Stronger Designs for Research on Educational Uses of Technology: Conclusion and Implications

Geneva Haertel and Barbara Means
SRI International

No information system or database maintained today, including the National Educational Longitudinal Study (NELS) and the National Assessment of Educational Progress (NAEP), has the design and content adequate to answer vital questions about technology’s availability, use, and impacts on student learning. NAEP, for example, while suitable for its primary purpose of collecting achievement data, is flawed as a data source for relating achievement to technology availability and use (see the paper by Hedges, Konstantopoulos, & Thoreson) The NAEP design is cross-sectional and thus unsuitable for revealing causal relationships between technology and student achievement; its survey questions are inconsistent across surveys in different years or subject areas and insufficiently specific about technology use. Thus, piggybacking a study of technology use and impact on NAEP as it exists today is unlikely to produce the type of unambiguous information that is needed about the impact of technology on student learning.

Given the insufficiency of current large-scale data collections for answering questions about technology effects, ten research methodology experts were commissioned to write papers providing guidance for a major research program that would address these questions. (See Table 1 for a list of authors and paper titles.) This synthesis uses the key arguments and convictions presented in the ten commissioned papers as a basis for making recommendations for educational technology research approaches and research funding priorities. Our discussion centers around the area of technology research that is regarded as both most important and most poorly addressed in general practice—the investigation of the effects of technology-enabled innovations on student learning. We have looked for points of convergence across the commissioned papers and have used them as the basis for our recommendations. Our synthesis is based on the ideas within the individual papers and those discussed at the authors’ design meeting held at SRI in February 2000. The interpretation and synthesis are our own, however, and individual
paper authors should be “held harmless” of responsibility for the design and policy implications we have drawn from their work.

**Cross-Cutting Themes**

Three themes appeared and reappeared in nearly all of the commissioned papers. The first and most prevalent theme was the need for new assessment approaches to measure student learning outcomes that are not well represented on traditional standardized achievement tests. Two other recurring themes were the call for careful measurement of implementation and context and the advantages of conducting coordinated or clustered studies that share approaches, measurement instruments, and research infrastructure. In the remainder of this section, we will treat the first of these cross-cutting themes at some length and touch on the latter two more briefly because they will be covered at greater length when we discuss proposed research strategies.

**Need for New Assessment Approaches to Measure Outcomes**

In the past, evaluations of technology effects have relied heavily on norm-referenced, standardized tests as learning outcome measures. While standardized achievement tests may be effective measures of basic skills in reading and mathematics, they generally do not tap higher-level problem-solving skills and the kinds of deeper understandings that many technology-based innovations are designed to enhance. Many technology-based interventions were designed around constructivist theories of learning. These interventions often have goals for students that include the production of enduring understandings, the exploration of essential questions, the linking of key ideas, and rethinking ideas or theories. The instructional activities that accompany these interventions focus on increasing students’ capacity to explain, interpret, and apply knowledge in diverse contexts. Standardized achievement tests are ill-suited to measuring these types of knowledge. Evaluations of technology effects suffer from the use of scores from standardized tests of content unrelated to the intervention and from the substitution of measures of opinion, implementation, or consumer satisfaction for measures of student learning.
Evaluations of technology-supported interventions need a wide range of student learning measures. In particular, performance measures that can more adequately capture outcomes of constructivist interventions are needed. Measures within specific academic subject areas might include level of understanding within the subject area, capability to gain further understanding, and ability to apply knowledge in new contexts. Other competencies that might be assessed are relatively independent of subject matter; for example, acquiring, evaluating, and using information and collaboration, planning, and leadership skills.

The papers by Becker and Lovitts and by Mislevy et al. anticipate the nature, as well as some of the features, of new learning outcome measures. While any single learning outcome measure is unlikely to incorporate all of the features specified below, many will include several.

The new learning assessments should include:

- Extended, performance tasks
- Mechanisms for students to reveal their problem-solving, to describe their rationale for proceeding through the task, and to document the steps they follow
- Opportunities to demonstrate social competencies and collaboration
- Scoring rubrics that characterize specific attributes of performance
- Scoring rubrics that can be used across tasks of varying content
- Integration with curriculum content
- Links to content and performance standards
- Content negotiated by teachers

These features are delineated in more detail, and contrasted with the characteristics of traditional standardized tests, in Table 2. Exhibit 1 described a prototype performance assessment task incorporating many of these features.

Even as consensus builds among the educational research and assessment communities that new measures of this sort are needed, policymakers are likely to want to continue using standardized test results to inform their decision-making. As a practical matter, we recommend including data from standardized tests as part of the array of outcome measures collected in evaluations of educational technology. This is not to say
that improvement in standardized test scores is a goal of every intervention, but rather a recognition of the importance of demonstrating that any improvements observed on assessments of higher-order skills are not obtained at the expense of the more basic skills measured on standardized tests.

**Development and Validation of Measures.** Principled design of questions, items, and tasks should be applied to the range of measures used in evaluations of technology-supported interventions. (This admonition applies not only to content assessments designed to measure student learning outcomes, but to surveys, implementation, and context measures, as well.) A clear example of the rationale, reasoning, and procedures followed in principled assessment design is provided in the Mislevy et al. paper. As that paper illustrates, principled assessments are based on a chain of reasoning from the evidence observed to what is inferred. Building, in part, on Samuel Messick’s (1989) concept of validity, this construct-centered approach guides the construction of relevant tasks and provides a rationale for the development of scoring criteria and rubrics. In studies of student-learning effects associated with new technologies, attention must be paid to the technical qualities of the outcome measures used, especially the performance assessments.

The need to obtain and score more complex performances notwithstanding, Shavelson and his colleagues (Shavelson, Baxter, & Gao, 1993; Shavelson, Baxter, & Pine, 1991) have documented the limited generalizability and reliability of performance assessment scores for individual students. This problem is of great importance when scores from performance assessments are being used to make decisions about the education of individual students. However, when the assessments are designed primarily for research purposes and the results will be aggregated across groups of students and not used to influence the educational history of individual students, relaxing the reliability requirements is of less consequence. Becker and Lovitts present this argument as part of the rationale for their project-based assessment design.

**Technology Affordances for Improved Assessment of Student Learning.** Technology has the capacity to “break the mold” of traditional assessment by supporting the development of new assessment forms that can measure higher-level inquiry
processes (Quellmalz & Haertel, submitted for publication). In the area of science inquiry processes, for example, technology can support ways to present and measure the entire range of inquiry processes from generating research questions, to planning an experiment, conducting the experiment, collecting and organizing data, analyzing and interpreting data, drawing conclusions and communicating results. New technology-based assessments, such as those developed by VideoDiscovery or Vanderbilt’s Cognition and Technology Group (CTGV) reveal how technology can be a means for students to solve science inquiry problems that they could not do in a hands-on situation because of issues of scale, complexity, expense, risk, or timeframe. In these computer-based assessments, students are presented with challenging content, authentic tasks, and a resource-rich assessment environment.

Becker and Lovitts devote considerable attention to the question of whether evaluations of technology effects should employ student assessments that permit students to use the technology in the assessment task or whether a conventional, minimum-resource “standardized” testing environment should be maintained. A minimum-resource assessment environment denies computer-capable students the ability to demonstrate important competencies they may have acquired as a result of the intervention. Becker and Lovitts point out that typical measures used in technology evaluations omit tasks where computers may be important tools for the accomplishment of the task. (Recent analyses of writing assessment data by Russell, 1999, suggests that students who are accustomed to composing on computers attain scores a full grade level higher when their writing is tested on computer than when it is tested in a paper-and-pencil exam.) Innovative technology-exploiting curriculum development projects, on the other hand, often define assessment tasks that require technology-specific competencies, thereby rendering comparisons with students without computer experiences irrelevant or unfair. Becker and Lovitts resolve this dilemma by defining outcome competencies and skills at a level of abstraction that permits computers to be used but does not require their use.

As Mislevy et al. point out, technology has many affordances that can transform the entire assessment process, not only task presentation. Technology can alter the way assessments of student learning are constructed, the management of the assessment
process, how tasks are scored, the extraction and evaluation of key features of complex work products, and the archiving of results.

**Need for Better Measures of Context and Implementation**

In studies of technology’s effects, as in all intervention or reform efforts, it becomes important to determine not just that an intervention *can* work but the circumstances under which it will work. Most paper authors stress the need for better and more comprehensive measures of the implementation of technology innovations and the context or contexts in which they are expected to function.

Rumberger, Means et al., Lesgold, and Culp, Honey, and Spielvogel articulate the need for studies to be sensitive to key contexts, such as the household, classroom, school, district, community, and state. Some of the understandings about context that need to be incorporated in research include:

- Technology is only one component of the implementation and usually not the most influential.
- Technology innovations must be carefully defined and described.
- Great variation occurs in students’ exposure to technology due to their participation in classes of teachers who differ in their levels of “technology comfort” and in the supports they have for implementing technology.

Paper authors place a premium also on careful definition and measurement of the technology innovation as it is implemented. Candidates for use in measuring implementation include surveys, observations, interviews, focus groups, teacher logs, records of on-line activity, and document reviews. Paper authors recommend combining various methodologies in order to increase the richness, accuracy, and reliability of implementation data.

The first line of attack in determining the effectiveness of an innovation is to document the way it is introduced, how teachers are trained to use it, whether resources are available to support its use, and the degree to which teachers faithfully implemented it. Lesgold stresses the importance of gathering data describing teacher professional development when documenting the implementation of an technology-based intervention.
Information should be collected on the nature, quality, and amount of professional development that was made available to teachers and the degree to which they took advantage of the resources. In addition the availability of resources to support the innovation and timely delivery of innovation materials should also be documented.

**Need for Clustered Studies**

Numerous authors recommended coordinating evaluations of related technology innovations or issues. Although we have almost as many variants of this idea and as many new terms (“partnership research,” “firms,” “testbeds,” “embedded experimental studies within a larger sample,” “heterogeneity of replication model,” and “sentinel schools”) as papers, all exemplify the desire for integrating a series of studies.

No single study, by itself, can disambiguate the relationship among the many influences that affect student and teacher outcomes in a myriad of relevant contexts. Thus, most of the paper authors envision a program of inter-related studies to be linked not only to prior research, but also to other studies that would be conducted in tandem or in sequence as part of a more comprehensive research agenda.

