614) 466-5866

Sandra Calvert
Georgetown University
309 C White Gravenor 37th &O Streets, NW
Washington, DC 20057
Phone (202) 687-3968
Fax (202) 687-6050

Doug Clements
State University of New York at Buffalo
Department of Learning and Instruction
Graduate School of Education
505 Baldy Hall
Buffalo NY, 14260
phone (716) 645-2455 ext. 1124
fax (716) 645-3161

John Cooney
Department of Educational Psychology
Campus Box 124
University of Northern Colorado
Greeley, CO 80639
phone (970) 351-1642
fax (970) 351-1622

Christopher Dede
Harvard Graduate School of Education
323 Longfellow Hall
13 Appian Way
Cambridge, MA 02138
phone (617) 495-3839
fax (617) 495-9268

Kathleen Fulton
National Commission On Teaching
and American's Future
2010 Mass Avc., N.W. Suite 210
Wahington, DC 20036
phone (202) 416-6187
fax (202) 416-6189

Anita Givens
Texas Education Agency
1701 North Congress Ave.
Austin, TX 78701
phone (512) 463-9400
fax (512) 463-9090

Dustin Heuston
Waterford Institute
1590 East 9400 South
Sandy, Utah 84093
phone (801) 576-4900
fax (801) 572-1667

Michael Kamil
Stanford University
School of Education
123 Cubberley
485 Lausen Mall
Stanford, CA 94305
phone (650) 725-5452
fax (650) 725-7412

Ken Koedinger
Carnegie Mellon University
3531 Newell Simon Hall
Pittsburgh, PA 15213
phone (412) 268-7667
fax (412) 618-5739

Cheryl Lemke
Metiri Group
1801 Avenue of the Stars, Suite 426
Los Angeles, CA 90067-5911
phone (310) 286-7944
fax (310) 286-7941

Christopher Lonigan
Florida State University
Department of Psychology
1 University Way
Tallahassee, FL 32306
phone (850) 644-7241
fax (850) 644-7739

Steven Ross
University of Memphis
Center for Research in Educational Policy
325 Browning Hall
Memphis, TN 38152-3340
phone (901) 678-3413
fax (901) 678-4257

Steven Sanchez
New Mexico State Department of Education
300 Don Gaspar, Room G-1
Santa Fe, NM 87501-2786
phone (505) 827-3644
fax (505) 827-7611

Appendix B

Estimates of Intra-Cluster Correlation Coefficients for Schools and Classrooms

Evaluations of educational interventions are often based on clustered data. For example, interventions often are provided to students who attend the same schools. Data for this sample are clustered because students within each school tend to be similar to each other.

When calculating the statistical power of a clustered design, an important factor in the calculation is the portion of variation in an outcome—such as test scores—that can be attributed to the cluster (hereafter, the intra-cluster correlation coefficient). This factor plays an important role because, given a particular sample, the minimum effect size that can be detected increases as the correlation coefficient increases.

This appendix presents estimates of the correlation coefficient based on a sample of students that is similar in many ways to the sample that will be used for the educational technology evaluation. Estimates were computed assuming that random assignment will occur at two different levels: (1) school and (2) classroom.

THE INTRA-CLUSTER CORRELATION COEFFICIENT

Assume that y_{ijk} represents the test score of student *k* in classroom *j* and school *i*, and can be written as:

$$y_{ijk} = \mu + \alpha_i + \gamma_{ij} + \varepsilon_{ijk}$$

where, i=1,2,…,s (s equals the number of schools in the sample); j=1,2,…,c_i (c_i equals the number of classrooms in school i); and k=1,2,…,n_i (n_i equals the number of students in classroom j and school i). Also assume that α_i, γ_{ij}, and ε_{ijk} are independent random variables with zero mean, and variance equal to σ_α^2, σ_γ^2, and σ_ε^2, respectively. This model assumes that the test score of student *k* in classroom *j* and school *i* (y_{ijk}) equals the sum of four components: (1) μ = the average test score of all students; (2) α_i = the difference in the average test score between all students, and students in school *i*; (3) γ_{ij} = the difference between all students in school *i*, and students in classroom *j* and school *i*; and (4) ε_{ijk} = the difference between all students in classroom *j* and school *i*, and student *k* in classroom *j* and school *i*. Based on this model, the variance of y_{ijk} equals $\sigma_\alpha^2+\sigma_\gamma^2+\sigma_\varepsilon^2$. The intra-school and intra-classroom correlation coefficients equal $\sigma_\alpha^2/(\sigma_\alpha^2+\sigma_\gamma^2+\sigma_\varepsilon^2)$ and $\sigma_\gamma^2/(\sigma_\alpha^2+\sigma_\gamma^2+\sigma_\varepsilon^2)$, respectively.

THE LONGITUDINAL EVALUATION OF SCHOOL CHANGE AND PERFORMANCE

We estimated the correlation coefficient using data from the Longitudinal Evaluation of School Change and Performance (LESCP). The LESCP was designed to examine whether learning improved after 1994 among students in high-poverty schools—an objective of Title I of the Elementary and Secondary Education Act, as amended in 1994. The LESCP design also paid particular attention to learning among those students with low levels of prior achievement.

The LESCP data contain 71 schools from 18 districts in 7 states. The sample of schools was not chosen to be nationally representative, but rather to include high- and low-poverty schools in states and districts that had enacted standards-based reform several years earlier. This criteria was used to select the schools because the provision of Title I enacted in 1994 encourages states, districts, and schools to pursue a standards-based approach for improving learning among students in high-poverty schools.

Data were collected during spring 1997, spring 1998, and spring 1999. Information was obtained from district administrators, school principals, teachers, and students. Table B.1 shows the students for whom data were collected during each of the three data collection years.

Table B.1
LESCP Student-Level Data Collection

Spring	Grade 3	Grade 4	Grade 5
1997	X	X	
1998		X	
1999		X	X

Student learning during each of the data collection years was measured using two of the Stanford Achievement Tests, Ninth Edition (SAT-9), subject tests: (1) math and (2) reading.

In each of these two subjects, an open-ended and a closed-ended test was administered, for a total of four tests during each of the data collection years.[14]

This data collection scheme makes it possible to examine math and reading achievement of three grade levels: (1) spring 1997 third graders; (2) the combined sample of spring 1997, spring 1998, and spring 1999 fourth graders; and (3) spring 1999 fifth graders. It also makes it possible to examine math and reading achievement growth of Spring 1997 third graders who remained in the sampled schools and progressed to the fourth and fifth grades—the shaded boxes in Table A.1. LeBlanc and Thomas (2002) provide more details about the LESCP design.[15]

ESTIMATES OF THE CORRELATION COEFFICIENT BASED ON THE LESCP DATA

We estimated the intra-school and intra-classroom correlation coefficient based on the test-score/grade-level combinations available in the LESCP data—that is, separately for the math and reading scores, and separately for third, fourth, and fifth graders. We used the scores that students received on the closed-ended tests in both math and reading. When analyzing the math score, only students who have a valid value for that score were included in the analysis. The same inclusion criterion was used when analyzing the reading score.[16]

The results indicate that the intra-school correlation coefficient equals about 0.12 when based on the LESCP third graders, and increases to about 0.15 and 0.18 when based on the LESCP fourth and fifth graders, respectively (Table B.2). These results are similar whether the calculation is based on the math or reading score.

An important issue is whether the correlation coefficient is smaller when achievement *gains* are analyzed, instead of achievement *levels*. Students within a particular school tend to have similar levels of achievement. However, achievement gains within a particular school may be more different across students, than achievement levels.

[14] Students were tested using the level of the SAT-9 that is appropriate for their grade. In particular, third graders were administered the "Primary 3 Level" test, fourth graders were administered the "Intermediate 1 Level" test, and fifth graders were administered the "Intermediate 2 Level" test.