[Hedges, Konstantopoulos, & Thoreson](http://example.com) propose a network of “sentinel schools.” This network is similar, in purpose and design, to [Lesgold’s](http://example.com) “testbeds;” [Moses’](http://example.com) “firms;” [Culp, Honey, and Spielvogel’s](http://example.com) “partnership research;” and [Means et al.’s](http://example.com) “embedded studies.” Each of these arrangements would provide an opportunity for researchers, practitioners, and policymakers to design, conduct, and collaborate on a family of studies and to share their results. Such arrangements could provide evidence of emerging trends and could make available to researchers a set of study sites willing to participate in sustained studies of technology effects.

A corollary of the proposed establishment of programs of inter-related studies is the need for “intermediary organizations.” Such organizations would provide the infrastructure to support the inter-related program of studies. This type of organization was most fully described by [Culp, Honey, and Spielvogel](http://example.com). Intermediary organizations could provide a variety of research functions such as reviewing existing research, identifying research questions, synthesizing results from other studies being conducted,
creating templates or forms for data collection instruments, and supporting local researchers in their efforts.

Intermediary organizations and networks of participating schools would bring together the resources of school systems, research organizations, universities, and government agencies. Such a consortium of collaborating institutions would provide the multiple capacities needed to achieve the overall goal of conducting programmatic research to determine the impacts of technology on educational outcomes. Target populations of students and teachers would be present and readily accessible. Manpower would be available to gather, score, and code large amounts of data, if needed. Given agreement on core sets of context variables, the intermediary organization could make available data collection instruments for use across multiple studies. The methodological expertise needed to conduct rigorous research on learning technologies could be made available to all participating research organizations. The nature of the arrangement would be conducive to disseminating new knowledge to diverse target audiences, including practitioners, researchers, and the policy communities. The primary purpose of the intermediary organizations, however, would be support of quality research in this area, as opposed to the professional development, research dissemination, and technical assistance functions of today’s Regional Technology in Education Consortia (R-TECs).

**Multiple and Complementary Strategies**

Paper authors were in agreement that multiple and complementary research strategies are needed to measure the implementation and impact of learning technologies. No single study, genre of studies, or methodology is adequate to the task. While formative studies provide information to refine particular technology innovations, the evaluation of technology’s effects requires studies of mature innovations that have been implemented in diverse settings, including schools in high-poverty neighborhoods and schools that are not atypically rich in technology resources and support systems.

As a group, and for the most part individually, the authors embrace:

- Collection of both qualitative and quantitative data
• Assessment of a wide range of student learning, attitude, and behavioral outcome measures
• Assessment of both context and implementation, as well as the primary intervention
• Design of both small- and large-scale studies

Across the range of papers, no single research strategy was endorsed as most promising. As we reviewed the commissioned papers, three promising general strategies for research designs emerged:

• **Multiple, Contextualized Evaluations** bring together the qualities of contextualized research and the strategy of clustering studies that were two of the cross-cutting themes discussed above. The linking of multiple intensive studies of technology effects in schools and classrooms would be supported by an intermediary organization helping to orchestrate consensus-building around variables and methods, providing infrastructure, and storing and analyzing data from all the evaluation studies.

• **Multi-Level, Longitudinal Research.** This research strategy takes into account: (1) the multiple contexts of students’ learning environments, (2) the innovation’s cumulative effects, and (3) the direct and indirect effects of contextual variables on outcomes and implementation.

• **Random-Assignment Experiments.** Random assignment of subjects to control and experimental groups is a necessary criterion of a true experiment. With this design, one can be more certain than with any other design about attributing cause to the independent variables. The greatest limitation may be external validity (i.e., the ability to generalize results beyond the explicit conditions of the experiment).

Each of these three designs was recommended, in one form or another, by multiple authors, although no single design was the method of choice for all the authors. Below we provide a description of the characteristics of each design and its application in studies of educational technology.
Multiple, Contextualized Evaluations

Culp, Honey, and Spielvogel; Baker and Herman; Moses; Means et al.; and Lesgold all advocate some form of linked, contextualized evaluations. We begin our discussion of this strategy with a consideration of the concept of context in studies of educational uses of technology.

Meaning of Context. In most educational research studies, context is the environment within which the student learns. Context influences both the implementation of the technology intervention and its impact. In educational technology studies, context includes: (1) the vision of the innovation and its perceived value; (2) the physical facilities available, in particular the technology infrastructure of the class, school, and district; (3) the availability of resources, such as staff with technology expertise, curricular materials, and time allocated to technology-based instruction; (4) the climate toward technology, learning, and educational reform that exists in the classroom, school, and district; (5) degree of support from leaders regarding the technology innovation; (6) school board policies that shape technology use; and (7) demographic characteristics of the classroom, school, or community organization, as well as the student’s home. Traditionally, the demographic characteristics of the classroom, school, and community included in education research encompass geographic location and urbanicity or district size as well as ethnic, linguistic, economic, and gender compositions.

Lesgold sets forth maturity models as an approach to documenting the context that surrounds an innovation. A maturity model is a hypothetical set of stages of quality (or “levels of maturity”) through which an individual, product, or organization advances. Lesgold enumerates four dimensions of context that can be described in terms of level of maturity: instructional, technology infrastructure, educational software product, and people maturity. Each stage in a maturity model is defined as a set of features that can be scored using a rubric. The rubrics and stage definitions support reliable determination of the level of maturity that has been attained on the dimension of context under study. In Lesgold’s approach, an evaluation would establish empirically the levels and forms of maturity that influence the extent to which the technology innovation under investigation
results in improved learning. Maturity models can be combined with almost any of the design approaches recommended by other authors.

Baker and Herman’s distributed evaluation model is a good illustration of the range of data that can be gathered to capture the educational experience being delivered at a local site. Among the types of data that might be collected as part of this approach are: assessment data reflecting students’ cognitive performances, archival data on student outcomes, and common and site specific implementation measures, all of which would be integrated with relevant standards and local goals.

For both Baker and Herman and Culp, Honey, and Spielvogel, the role of local engagement, collaboration, and feedback is paramount. Both point out that local school communities need support to think about evaluative questions and evidence. Teachers and administrators at the local site should be participants in, rather than recipients of, the evaluation. In such cases, information generated by the evaluation is particularly valuable for users of the innovation and for program managers, who gain information to support reflection on their experiences and identification of promising paths toward successful change. These authors conclude that evaluation research that is responsive to local concerns, constraints, and priorities can be structured and synthesized to produce knowledge about effective uses of educational technology that has high face validity within local communities and still informs wider research, practitioner and policy audiences.

Arguments for a Focus on Context. Members of the evaluation community disagree on the value of conventional, formal evaluation studies (i.e., those third-party evaluations conducted by individuals at a distance from the innovation or intervention). Stake (1967) has been skeptical for years about the ability of such formal evaluations to help stakeholders make decisions about program continuation or refinements. Culp, Honey, and Spielvogel assert that large-scale and summative evaluations have not traditionally been expected to answer questions about why an outcome occurred. Only designs that are highly contextualized—that include the “why” question from the start—will be able to inform decision-making about the effectiveness of technology in educational settings.
For many policymakers the decision to be made is not whether to invest in technology or not, but rather how best to integrate technology with local educational goals. Highly contextualized evaluations are well suited to this purpose. They can be responsive to local needs because they typically produce descriptive, complex models of the role that an intervention or program plays in the existing system and how effectively it matches the system’s needs and resources. These models can help practitioners make informed decisions about technology implementations. Exhibit 2 provides an illustration of an evaluation effort designed to serve local needs while connecting with broader research and policy issues.

The most effective evaluations produce both research-based knowledge of what kinds of technology-based innovations work best in what educational environments and practice-based knowledge of how the technology-integration process can meet the learning goals of classrooms or schools with particular characteristics.

Need for Multiple, Linked Contextualized Evaluations. To advance what we understand about technology use and effects, the results of multiple, contextualized evaluations must be combined and accumulated. The intent in such an effort is not to find uniform results, but rather to aggregate findings across studies to enable inferences relating features of contexts to successful or unsuccessful implementations and degrees of impact. An existing methodology, meta-analysis, is available for aggregating quantitative results across studies (Hedges & Olkin, 1985; Lipsey, 1993). Other methodologies have been used to produce qualitative summaries of multiple studies. Cross-case analysis methodologies, used to synthesize case study data or ethnographies from multiple sites, can be used to aggregate qualitative data. Elements of both of these methodologies may be required to combine results of these multiple, contextualized evaluations. Whatever approach is used to combine the multiple and cumulative results of the contextualized evaluations, it would be greatly facilitated by the use of common definitions, a common framework, and common instruments and data collection procedures. It should be noted also that simply applying these techniques to the extant research base will not suffice. As Baker and Herman point out, the studies submitted to meta-analysis fall far short of a representative sample of current educational uses of
technology. An array of contemporary, parallel studies capturing current and emerging technology practices is a prerequisite for this approach.

**Multi-Level Longitudinal Studies**

Newer forms of technology, many of which are student-centered, constructivist paradigms (e.g., exploratory computer environments, communication programs that promote collaboration among students), require teachers who are skilled users of the technology and school and district infrastructures that support technology-based innovations through resource allocation and a climate of collaboration. Thus, research to evaluate the use and effects of educational technology must take into account the fact that it is a multi-level phenomenon that operates at the district, school, classroom, and student levels. In addition to the need for multi-level modeling, evaluations of technology use and effects must also take into account the short- and long-term outcomes of these innovations. Research tells us that there is a substantial period of time needed for a technology innovation to become well enough integrated into a school setting to be used to its greatest advantage (Sandholtz, Ringstaff, & Dwyer, 1996). We want to evaluate technology-based innovations when they are mature and to continue data collection for several years to determine whether the outcomes they generate are sustained. Especially in cases where we cannot randomly assign students or classes to treatments, there is a need to measure desired learning outcomes at multiple points in time. The use of multi-level modeling and longitudinal designs can address these concerns. Papers by Rumberger, Means et al., and Lesgold advocate the use of these designs and statistical methodologies.

*Rationale for Incorporating Multiple Levels.* Empirical investigations of educational innovations need to take into account the hierarchical nature of school systems and the multiple, overlapping contexts in which learning occurs. Students are nested within classrooms within schools within districts. As individuals, these same students are members of families, peer groups, and communities. Technology may be used and may exert an influence within any of these contexts. Sophisticated modeling
approaches, such as hierarchical linear modeling (HLM) were designed to address such conditions.

**Importance of the Longitudinal Design.** By collecting outcome, contextual, and implementation variables at multiple points in time, we can provide a basis for ruling out many alternative hypotheses concerning differences in outcomes. This is especially important in educational settings, where the use of experiments with random assignment has been infrequent. Although the longitudinal elements of the design do not confer the benefits of random assignment, the use of a longitudinal design where each student serves as her or his own control does eliminate many threats to the validity of the study. Exhibit 3 describes a widely disseminated longitudinal study of the results associated with a statewide technology implementation effort. Differences in student growth over time, when based on large samples, can be related to the influence of the intervention and contextual variables at multiple levels of the educational system. (An analytic approach not used in the study described in Exhibit 3.) If probability samples are used in multi-level longitudinal studies, then the results can be generalized to the populations from which the samples were drawn, an important advantage over the typical random-assignment experiment. (See Means et al.; Hedges, Konstantopoulos, & Thoreson; and Rumberger.)

**Specifying a Multi-level Framework to Guide Study Design.** To guide the design of a multi-level, longitudinal study, a comprehensive, conceptual framework should be developed. Such a framework would describe the multiple levels and contexts of the educational system. It would also articulate the inputs, processes, and outcomes that comprise the technology innovation’s theory of change. Context variables and implementation milestones and measures should be specified, as well. Presumed causal and temporal relationships among the influences, processes, and outcomes might also be depicted.

Three of the commissioned papers present frameworks that specify the relationships among influences, contextual variables, implementation, and outcomes in different ways. Rumberger’s framework is similar to those used in economics and studies of educational productivity. In such studies, relationships are drawn among the system’s inputs,
processes, and outcomes. Rumberger’s framework is depicted graphically in a way that demonstrates the many interconnections among the levels and outcomes of the system. Means et al.’s conceptual framework identifies several contexts, the variables that would be measured in those contexts, and a comprehensive list of outcomes to be measured, as well. Lesgold puts forth a preliminary causal model, which relates different levels of the educational environment to context measures and outcomes. Indirect and direct influences of independent variables are specified.

The authors agree that the technology innovation itself should be carefully defined as part of the framework. Beyond providing a complete description of the innovation, the relationship of the innovation to other instructional activities and materials that are part of the instructional package must be specified. In addition, the short- and long-term outcomes that are expected to occur as a result of the innovation should be specified. If there are interim outcomes that must occur, they should also be identified. This is especially important if those interim outcomes are related to the technology intervention’s implementation at several levels of the educational system.

The framework should also take into account the implementation of the innovation. Measures of implementation should go beyond simple checklists. In particular, the amount, nature, and quality of professional development provided to teachers should be documented, using both qualitative and quantitative methods. Amount of hands-on training, availability of just-in-time support, and the technology infrastructure available in practice to support the implementation should be described.

**Random-Assignment Experiments**

Random assignment is widely acknowledged as the premier research method for making an unambiguous case that a given treatment caused an observed difference in outcome. While experiments conducted with random assignment are a mainstay of research methods in medicine, agriculture, and psychology, such experiments have rarely been used in education. Among the paper authors, Cook and Moses argue that it is time for educators to reconsider their disaffection with this powerful methodology.
Educators’ Arguments against Experiments with Random Assignment. There are several reasons for educators’ disaffection with random experiments. One key source is skepticism about the application of rigorous, quantitative methods to the study of complex social environments, such as schools and classrooms (Fetterman, 1984; Guba & Lincoln, 1982; House, 1993; Stake, 1967). Many educational evaluators reason that the theory of causation that buttresses experiments is too naïve; experiments are too difficult to implement in school contexts; experiments require trade-offs that lower the quality of answers to non-causal questions; the information provided by experiments is not relevant to policy decisions; and the information that is needed for evaluation purposes can be gained using different and more flexible methods. Because of the perceived and real difficulties with “true” experiments, researchers have put forth alternative designs, such as intensive case studies, causal modeling, quasi-experiments, and “design experiments.” These alternative designs, many education researchers argue, have advantages that offset the true experiment’s capacity to provide evidence of cause and effect relationships. (Among the commissioned papers, this position is articulated most clearly by Culp, Honey, and Speilvogel; Lesgold; and Rumberger). Cook provides a thorough and fair-minded appraisal of the concerns that education researchers have raised regarding the appropriateness and feasibility of experiments in education.

Arguments in Favor of Experiments with Random-Assignment. Adherents of experimental research respond that no alternative research design provides as convincing a causal counterfactual as the randomized experiment. Random assignment is the only method that can lay claim to testing what would have happened to the treatment group if it had not received the treatment. To summarize the nature and key points of the argument surrounding the use of random assignments in education, we present Table 3, entitled “Arguments For and Against Random Assignment.” The table presents a series of arguments made by those who are skeptical about the feasibility and value of experiments in education and rebuttals to the arguments drawn from Cook’s paper.

To overcome the lack of contextual and implementation information that is used in experiments, mixed-methods designs can be employed. Such designs, incorporating qualitative and quantitative measures in the same study, can supplement what the experiment reveals about program theory, implementation, and context. Since many of
the objections to random assignment experiments were identified through experience in school-based studies; we conclude that random-assignment experiments can be done in education, but implementing them remains a significant challenge. Exhibit 4 describes a study that the American Institutes for Research identified as “the closest to an experiment” in the last five years of research on technology’s impact on student learning (B. Lovitts, personal communication). The study is not a true experiment, however, because the participating districts insisted on selecting the classrooms for the experimental and control groups, rather than allowing them to be assigned at random.

Cook and Moses, as adherents of the use of randomized experiments, leave us with a trinity of convictions about the value of experiments. They argue that experiments have: (1) greater validity (specifically related to the presence of a counterfactual), (2) greater efficiency at reaching defensible causal conclusions compared to other methods, and (3) greater credibility in the scholarly and policy communities. Cook cautions that a single, large experiment may not be a feasible or desirable course of action when addressing the effects of technology on student outcomes. He offers a “heterogeneity-of-replication” model, in which many smaller experiments are conducted in different settings, time periods, and regions of the country, using different operationalizations of cause and effect.

Matching Methods to Purposes

Some recent documents and policy pronouncements have implied that a single theoretical approach is the only defensible stance for education research. Both the Office of Management and Budget (OMB) and members of Congress have made assertions to the effect that only experimental studies, employing random assignment to treatment and control groups, constitute scientifically defensible quantitative research methods. The House bill to reauthorize the Office of Educational Research and Improvement (H.R. 4875), for example, contains language calling for “scientifically based quantitative research” to obtain “understanding of the truth of a particular educational theory, practice or condition.” The legislation goes on to define scientific research as studies “in which individuals, entities, programs, or activities are assigned to different conditions with
appropriate controls to evaluate the effects of the conditions of interest through random assignment experiments, or other designs to the extent such designs contain within-condition or across-condition controls.”

The rising call for true experiments in education research was reflected also in a forum (“Can We Make Education Policy on the Basis of Evidence? What Constitutes High Quality Education Research and How Can it Be Incorporated into Policymaking?”) held at the Brookings Institution in December 1999. At the Brookings meeting, Tom Cook and Robert Boruch documented the scarcity of random-assignment experiments in educational research and argued that such experiments are both necessary to obtain sound evidence of causal effects and far more likely than quasi-experiments or other designs to influence policymakers’ decisions.

In striking contrast to the dominant tenor of the Brookings dialogue and H.R. 4875, the March 1999 advisory report submitted by the National Academy of Education (NAE) to the National Educational Research Policy and Priorities Board (the Congressionally mandated oversight body for the Office of Educational Research and Improvement) asserts that “progress toward high achievement for all students has been impeded by the belief that research, students’ learning, and teachers’ learning can be studied in isolation from important matters of context” (p. 8). The NAE report calls for what they term “collaborative problem-solving research and development,” which is defined as efforts focused on “solving specific current problems of practice and at the same time. . . developing and testing general principles of education that can be expected to apply broadly beyond the particular places in which the research is done” (p. 9). These projects would be joint efforts between researchers and professional educators to combine improvements in practice and research—intense collaborations difficult to reconcile with the tenets of random-assignment experiments.

The debate over methods reflected in these documents and proceedings may instigate a healthy dose of reflection and questioning of assumptions within the educational research field. For federal agencies, we would recommend an eclectic perspective with

---

1 The transcript is available at: http://www.brook.edu/comm/transcripts/19991208.htm.
respect to research methods rather than championing of one approach or another as “the
gold standard.” As the discussion in the introductory chapter of this volume suggests,
studies of technology-supported educational practices are performed in many different
contexts for many different purposes. The degree of definition and control of the
practices under study differs markedly from case to case. We simply do not believe that
any one research approach will cover all cases. Rather, we recommend an effort to
clarify the purposes, constraints, and resources for any given piece of research or research
program as a basis for choosing among methods.

Although there are no cut-and-dried rules for when to choose which method, we will
try to elucidate some general rules of thumb based on our own and others’ experience.
We have organized the discussion below in terms of broad categories of research goals
and circumstances with implications for the choice of methods. Like any categorical
scheme, ours is an over-simplification that sounds neater in theory than it is in practice.
Nevertheless, we have found the distinctions useful in matching research methods to
purposes. The over-arching distinction in our scheme is between investigations of the
workings and effects of specific projects (what we have called “project-linked research”)
versus studies of a range of “naturally occurring practices.” In the first case, a particular
initiative, approach, or project has been defined and is the focus of the research. In the
second case, the researcher is seeking to understand “what’s out there,” defined in terms
of practices or access to technology rather than examining a particular project or funding
stream

**Project-Linked Research**

For simplicity’s sake, we refer to this category simply as “project-linked” research, but
we intend the term to include any defined innovation, regardless of whether or not the
implementers of the innovation share formal membership in or funding from a given
project. Examples in the educational technology area would include the GLOBE
program, in which students and teachers collect scientific data on their local
environments and submit their data to a central program-run Web-based data archive; the
adventure learning resources offered by the Jason Foundation; and the Generation WHY

---

2 We can relate our scheme to what may be a more familiar distinction between evaluation and research: Evaluations, and certainly the narrower classification of program evaluations, are “project-linked,” but there are many project-linked studies that would not qualify as evaluations.
Technology Innovation Challenge Grant that trains students to provide technical support and consulting for teachers who want to use technology in their instruction.

**Early-Stage Projects**

In the case of evaluation studies conducted in conjunction with an evolving technology-supported innovation, contextualized evaluation studies will usually be the method of choice. At this early stage of work, it is important to understand how the innovation plays out in real classrooms, and the evaluator needs to be alert to unintended interactions with features of the environment that program designers may not have taken into consideration. Providing useful feedback to program developers and developing an understanding of project implementation in context—that is, how the elements of the innovation influence teacher and student behavior—will be paramount concerns at this stage. Exhibit 5 describes a project-linked, formative evaluation of an early-stage innovation.