[15] LeBlanc, Linda, and Dawn Thomas. "The Longitudinal Evaluation of School Change and Performance (LESCP) in Title I Schools: Report of Study Methods." Report submitted to the U.S. Department of Education, Office of the Under Secretary. Rockville, MD: Westat, October 2002.

[16] Students with a disability may be more likely to use an assistive technology, which the educational technology evaluation would not include. Therefore, students with an Individualized Education Plan (IEP) were deleted from the analysis. Students in 67 of the 71 LESCP schools met the sample inclusion criterion. The actual number of students included in the analysis depends on the measure of achievement (math or reading), but always included at least 2,710 students.

We examined this issue by analyzing achievement gains made by the LESCP third graders as they progressed to the fourth grade, and as they progressed from the fourth to the fifth grade. When analyzing third-to-fourth grade achievement gains, only students who have a valid value for both scores were included in the analysis. The same inclusion criterion was used when analyzing fourth-to-fifth grade achievement gains.[17]

Table B.2
Intra-Cluster Correlation Coefficient
Cluster=School
(standard errors are in parentheses)

	SAT-9 Score	
Grade-Level	Math	Reading
Third	0.12 (0.02)	0.11 (0.02)
Fourth	0.15 (0.03)	0.14 (0.03)
Fifth	0.16 (0.03)	0.18 (0.03)

Note: Author's calculations based on the LESCP data.

Using the math score, the results indicate that the correlation coefficient based on achievement gains is similar to the one based on achievement levels; however, using the reading score, the correlation coefficient is smaller when based on gains, than when based on levels (Table B.3). In Table B.2, we saw that the correlation coefficient equals between 0.11 and 0.18 (depending on the grade level) when we analyze reading levels. When we analyze reading gains, the correlation coefficient equals between 0.7 and 0.8 (Table B.3).

Our results thus far assume that the point of random assignment will be at the school level. We also produced results that assume the point of random assignment will be at the classroom level. Tables B.4 and B.5 present those results when we analyzed achievement levels and achievement gains, respectively.[18]

[17] Students in 67 of the 71 LESCP schools met the sample inclusion criterion. The actual number of students included in the analysis depends on the measure of achievement, but always included at least 1,721 students.

[18] The number of classrooms and students that met the sample inclusion criterion for the analysis of achievement *levels* depends on the grade level (third, fourth, or fifth) and measure of achievement (math or reading) analyzed, but always included at least 182 classrooms and 2,710 students. The number of classrooms and students that met the sample inclusion criterion for the analysis of achievement *gains* (which was limited to third graders) depends only on the measure of achievement (math or reading) analyzed, but always included at least 171 classrooms and 1,721 students.

Table B.3
Intra-Cluster Correlation Coefficient
Cluster=School
(standard errors are in parentheses)

	SAT-9 Gain Score	
	Math	Reading
3 to 4 Grade	0.16 (0.03)	0.07 (0.02)
4 to 5 Grade	0.14 (0.03)	0.08 (0.02)

Note: Author's calculations based on the LESCP data.

The intra-classroom correlation coefficient is smaller than the intra-school correlation coefficient—about 0.10 for the classroom level versus 0.12 to 0.18 for the school level. Like the school-level results, the intra-classroom correlation coefficient is smaller when we analyze achievement gains (about 0.04) instead of achievement levels, but only when based on the reading score.

Table B.4
Intra-Cluster Correlation Coefficient
Cluster=Classroom
(standard errors are in parentheses)

	SAT-9 Score	
Grade-Level	Math	Reading
Third	0.11 (0.02)	0.10 (0.01)
Fourth	0.11 (0.01)	0.09 (0.01)
Fifth	0.10 (0.02)	0.11 (0.03)

Note: Author's calculations based on the LESCP data.

Table B.5
Intra-Cluster Correlation Coefficient
Cluster=Classroom
(standard errors are in parentheses)

	SAT-9 Gain Score	
	Math	Reading
3 to 4 Grade	0.08 (0.02)	0.03 (0.01)
4 to 5 Grade	0.09 (0.02)	0.04 (0.01)

Note: Author's calculations based on the LESCP data.