Our methodologists’ papers would suggest, however, that where possible, these evaluations should be conducted using common instruments and outcome measures and within a consortium that shares and aggregates data from individual projects. Such a step would make it much easier to achieve a higher, more uniform level of quality across individual evaluations and to combine findings across studies. Thus, if a funding agency were to follow this recommendation when launching a new school technology initiative on the order of the Technology Innovation Challenge Grants or Preparing Tomorrow’s Teachers to Use Technology (PT3) program, they would solicit proposals addressing one or more pre-selected types of outcomes (e.g., early literacy or mathematics problem-solving skills) and require use of some agreed-upon instruments for documenting contextual variables and for measuring key classroom processes and outcomes.

The National Academy of Education (NAE) made a similar recommendation for coordinating studies in its recent report to the National Educational Research Policy and Priorities Board. “The recommendations include supporting federations of problem-solving research and development projects, linked in a hub-and-spoke relationship. The goal would be simultaneously to develop improved educational success in specific
settings (the spokes) and to identify issues of common concerns [sic] and to carry out theoretical analyses and construct tools that are supported by and facilitate the work of the several projects in integrative ways (the hub)” (p. 11).

Another point about these studies, made strongly by Lesgold is that it is important to study an innovation in a range of contexts, including those most critical from a national policy perspective, and to measure elements of the context within which each implementation occurs. From a policy perspective, critical contexts include classrooms serving students from non-English-speaking or economically impoverished backgrounds, students with disabilities, and schools low in technology resources. Almost any approach produces good results in some settings with some kinds of students and supports. Before recommending particular approaches for broader implementation, we need a basis for understanding the range of contexts within which desired results are and are not likely to be forthcoming.

**Mature Projects**

As individual projects become more mature and more widespread, there will be cases where further research is warranted. By a mature project, we mean one where the intervention has been fairly well specified, such that its elements can be delineated and an observer can make judgments as to the extent to which they are being implemented. Further, mature projects are ones whose model for producing desired changes is understood, at least in theory. That is, the innovation is not just a black box placed between inputs and outputs. There is some understanding of what classroom elements or processes the inputs are supposed to alter and of how those altered processes (or interim outcomes) produce the targeted student outcomes that are the project’s ultimate goals.

The question raised by the recent debate among national policymakers and discussed intensively at our authors’ meeting is whether the random-assignment experiment is the method of choice when the research question involves a mature innovation’s effects. Several of our authors (Cook and Moses) strongly support the position that the experiment is the only unimpeachable source of information about causal relationships and that such experiments are eminently feasible within the educational domain. While there was general agreement among authors that random-assignment experiments are
When Random Assignment is Preferred. As we have grappled with the issue of the value and feasibility of random-assignment experiments for studies of technology’s effects on students, we have found the points made by Judy Gueron at the Brookings forum cited above extremely helpful. Gueron addresses the issue of when random-assignment experiments are more and less appropriate and feasible on the basis of her experience at the Manpower Research and Demonstration Center (MDRC), an independent research organization known for its running of large-scale field trials, principally in the employment and training arena. Based on MDRC’s experience running 30 major random-assignment experiments over the last 25 years, Gueron provides eight guidelines for determining when random assignment designs are appropriate:

- The key question is one of program impact.
- The program under study is sufficiently different from standard practice and you can maintain the distinction over time.
- You are not denying anyone access to an entitlement.
- You are addressing an important unanswered question.
- You include adequate procedures to inform program participants in advance and to insure data confidentiality.
- There is no easier way to get a good answer.
- Participants are willing to cooperate in implementing the assigned conditions.
- Resources and capacity for a quality study are available.

We believe that Gueron offers a useful set of guidelines, some of which will be easier and some harder to achieve in designing studies of the impacts of technology-supported educational innovations. Questions of program impact are likely to be less central in research on newly developed (or developing) technology-supported innovations. They are likely to be regarded as critical, however, in cases of well-established innovations, particularly those that are candidates for wide implementation and expensive to implement. Addressing an unanswered question concerning impact will be an easy criterion to fulfill in the case of educational uses of technology. Much harder to meet in some cases will be the criterion that the “experimental” program be distinct from practice as usual and that the practice be maintained over time. If the innovation under study is a desirable under circumstances where the nature of the innovation is well understood and the experiment’s implementation is feasible, there were concerns about feasibility.
circumscribed curriculum unit supported by a particular piece of software, such a distinction may not be hard to enforce. (For example, the science of water quality can be learned using *Model-It* simulations or from a chapter in a conventional text.) If, on the other hand, the innovation is broad-ranging in scope and long term in duration, something on the order of process writing supported by word processors or the use of Internet resources to support learning and research skills, these conditions will be more difficult to satisfy. First, the open-ended nature of the technology will make it less likely that teachers will really be doing something distinctive from conventional practice. Descriptive studies of the use of technology tools, such as word processing and spreadsheet software, suggest that teachers initially tend to incorporate the technology into their existing pedagogical practices and only over time evolve new, more student-centered practices (Sandholtz, Ringstaff, & Dwyer, 1996). Second, over time, it will be difficult to keep students, classes, or schools assigned to the control condition from having access to and making use of the same technology resources, both in and outside of school. Although technology is not an *entitlement* in a legal sense, members of the public and educational administrators increasingly think of it as an entitlement in an ethical sense. Given the fact that more affluent students already have access to technology resources in their homes, many argue that students from less wealthy backgrounds are entitled to have these resources available within their schools and public libraries. It would be difficult indeed for principals or superintendents to commit to an experiment that might deny their students access to technology resources for any extended period of time. Thus, we conclude that studies on relatively small units of instruction (such as the civil rights unit described in *Exhibit 4*), specific pieces of software, or new technologies, regarded as less basic (e.g., handheld computing devices), will be more readily examined in experimental designs.

Further discussion of the place for random-assignment experiments in education research occurred at a July 2000 open session of the National Academies’ Board on Testing and Assessment. Robert Boruch gave a presentation to the board in which he pointed out that national random assignment experiments on the effects of interventions, of the sort done in health, juvenile justice, and employment and training fields, cost on the order of $10-12 million if individual students are assigned to treatments at random.
and $20–25 million if classes, schools, or districts are assigned at random. Laurie Bassi, an economist formerly at the Department of Labor (DOL), noted that in DOL’s experience, random-assignment experiments often consume all available research resources and take so long to run that the public policy questions they have been designed to address get acted upon prior to the availability of the research results. Bassi noted also that the fidelity of implementation of an intervention over time has been a serious problem and that differential attrition from either the experimental or the control group can introduce bias into experimental results. (Statistical techniques can be introduced to counteract such bias, but in this case the researcher is relying on the same kinds of corrections used in quasi-experiments.)

Richard Shavelson of Stanford University argued that the pendulum in educational research methods need to swing not to the extreme of doing only random-assignment experiments but to a middle position of asking whether an experiment is appropriate and feasible before moving to other approaches. Shavelson suggested that experiments are more likely to be feasible in the case of small studies of shorter-term, more discrete innovations. Shavelson’s argument echoes our own suggestion that random-assignment experiments will be more feasible in research on particular pieces of software and new devices than when answers are sought to more macro questions about core technology infrastructures or technology-supported whole-school reforms.

In summary, we conclude that experiments with random assignment are an underutilized design in educational research. In combination with other designs, random-assignment experiments would add information about cause-effect relationships in educational technology. This design, by itself, provides little information about the conditions of applicability that support any given technology innovation or intervention, however. Implementation and context data are needed to increase the interpretability of the experimental outcome data.

---

3 Bassi’s experience-based concerns are not new ones; Cronbach (1982) raised similar concerns nearly two decades ago. As Cook points out in his chapter, careful monitoring of an experiment’s implementation will reveal the extent to which differential attrition and treatment contamination or degradation are occurring.
Research on Naturally Occurring Practices

In many cases the question researchers are asked to address does not concern a specific project or innovation but rather a broad range of practices found in various schools to a larger or smaller degree. Here we have in mind questions such as “Does putting Internet-connected computers into instructional classrooms make a difference?” or “Do students who use graphing calculators learn more in high school mathematics?” Because the practices or resources that are the focus of study are arising from within disparate parts of the education system and not out of a particular innovation with a particular theory of change, they will not meet the criteria for an innovation distinct from standard practice that can maintain its distinctive character over time. Thus, these questions are difficult to address with random-assignment experiments.

Many studies of naturally occurring practices have a strictly descriptive purpose—that is, they seek to describe the frequency of various technology uses rather than the effects of those uses. The statistics on Internet connections and technology use gathered by the National Center for Education Statistics and Becker and Anderson’s 1998 Teaching, Learning, and Computing Survey, would fall into this category. Other studies go beyond reporting technology access and usage frequencies per se to correlate degree of access or use with student outcomes. Exhibit 6 describes such a study. Such correlations often feed into arguments about the changes caused by technology, an interpretation that is hazardous, given the many other factors that might account for observed relationships.

Several of the papers in this volume offer designs that can be applied to studying naturally occurring practices. The designs share the features of:

- looking at student performance longitudinally rather than at a single point in time,
- careful delineation and measurement of variables that may be alternative causes of the outcomes to be measured, and
- the use of analytic techniques that permit an estimation of effects at different levels of the education system (e.g., classroom, school, and district effects).
Considerations for a National Research Agenda

In this section of the chapter, we first describe existing federal funding for research on the effects of educational technology and then start sketching out recommendations for a greatly expanded research program. We discuss organizational considerations surrounding such a research effort and conclude with a description of the major types of research we believe are needed.

Existing Federal Support for Learning Technology Research

While we advocate the initiation of a major, coordinated federal investment in research on the effects of educational technology, it is important to recognize that the federal government already supports a number of relevant research programs. We describe the programs here to remind the reader of efforts on which new research can build and to provide a point of comparison for the types of new research programs we propose below.

U.S. Department of Education Funding for Technology Research

The Department of Education funds a variety of research efforts related to the issue of the effects of learning technology. Most of this funding comes in the form of support for evaluation activities conducted as part of implementation programs involving technology. Thus, these efforts are project-linked studies on early-stage innovations, to use the terminology we introduced above. The Technology Innovation Challenge Grants, for example, fund LEAs to partner with universities, businesses, and research organizations to develop and demonstrate “creative new ways to use technology for learning.” Since its inception in 1995, this program has funded 96 projects at a combined level of over $500 million. Although the funding primarily supports design and implementation activities, grantees are required to spend 10% of their funds on evaluation activities, thus generating something in the neighborhood of $50 million for TICG-related evaluation studies over the last six years. Similarly, Star Schools distance learning project grantees and teacher preparation programs receiving grants for Preparing Tomorrow’s Teachers to Use Technology (PTTT), are required to include evaluation as one of their project components.
These project evaluations are an important source of information for program refinement and in some cases have highlighted programs that appear to be particularly effective. It has, however, proved difficult to integrate evaluation findings across projects to provide any coherent set of empirically derived “lessons learned” with respect to the effects of technology on student learning. The various programs have different goals, use different outcome measures, and in many cases, do not incorporate strong evaluation designs with measures of student learning.

Another way in which research on the effects of learning technology might be funded is through the Office of Educational Research and Improvement (OERI) field-initiated research program. A review of the abstracts for 1999 grantees found that in practice relatively little research involving technology is funded through this mechanism: Only 1 of 20 funded field-initiated projects involves the use of technology in the approach under investigation. Moreover, with a funding pattern of just $5-10 million per year (and 0-funding in some years), this program in its present form could not support a large-scale investigation of technology effects on the order of the integrated efforts proposed by the methodology experts featured in this report.

Recently, the Planning and Evaluation Studies office within the Department of Education has supported several efforts in the educational technology arena not tied to individual projects. These include the High-Intensity Technology Study (HITS) being planned by Becker and Lovitts (this paper). This project is designing a three-year evaluation of technology’s impact on student outcomes in classrooms with a high level of technology use. While HITS is large in scope and attempts to examine the effects of technology use more broadly (rather than the impacts of a single program), it is still a single study and cannot be expected to serve all the purposes of a coordinated program or portfolio of research. Another Department of Education-funded project, Evaluation of Educational Policy and Practice, will synthesize the evidence of impact on student outcomes amassed by projects receiving support from the Technology Literacy Challenge Fund. As a formula grant program, the Literacy Challenge Fund gives every state money to help schools integrate technology by supporting improved applications of technology and teacher training and preparation. This synthesis will be dependent on the availability and quality of outcome data collected by individual grantees. Because individual grants
are limited in size and focused primarily on program design and implementation, both the quality of data and the comparability of data across widely disparate programs are at issue.

**National Science Foundation Research Support**

The Education and Human Resources Directorate within the National Science Foundation has a track record of several decades of funding for the development of new curricular and instructional approaches supported by technology in the areas of mathematics and science education. Most recently, this tradition is carried on through NSF’s new Research on Learning and Education (ROLE) program, which issued its first call for proposals in November 1999. ROLE’s stated aim is to enable the integration of research on learning into broader educational and social contexts. The ROLE research program is organized around the context and grain-size for the research (e.g., brain research, fundamental learning research, research in formal and informal educational settings, learning in complex educational systems) and does not include a separate category or priority for technology-related research. However, the solicitation does cite technology as a cross-cutting theme. Technology-related proposals are encouraged: “In order to improve quality, accessibility and efficiency of SMET [science, mathematics, engineering and technology] education, ROLE promotes the use of new and evolving information technologies.” Given the NSF tradition for funding technology-supported innovations, ROLE has attracted many proposals involving the use of technology. Like NSF-funded projects in the past, however, to the extent that ROLE projects incorporate research on the effects of learning technology on student learning, this research is likely to be what Becker and Lovitts call “research-oriented project-based evaluations.” While helpful in informing the R&D efforts, these projects are unlikely to provide evidence of a type or scale that policymakers demand. In these grants, the research on effectiveness is just one component, and often a fairly small proportion, of the entire project. The ROLE solicitation explicitly discourages proposals whose primary emphasis is evaluating the effectiveness of a given innovation: “ROLE is not an evaluation program; rather, it discourages submissions of proposals whose primary purpose is to conduct evaluations of other projects, including activities that EHR Divisions support.” (p. 10). With an
FY2000 annual budget of $8 million, ROLE supports research on new educational approaches supported by technology but is not designed to address questions concerning the effectiveness of educational technology more generally.

**Interagency Education Research Initiative (IERI)**

In 1999, the U.S. Department of Education, National Science Foundation, and National Institute of Child Health and Human Development initiated a joint research program focusing on reading, mathematics, and science with an emphasis on projects that integrate technology. The IERI program announcement is explicit in targeting projects with an articulated theoretical foundation and causal model as well as preliminary evidence of effectiveness. Moreover, proposals are required to provide plans for scaling implementation and research to a level where “questions regarding implementation and fidelity, effectiveness, individual differences… and environmental and policy factors” can be addressed. Thus, this research program seeks to fund research on the effectiveness of what we have called “mature” projects. The program solicitation explicitly encourages (but does not require) experimental designs involving random assignment. The IERI program is quite consistent with the themes stressed by the methodology experts in this volume. Compared to the PCAST report’s call for $1.5 billion annually in research on teaching and learning with technology, however, the IERI funding levels are modest indeed. Some $30 million was awarded under this program in 1999 and $38 million in 2000.

**Considerations for Organizational Structure**

Technology can potentially support any educational function, content area, or grade level. Thus, technology is what Scriven has called a “transdiscipline,” (Scriven, 1991). We could easily take the foci for the various Office of Educational Research and Improvement (OERI) institutes and create a research program entitled “Technology and . . .” for each of them (e.g., Technology and Student Achievement, Curriculum, and Assessment; Technology and Postsecondary Education, Libraries, and Lifelong Learning). And in fact, when the institutes were set up, technology was considered a “cross-cutting theme.” Ideally, the study of technology supports would be integrated
with research on critical questions in every area of teaching and learning. Often this has not happened in practice, however. The relatively small emphasis on technology in many subject area content standards, discussions of teacher preparation, and education reform initiatives outside those explicitly labeled as “technology” initiatives suggests that the question of the organizational “home” for research on teaching and learning with technology is not a trivial one. ⁴

The potential pitfall in setting up a separate technology research program (or for that matter, a separate technology curriculum or assessment) is the risk that technology will become a separate track, poorly integrated with core educational endeavors. Those with strong technology backgrounds are likely to be attracted to the research program, but there is danger of begetting an engineering emphasis rather than an interplay between technology and core teaching and learning issues. On the other hand, when educational technology research is made a part of a research program defined on the basis of a subject area (e.g., early reading or history) or target population (English language learners), opportunities for integration increase but technology may get token treatment or ignored completely. Researchers interested in technology’s contribution to the area may be discouraged from working with the program or may find it difficult to win support for their ideas. Peer review panels set up by such programs often lack individuals with a technology background, meaning that panelists are either uninterested in technology or unaware of what has already been done. In the latter case, panelists have a hard time distinguishing technology-based proposals that are both feasible and potentially ground-breaking from those that are technically unrealistic or mere rehashes of relatively common practice.

Some version of a “partnership” model, with a specifically designated program of research on learning technologies but requirements for coordination with the overall educational research and reform agendas, appears the most promising strategy overall. In our recommendations below, we envision some of the components of the research program being integrated with existing educational research units and some existing as

---

⁴ Our review of the Department of Education’s Catalog of School Reform Programs, for example, found that technology was a significant feature in less than a third of the 33 whole school reform models. Technology receives even less consideration as a force for school improvement in the widely influential document Turning Around Low-Performing Schools: A Guide for State and Local Leaders.
identifiable technology and education initiatives with their own visibility and support. Care will have to be taken to make sure that the technology research agenda is well coordinated with what we have called the “mainstream” research in each of the areas targeted for federal investments in education research.

**Considerations for Degree of Direction**

Another issue that needs to be considered in planning a major program of research is the extent to which the focus and methods of that research arise out of federal planning efforts versus coming from the field. Policymakers have to make tradeoffs between the desire to have certain kinds of research done and the desire to be open to good ideas arising from the individual investigators in the research community. Some of the current federally funded research on educational technology is performed under contract, with the government stipulating the nature and scale of the data it wants collected. In the past, most of this work has been the collection of survey data on technology access and frequency of use, or compilations of previously collected information. Other federal research programs have employed the opposite strategy, supporting field-initiated research, that is those research proposals coming from outside the government that receive the highest ratings from panels of reviewers. In education, most field-initiated research programs have not entertained proposals of a size commensurate with the research strategies recommended by our paper authors.

A major program of research on learning technology including all of the components we describe below would probably employ a wide range of contractual arrangements. A vehicle often used by the Department of Education that has not been used in the field of research on learning technology effects is the funding of a lab or center with this mission. (Centers focused on educational technology *implementation* have been funded.) Center proposals respond to federal agency statements of need for research in a priority area, but leave the proposing organizations substantial room for setting the particulars of their own research programs.

In the case of research on the effects of technology-supported educational innovations, the Department of Education may want to look to practices of the National Institutes for Health (NIH). The NIH uses two primary strategies for harnessing the ideas and energies
of multiple research organizations to an over-arching research program with common measures and shared data sets. Under cooperative agreements, the NIH sets up what we have called an “intermediary organization” within one of its own institutes. NIH researchers stipulate measures and data collection protocols and maintain a central data repository at NIH. This approach requires the availability of a set of practicing research scientists within the government agency. Alternatively, for major health studies (in the $50 million range), an NIH institute typically releases a separate announcement for a coordinating center (housed outside the government) that will serve this function for multiple research and data collection organizations, also working under contract. The coordinating centers typically have the research qualifications to be a data collection center themselves (and sometimes the same organization will win both types of contract). The coordinating center develops instruments, writes data collection protocols, serves as a data repository, runs core data analyses, and makes the data available to the other investigators for their analyses. The coordinating center supports the latter activity by making sure that analysts using the data set define variables in the same way, so that seemingly contradictory results are not caused by differences in variable labeling or definition.

Proposed Five-Part Technology Research Agenda

In the remainder of this chapter, we will make a case for a five-part federal educational technology research agenda, designed to address the larger research questions that have not been answered by individual project-linked research or evaluation studies. We propose five distinctive but inter-related research and development missions:

- Information System for Educational Context Measures
- 21st Century Skills, Indicators, and Assessments
- Research on Technology Use in Schools
- Research on Teaching & Learning with Technology
- Research on Technology and Teacher Professional Development
Information System for Educational Context Measures

Many of the authors in this volume call for carefully documenting the context within which technology-supported teaching and learning occur and for using the same measures in sets of coordinated, linked, or embedded studies. Examples of important contextual variables include teacher characteristics, teacher pedagogical beliefs, professional development supports, school leadership, community engagement, technology infrastructure, and the accountability system in place. The importance of these factors in influencing educational outcomes is not limited to interventions involving technology, of course. The compilation of a set of standard definitions and instruments for measuring such contextual variables would be a major support for educational research generally. To gain acceptance, the core set of contextual variables and associated definitions and instruments would have to be developed through an iterative consensus process. Educational research associations, education leadership and policy organizations, and agencies sponsoring teaching and learning research (not just research involving technology) should all be involved. To get the broadest possible benefit, this work should be carried out from an organizational home that spans the gamut of educational research (perhaps the National Center for Education Statistics). Definitions, rubrics, and instruments could be made available through the World Wide Web (the OERL site at http:oerl.sri.com provides an example of the kind of easy-to-navigate interface that would be needed).

Initiative for 21st Century Skills, Indicators, and Assessments

Many studies of the effects of technology-supported innovations are hindered by a lack of measures of student learning commensurate with the initiatives’ goals. The kinds of mathematical problem finding and planning skills that are among the key objectives for the Adventures of Jasper Woodbury (CTGV, 1997), for example, get little or no coverage in widely available standardized tests. High-stakes testing programs that emphasize basic skills and factual knowledge concerning a broad range of topics (as opposed to deeper conceptual knowledge in a narrower range of fields) serve as disincentives for the use of innovative technology-supported programs that stress deep understanding of a few topics and advanced problem solving and communication skills.

The development and field testing of assessment instruments that are valid, reliable, and sensitive to instruction is a complex, time-consuming effort, and one that is not easily
mastered by organizations with little experience in this area. It is unrealistic to expect individual technology projects that support types of learning not well represented in off-the-shelf tests to develop their own measures in a context of limited funding. Moreover, neither private companies nor individual states are likely to amass the resources and expertise necessary to design, develop, field test, and disseminate such assessments for all the grade levels and skill and knowledge areas we hope to address with technology. For this reason, we see the development of high-quality assessments of the kinds of skills called for by these standards-setting bodies as appropriate for federal R&D support.

While the number of different content areas in which we might want to have better student assessments is virtually unlimited, there is a great deal of agreement around the need for certain generalizable information seeking, analysis, and communication skills. These skills would provide a good focus for the first stage of this assessment development initiative. These are skills that employers say they need in 21st century workers and that are stressed by standards-setting organizations both in content areas and in the area of technology itself. The standards of the National Council of Teachers of Mathematics (NCTM, 1989), for example, start with four process skills important at every grade level: problem solving, communication, reasoning, and connections (linking different subfields of mathematics and linking mathematics to other disciplines and real-world problems). The Benchmarks for Science Literacy (AAAS, 1993) include critical response skills (being able to judge the quality of claims based on the use or misuse of supporting evidence, language used, and logic of the argument) and communication skills as essential components of science literacy. The National Science Education Standards (NRC, 1996) includes eight standards that focus on science inquiry. The abilities and skills that underlie these standards include: (1) identifying research questions, (2) designing and conducting scientific investigations, (3) using appropriate tools and techniques, (4) developing descriptions, explanations, predictions, and models; (5) thinking critically and logically to relate evidence and explanations; (6) recognize and analyze alternative explanations and predictions; (7) communicate scientific procedures and explanations; and (8) use math in all aspects of science inquiry. The skills promoted by these disciplinary groups overlap with three of the five technology skills the International Society for Technology in Education (ISTE, 1998) says are required to
become “capable information technology users”: Namely, (1) information seeking, analysis, and evaluation, (2) problem solving and decision making, and (3) communicating, collaborating, publishing, and producing. Similarly, the National Research Council report *Being Fluent with Information Technology* (1999) stresses intellectual capabilities such as organizing and navigating information structures and evaluating information, collaborating, and communicating to other audiences. Figure 1 illustrates the overlap among these various standards. Despite the value placed on these skills, we lack widely available, high-quality assessments for gauging students attainment of them. The Secretary’s Commission on Achieving Necessary Skills and the New Standards Project have both done useful work that provides a foundation for further efforts, however. Prospects for improved and more widely available assessments could be enhanced by capitalizing on technology as a means of delivering, scoring, managing, and storing assessments (see Mislevy et al.’s paper).

The proposed national initiative would establish assessments that could be used in evaluation and research on technology-supported learning and education more generally. Many of the assessments should themselves be technology-supported. To be useful and credible, the assessments must have demonstrated technical quality and endorsement from an unbiased, prestigious organization (such as the National Research Council).

**Research on Technology Use in Schools**

This research program would examines the frequencies and correlates of common and emerging “naturally occurring” practices. These are practices that are being widely implemented in U.S. schools without special funding from programs such as the Technology Innovation Challenge Grants and without support from research or commercial organizations developing the technology. Research under this program would attempt to answer questions about how technology is being used in schools serving different populations and in different subject areas and grade levels and would relate these uses to observed outcomes. This component of the research agenda thus would
include major longitudinal studies of technology use in schools (as proposed by Rumberger and Means et al.), possibly in connection with an existing national longitudinal study.

More specific topics that might become areas of research within this program include Internet research, use of discrete educational software (including integrated learning systems) with Title I students, technology use for English language learners, and computer-based writing instruction.

**Research on Teaching and Learning with Technology**

Another important line of research would examine the student learning effects of well-defined projects or innovations involving technology. Given the size, breadth, and complexity of the research agenda, we propose setting up a responsible organization, a research institute or center sponsoring and overseeing multiple studies on the effectiveness of technology in addressing topics established as priorities for federal education research generally (e.g., early reading and middle school mathematics).

In addition to contextualized evaluations illustrating the interplay between research and practice described as Pasteur’s Quadrant (Stokes, 1997), the proposed Institute for Research on Teaching and Learning with Technology would support random-assignment and quasi-experiments to investigate the effects of mature innovations. Because it is important also to be looking to the future and to provide a research base that can influence both commercial and noncommercial technology innovators, this institute would also sponsor proof-of-concept studies exploring the value of new technologies (e.g., wireless Internet devices and services). A final thread would be studies linked to theories about how best to support learning in specific subject areas, such as literacy, middle school mathematics, and social studies.

Thus, we imagine this research institute sponsoring research employing a range of methodologies. Quality criteria and conditions of applicability for each methodology should be elucidated, as a guide both to individuals responding to program solicitations and to the proposal review process. Wherever appropriate, the research sponsored by this institute would incorporate the common context measures and the assessments developed
under the first two components of the research agenda described above. Aggregation of findings across studies could be further supported through clustering studies of innovations with similar learning goals and the efforts of a (nongovernmental) intermediary organization, as suggested above. This work could also be supported by a network of “sentinel schools” or testbeds, as suggested by several of the authors in this volume. The Institute for Research on Teaching and Learning with Technology would be the appropriate sponsoring agency for this network. These schools would become a testbed for coordinated studies of new approaches and innovations.

**Research on Professional Development for Instructional Uses of Technology**

The final component in our five-part agenda would focus on identifying effective approaches to providing training and continuous support for teachers’ integration of technology with instruction. Both pre-service and in-service education and support, and both technology-based and off-line forms of training and support would fall within the purview of this research program.

This research should be conducted with an eye toward informing policy discussions around state and district accountability systems which are providing rewards and sanctions related to the integration of technology and teachers’ demonstrated technology proficiency. An important research question given different state strategies for increasing teachers’ ability to use technology within classrooms (e.g., requiring a technology course as part of teacher preparation as opposed to requiring teachers to pass a technology proficiency test in order to obtain a credential) is the effect of any such system on the teaching and learning that occurs within those teachers’ classrooms. This same research program could encourage integration of graduate schools of education and local K-12 school systems through professional development programs that integrate research and practice with teacher learning.

**Conclusion**

In 1997 the Panel on Educational Technology of the President’s Committee of Advisors on Science and Technology (PCAST) issued its report asserting that “a large-scale program of rigorous, systematic research on education in general and educational
technology in particular will ultimately prove necessary to ensure both the efficiency and
cost-effectiveness of technology use within our nation’s schools.” The PCAST Panel argued that the investment in research in this area should be comparable in scope to that in pharmaceutical research—specifically calling for an annual investment of $1.5 billion. Given the fact that the current funding level for research on the learning impacts of technology-supported innovations (as described above) is closer to $50 million, any approximation to the PCAST recommendation would require a major change in the way the federal government thinks about and sponsors educational technology research. This synthesis is intended as a next step in conceptualizing the research needs, promising new approaches, and innovative research sponsorship arrangements to respond to that challenge.

References


Table 1
Commissioned Research Design Papers

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Institution(s)</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eva L. Baker and Joan L. Herman, CRESST</td>
<td></td>
<td>New Models of Technology Sensitive Evaluation: Giving Up Old Program Evaluation Ideas</td>
</tr>
<tr>
<td>Henry Jay Becker, University of California, Irvine and Barbara E. Lovitts, American Institutes for Research</td>
<td></td>
<td>A Project-Based Assessment Model for Judging the Effects of Technology Use in Comparison Group Studies</td>
</tr>
<tr>
<td>Thomas D. Cook, Northwestern University</td>
<td></td>
<td>Reappraising the Arguments against Randomized Experiments in Education: An Analysis of the Culture of Evaluation in American Schools of Education</td>
</tr>
<tr>
<td>Katie McMillan Culp, Margaret Honey, and Robert Spielvogel, Education Development Center/Center for Children and Technology</td>
<td></td>
<td>Local Relevance and Generalizability: Linking Evaluation to School Improvement</td>
</tr>
<tr>
<td>Larry V. Hedges, Spyros Konstantopoulos, and Amy Thoreson, University of Chicago</td>
<td></td>
<td>Designing Studies to Measure the Implementation and Impact of Technology in American Schools</td>
</tr>
<tr>
<td>Alan Lesgold, LRDC, University of Pittsburgh</td>
<td></td>
<td>Determining the Effects of Technology in Complex School Environments</td>
</tr>
<tr>
<td>Barbara Means, Mary Wagner, Geneva D. Haertel, and Harold Javitz, SRI International</td>
<td></td>
<td>Investigating the Cumulative Impacts of Educational Technology</td>
</tr>
<tr>
<td>Lincoln E. Moses, Stanford University</td>
<td></td>
<td>A Larger Role for Randomized Experiments in Educational Policy Research</td>
</tr>
<tr>
<td>Russell W. Rumberger, University of California, Santa Barbara</td>
<td></td>
<td>A Multi-level, Longitudinal Approach to Evaluating the Effectiveness of Educational Technology</td>
</tr>
<tr>
<td>Argument</td>
<td>Rebuttal</td>
<td></td>
</tr>
<tr>
<td>---------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Causation is more than the small subset of potential causes that can be tested in a randomized experiment; often only a single cause is tested.</td>
<td>Some causal contingencies, however, are of minor relevance to educational policy, even if they are useful for full explanation. The most important contingencies are those that, within normal ranges, change the sign of a cause relationship and not just its magnitude. Such causal changes indicate where a treatment is directly harmful as compared to having more of less benefit for one groups students than the other groups.</td>
<td></td>
</tr>
<tr>
<td>Random assignment was tried in education and has failed. Prior experiments experienced difficulties in how the random assignment was implemented and the degree of correspondence between the sampling particulars and likely conditions of application as new policy.</td>
<td>Experiments can overcome some of the past difficulties, by checking on how well the initial randomization process was carried out and whether treatment independence has been achieved and maintained.</td>
<td></td>
</tr>
<tr>
<td>Random assignment is not feasible in education.</td>
<td>In implementing randomization, the role of political will and disciplinary culture are critically important. Compared to research conducted in other fields, educational research accords little privilege to random assignment.</td>
<td></td>
</tr>
<tr>
<td>Random assignment is not the method of choice for studying many educational innovations, because the reform theories are under-specified, schools are chaotic, treatment implementation is very variable, and treatments are not theory-faithful.</td>
<td>For policy purposes, we have to assess what an innovation can do despite variation in treatment exposure within the comparison groups. Standard implementation will not be expected in the hurly-burly of real educational practice. With random assignment we can assess both the effects of treatments that are variably implemented and the more theory-relevant effects of spontaneous variation in the amount and type of exposure to program details.</td>
<td></td>
</tr>
<tr>
<td>Random assignment entails trade-offs not worth making. Often experiments reveal little that can explain the processes whereby effects are produced or provide guidance for effective implementation.</td>
<td>Although an experiment focuses on answering a causal experiment, that does not preclude examining reasons for variation in implementation quality or seeking to identify the processes through which a treatment influences an effect. The data analysis does not have to be restricted to the intent-to-treat group. Ethnographic data can be collected on treatment groups, in order to identify unintended outcomes and mediating processes.</td>
<td></td>
</tr>
<tr>
<td>Experiments assume an invalid model of rational decision-making on the part of policymakers.</td>
<td>Whether it’s experiments or surveys or cases studies, research utilization is multiply determined by politics, personalities, windows of opportunity, and values.</td>
<td></td>
</tr>
</tbody>
</table>
Table 3  
Contrasts Between Innovative Technology-Supported Assessments and Traditional Tests

<table>
<thead>
<tr>
<th>Assessment Features</th>
<th>Traditional Standardized Achievement Tests</th>
<th>Innovative Technology-Supported Assessments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administration</td>
<td>• Individual learners • No collaboration • One common setting • Standardized conditions and procedures</td>
<td>• Individual learners or small groups • Opportunities to demonstrate social competencies and collaboration • Multiple, distributed settings • Documented but flexible procedures</td>
</tr>
<tr>
<td>Item/Task Content</td>
<td>• Typically measures knowledge and facts • Rarely measure inquiry and communication, other than brief writing samples and simple calculations on small data sets</td>
<td>• Measure all aspects of inquiry • Linked to content, inquiry, and performance standards •</td>
</tr>
<tr>
<td>Item/Task Presentation, Format, and Scaffolding</td>
<td>• Discrete, brief problems • Decontextualized content • Mostly multiple-choice/&quot;fill-in-the bubble&quot; format • Limited number of constructed response items • Usually no external resources can be used in problem-solving</td>
<td>• Extended, performance tasks, including hands-on tasks with use of simulations, probeware, Web searches, visualizations, and multiple representations • Option for access to other resources, including software, the Internet, and remote experts</td>
</tr>
<tr>
<td>Scoring and Analysis</td>
<td>• Number and percent correct; percentiles; NCEs • Competency-based categorical ratings sometimes used</td>
<td>• Qualitative and quantitative data • Use of scoring rubrics that characterize specific attributes of performance; • Potential for automated scoring of natural-language responses (e.g., essays) and complex problem solving (e.g., diagnostic tasks)</td>
</tr>
<tr>
<td>Recording and Archiving of Responses</td>
<td>• Paper-pencil • Optical scan</td>
<td>• Mechanisms to reveal steps of problem-solving (e.g., Internet trace strategies, electronic notebooks for annotations and describing rationale and documentation of steps) • Web pages • Screen shots • May accumulate responses over time</td>
</tr>
</tbody>
</table>

Source: Adapted from Quellmalz and Haertel (submitted for publication)
Figure 1. 21st Century Skills in Relation to Selected Content and Technology Standards

21st Century Skills
- Seek & organize information
- Evaluate information
- Communicate
- Collaborate

NCTM Standards
- Problem solve
- Communicate
- Reason
- Connect

NSES Standards
- Identify questions
- Design/conduct investigations
- Develop descriptions, explanations, predictions
- Analyze explanations
- Communicate

AAAS Benchmarks
- Judge quality of claims based on supporting evidence, language used, & logic of argument
- Communicate

NCR Tech Fluency
- Organize & navigate information structures
- Evaluate information
- Collaborate
- Communicate

ISTE NETS
- Seek, analyze & evaluate information
- Solve problems
- Communicate, collaborate, publish, & produce
Exhibit 1
A Technology-based Assessment

Edys Quellmalz and her colleagues at SRI have designed and developed a Web-based assessment of Internet research skills. This assessment was designed to capture students’ ability to locate, navigate through, and organize information as well as their ability to evaluate that information and communicate their conclusions to other audiences. The recent National Research Council report *Being Fluent with Technology* (1999) argues that these intellectual skills are as essential to technology fluency as the more commonly measured skills in using contemporary software.

SRI’s on-line assessment presents a problem or challenge which can be responded to by individuals or pairs of students. The assessment task involves assisting a group of foreign exchange students who are planning a summer trip to the U.S. by helping them to pick one of several U.S. cities as the place to spend their summer. Given the city features of most concern for the foreign students (e.g., summer recreational opportunities), students taking the assessment task pour through complex sets of real Web resources to identify information on which to base a decision. In addition to the provided URLs, each student is required to formulate a search query to collect additional information. Students are also asked to identify information of dubious validity in the Web materials and to explain why they question the accuracy of that information or statement.

When the individual or pair of students taking the assessment determines that enough information has been collected to make a selection, they choose the city to recommend and compose a justification for their choice, which they enter into a text box in the assessment’s Web interface.

Finally, the students taking the assessment compose a letter to the foreign exchange students to inform them of the recommendation and the facts supporting their choice.

Each group’s work is scored using rubrics for the three areas of information search, reasoning with information, and communication. Scores for collaboration skills and for fluency using Web browser’s and word processors are also assigned by trained raters.
In 1991, Bell Atlantic-New Jersey began planning an initiative with the Union City, NJ, Board of Education to test the technical feasibility and educational benefits of offering multimedia on demand at school and at home. The Center for Children and Technology (CCT) was asked to join the partnership to help plan, support, and evaluate the initiative.

The technology trial, Union City Online, was launched under circumstances that posed many challenges yet offered broad opportunities. Just two years earlier, in 1989, the school district had failed 44 of 52 indicators used by the state of New Jersey to measure school system efficacy. The district would have to undergo state takeover if a demonstrably successful restructuring wasn’t implemented within five years. After an initial planning year, the district had begun implementing broad reforms, starting with grades K-3 and adding additional grades each year. Elements of the reform included an intensive focus on literacy, a whole-language approach to learning, elimination of pull-out programs, expansion of the annual number of teacher in-service hours from 8 to 40, and block scheduling. In addition to the reform efforts, the district benefited from a statewide reform of educational spending formulas which drastically increased Union City’s funding, making it possible to refurbish the district’s aging schools.

When the technology trial began its implementation phase in September 1993, the school reform effort and new curriculum were just starting pilot implementation in the middle school grades. The technology project was initiated in a newly re-opened building, the Christopher Columbus Middle School. Thus the technology infrastructure, organization of grades seven and eight, and curriculum were all changing simultaneously.

CCT researchers documented the context in which the technology was being used. They studied teachers’ practices and parents’ involvement. Details of whether teachers used inquiry-based curricula, the amount of professional development they received, quality of leadership at the building level, and the level of expectations that teachers held for students were recorded. In their observations, CCT researchers looked for the impact that the computer and networking technologies were having on students’ learning, teachers’ teaching, and parent involvement. Teacher interviews documented their perceptions that the technology increased students’ interest in writing projects, enhanced their writing abilities, and increased communication among teachers and between teachers and parents.

Quantitative indices of education quality were examined also. Seventh graders at Christopher Columbus performed better than other district seventh graders on state achievement tests; Christopher Columbus eighth graders were the only ones in the district to meet state standards for performance on reading, math, and writing tests and were more likely than their peers at other Union City schools to qualify for ninth-grade honors classes. Christopher Columbus also had the best attendance rate in the district for both teachers and students.

The evaluation design included in-depth analyses of a group of middle school students who started as seventh or eighth graders at Christopher Columbus and had sustained access to the networking technologies at home and school as well as a group of students who had access to the technologies at school only. Students with both school and home access to technology performed better than other district students at the same grade level in writing and mathematics during the first year of the project; in subsequent years, they continued to do better on the writing portion of state tests. The evaluators report that the technology facilitated increased communication among teachers, students and parents; additional opportunities to write and edit; and increased opportunities to participate in group multimedia authoring projects. Contextual factors contributing to the students’ higher test scores included the enthusiasm and dedication of the Christopher Columbus staff; high expectations set for students in the technology trial; and district programs to involve parents more directly in their children’s education (Chang et al., 1998).

In this contextualized evaluation, the district, school, classroom, and home settings were well documented. Outcome measures included those indices that made a difference in the political climate of Union City (e.g., state “early warning tests” that could lead to reconstitution). The identification of technology’s contributions was possible only through finer-grained analyses of teachers’, students’, and parents’ activities because so many efforts to improve academic performance were undertaken simultaneously (e.g., curriculum reform, block scheduling, increased funding).
Exhibit 3
A Contextualized Evaluation

In school year 1990-91, the state of West Virginia began statewide implementation of a systematic program to bring computer technology, basic skills software, and teacher training to every public school in the state. Under this Basic Skills/Computer Education (BS/CE) program, every public elementary school received 3-4 computers, a printer, and access to a schoolwide, networked file server for every kindergarten class during the program’s first year. As the cohort of 1990-91 kindergartners moved up in grade each year, the state provided an equivalent technology infrastructure for the grade they were entering. Schools were required to choose software systems from either IBM or Jostens Learning to implement using the new hardware and network access. Teachers in the target grade receiving new equipment and software were given intensive training stressing the relationship between the software offerings and the state’s basic skills standards and how to guide their students through use of the programs.

After eight years of the program, West Virginia knew that standardized test scores for students in the BS/CE program cohorts were higher than those of previous cohorts, but did not know how much of the improvement could be attributed to the technology program. It could be that the nature of the school population was changing over time or that other educational improvement efforts were producing the higher scores. Interactive, Inc. was hired to conduct analyses addressing this question (Mann, Shakeshaft, Becker, & Kottkamp, 1999).

The West Virginia case was unusual in that the intervention was clearly defined (use of basic skills software from one of two vendors) and was implemented in every school statewide. Schools did differ, however, in how intensively they implemented the program—how much time students were given to use the software and how involved individual teachers were in professional development and implementation. Mann and his colleagues designed a study capitalizing on this variation by relating it to the size of student gains on achievement tests. Eighteen schools were selected for study. Mann et al. report that the schools were selected with the help of a state education advisory group on the basis of achievement, perceived intensity of technology implementation, geography, vendor uses, and socioeconomic status. The schools covered the range from low to high standardized test scores and from low to high technology use. All fifth-grade students in the 18 schools were included in the study. Students were surveyed concerning their attitudes towards school and towards technology and their technology experiences each year since kindergarten. Surveys were administered to teachers in grades 3-5 to capture the attitudes and practices of teachers currently working with the fifth-grade cohort as well as those of teachers the students would have had in prior years. Principals, fifth-grade teachers, and some early-grade teachers were interviewed as well.

West Virginia’s introduction of the Stanford Achievement Test Ninth Edition (SAT-9) in school year 1996-97 meant that two successive years of test data were available for the fifth-grade students. Mann et al. computed student gain scores and analyzed them using a three-factor model comprising software and computer availability and use, student and teacher attitudes toward computers, and teacher training and involvement in technology implementation decisions.

Mann et al. found that the more of each factor students experienced, the greater their gains on basic skills from the end of fourth grade to the end of fifth grade. Multiple regression analysis suggested that 11% of students’ gains could be attributed to the model (i.e., technology use to support basic skills). The BS/CE program appeared to have larger effects for children who did not have computers at home, and for students who reported earning C grades rather than As or Bs. There were no differences in gain scores between white and non-white students nor generally between girls and boys.
Exhibit 4
A Quasi-Experiment

In 1996, the Center for Applied Special Technologies (CAST) conducted a quasi-experimental study of the effects of access to on-line resources on students’ content knowledge and presentation skills. A total of 28 classes, equally divided between the fourth- and fifth-grade levels, were drawn from seven urban school districts participating in the study, which was funded by Scholastic, Inc. and the Council of Great City Schools.

The primary contact for each district selected the two schools for study participation and worked with the two principals to select the experimental and control classes. Within each participating school, one class was assigned to the experimental group, which received on-line access to Scholastic Network and the Internet, and a second classroom at the same grade level was assigned to the control group, which did not have Internet access. CAST reports, “District administrators did not randomly assign schools and classes for the study due to logistical constraints” (p. 20).

Both sets of classrooms agreed to implement a unit of study on civil rights, culminating in student research projects. A curriculum framework, activities, worksheets, and an outline for the student projects were distributed to teachers of all participating classrooms. For the student projects, teachers were instructed to divide their class into small groups of three or four students. Each group was to conduct research, analyze information, and prepare a presentation. All classes were encouraged to have students use multimedia reference materials, but only the experimental classes could use on-line resources or communication activities. Teachers in the experimental group received on-line training in how to incorporate Internet resources into the unit. In addition, CAST provided half of the experimental teachers with two sets of in-person, two-day workshops and ongoing support through email and message boards. Participating classes were instructed to implement the unit during January and February and to submit student projects to CAST for scoring by mid-March.

Six classrooms were not included in the final data set on student performance: Four of these classes did not implement the civil rights unit within the study’s time parameters because of conflicting school priorities, and two classes had students do whole-class presentations rather than working in small groups, as instructed. The final analysis included 41 presentations from experimental classrooms and 19 from control classrooms at the fourth-grade level and 25 from experimental classrooms and 19 from control classrooms at the sixth-grade level. An experienced teacher was hired to serve as an “independent” scorer for the student presentations.

Student projects were scored on nine dimensions, using a four-point scale. Among fourth-graders, experimental student groups performed better than control groups on the two dimensions “effectiveness of bringing together different points of view” and “presentation of a full picture.” Sixth-graders in the experimental group performed significantly better on “completeness,” “presentation of a full picture,” “accuracy of information,” and “overall effectiveness of presentation.” None of the 18 T-tests found a significant advantage for control group students.

Within the experimental group, students whose teachers received the extra training and support performed more poorly than other student groups in the experimental condition, a difference CAST attributed to extenuating circumstances such as a teachers strike in one of the districts. An analysis relating the amount of time students within the experimental group were logged onto the Internet to the performance scores found no relationship.
Exhibit 5  
A Formative Evaluation

Classroom Connect, a company developing subscription-based Web educational resources, offers a product line called *Quest* which allows students to use the Internet to follow an expedition exploring a central question or mystery. Quests extend for 4-5 weeks and students follow the progress of, and make suggestions to, a team of scholars and educators travelling by bicycle as they pursue evidence related to questions such as “Did Marco Polo really go to China?” Classroom Connect asked the Center for Technology in Learning at SRI International to evaluate the quality of learning stimulated by the Quests.

The company needed to know how its product was being used in classrooms, and whether any particular kinds of classrooms (for example, those at certain grade levels or with limited technology) were having difficulties using the Web resources as intended.

SRI researchers helped Classroom Connect more clearly define its learning goals for the product in terms of both content knowledge and problem solving skills. Based on the research literature, SRI suggested a hierarchy of increasingly complex student outcomes in each of these areas and then initiated field visits to classrooms conducting Quest activities. Field notes were largely qualitative in nature, but each observation covered the issues of technical configuration of the classroom, student demographics, assigned student activities, teacher facilitation activities, curriculum integration, and observable evidence of the kinds of learning students were experiencing, using the content and problem solving hierarchies.

Observations suggested that different classrooms were using the Quest resources in vastly different ways. Some teachers turned students loose “to explore” while others sent them to find specific pieces of information. Some teachers developed their own off-line activities to help focus their students’ attention on the central question in the Quest and to help them relate evidence to competing hypotheses. In some classrooms the program was well integrated with the curriculum; in others it was viewed as a supplemental “fun” activity unrelated to other student work. Classroom observations suggested that depth of student inquiry was particularly variable, with some students looking for quick ways to get to “the answer” and others surfing for engaging videos. Researchers also found that the program could be effectively implemented with a single computer in the classroom, a configuration which often promoted more effective group inquiry than a separate computer for each student.

Since the main goal of the evaluation activities was to inform product refinement, data and design recommendations were communicated quickly and informally in oral briefings and letter reports. Based on what was learned from the initial evaluation activities, the Classroom Connect development and expedition teams refined their approach in developing the next Quest. This Quest was designed to give more prominence to the central mystery throughout the Quest; provide more extensive modeling of the inquiry process and learning activities that would promote student inquiry; add prompts to encourage students to research their responses in more depth and to support their conjectures with evidence; and offer more tips and tools to support teachers’ curriculum planning. An on-line survey was administered to participating teachers and student inputs to the Quest Web site were analyzed in terms of demonstrated depth of inquiry. The analysis of the Quest content confirmed that the team had indeed made evidence a more prominent part of the most recent Quest. Student inputs posted on the Web site were much more likely to display evidence-based reasoning than were the inputs to the prior Quest. Classroom Connect decision makers reported an increased commitment to using formative evaluation data as part of the product design and development process.
## Exhibit 6

### A Correlational Analysis

The Educational Testing Service conducted an analysis of survey and assessment data from the 1996 National Assessment of Educational Progress (NAEP) in mathematics. Two student samples were part of the analysis: 6,227 fourth graders and 7,146 eighth graders. A four-factor model was tested against the data. Factors in the model were frequency of school computer use for mathematics; access and use of computers at home; professional development for math teachers in use of technology; and higher-order and lower-order uses of computers by math teachers and their students. Computer uses considered “higher order” were “mathematical/learning games” for fourth graders and “simulations and applications” for eighth graders. Use of “drill and practice” software was considered “lower order” use at both grade levels. Outcome variables analyzed were performance on the NAEP mathematics achievement items and school social climate, a variable derived from measures of student tardiness, student absenteeism, teacher absenteeism, teacher morale, and student regard for school property.

After controlling statistically for characteristics of students and schools (i.e., socioeconomic status, class size, and teacher characteristics), the analysis found that amount of school time students spend on computers in total does not predict greater mathematics achievement (in fact there is a small negative effect) but that certain uses of technology are associated with higher achievement, particularly at the eighth-grade level. Eighth graders whose teachers mostly used computers with them for simulations and applications had higher mathematics scores. Eighth graders whose teachers mostly used computers with them for drill and practice programs had lower scores. Among fourth graders, there was a smaller positive association between the use of mathematical/learning games and NAEP math scores. Fourth-grade use of drill and practice appeared to have no effect on scores after controlling for student and school characteristics. At both grade levels, teachers’ receipt of professional development on the use of technology was associated with higher student scores and with a more positive school climate. Teacher use of technology to promote higher-order skills was also associated with more positive school climates.

The published report of this analysis (Wenglinsky, 1999) suggests that use of technology to support higher-order skills at the eighth-grade level raises mathematics achievement. The author acknowledges, however, “There are no prior measures of mathematics achievement, making it difficult to rule out the possibility that positive educational outcomes are conducive to certain aspects of technology use rather than the other way around.” That is, it may be that teachers who perceive their students are doing well in mathematics provide them with experience with simulation and applications programs while those who perceive deficiencies use drill and practice software for remediation.