EVALUATING EDUCATIONAL TECHNOLOGY

EFFECTIVE RESEARCH DESIGNS FOR IMPROVING LEARNING

GENEVA HAERTEL AND BARBARA MEANS

Evaluating Educational Technology

Effective Research Designs for Improving Learning

Evaluating Educational Technology

Effective Research Designs for Improving Learning

Geneva D. Haertel
Barbara Means
Editors

Foreword by Linda G. Roberts

TEACHERS COLLEGE PRESS

Teachers College
Columbia University
New York and London

Published by Teachers College Press, 1234 Amsterdam Avenue, New York, NY 10027

Library of Congress Cataloging-in-Publication Data

Evaluating educational technology : effective research designs for improving learning / Geneva Haertel, Barbara Means, editors.
p. cm.
Includes bibliographical references and index.
ISBN 0-8077-4330-5 (cloth: alk. paper)
1. Educational technology—United States—Evaluation. 2. Educational innovations—United States. I. Haertel, Geneva D. II. Means, Barbara, 1949–

LB1028.3.E962 2003
371.33—dc21 2003044046

ISBN 0-8077-4330-5 (cloth)

Printed on acid-free paper

Manufactured in the United States of America

10 09 08 07 06 05 04 03 8 7 6 5 4 3 2 1

Contents

Foreword vii
Linda G. Roberts

Introduction to Part I:
Evaluating the Effects of Learning Technologies **1**
Barbara Means, Geneva D. Haertel, and Lincoln Moses

1 The Case for Randomized Experiments 15
Thomas D. Cook, Barbara Means, Geneva D. Haertel, and Vera Michalchik

2 Detecting Technology's Effects in Complex School Environments 38
Alan Lesgold

3 Achieving Local Relevance and Broader Influence 75
Katherine McMillan Culp, Margaret Honey, and Robert Spielvogel

4 A Distributed Evaluation Model 95
Eva L. Baker and Joan L. Herman

Introduction to Part II:
Measuring Important Student Learning Outcomes **121**
Barbara Means and Geneva D. Haertel

5 A Project-Based Approach to Assessing Technology 129
Henry Jay Becker and Barbara E. Lovitts

6 Improving Educational Assessment 149
Robert J. Mislevy, Linda S. Steinberg, Russell G. Almond, Geneva D. Haertel, and William R. Penuel

Introduction to Part III: Studying Long-Term Effects Within a Complex, Multilevel System **181**
Barbara Means and Geneva D. Haertel

7 Studies of Technology Implementation and Effects 187
Larry V. Hedges, Spyros Konstantopoulos, and Amy Thoreson

8 The Advantages of Longitudinal Design 205
Russell W. Rumberger

9 Studying the Cumulative Impacts of Educational Technology 230
Barbara Means, Mary Wagner, Geneva D. Haertel, and Harold S. Javitz

Conclusion: Cross-Cutting Themes in Research Designs for Technology 257
Barbara Means and Geneva D. Haertel

About the Contributors 271

Index 279

Foreword

Linda G. Roberts

Dramatic advances in computer and telecommunications technology have occurred over the past decade. These advances have lowered the cost of technology while increasing capabilities for applications that cut across society, including education. If we were to draw a map of the "technology horizon," we would see tremendous forces shaping the future of teaching and learning (Grove Consultants International and the Institute for the Future, 2000), including massive amounts of information accessible via the Internet; smarter appliances and devices that are becoming increasingly interconnected; miles of fiber-optic cables that are bringing high-speed access to more and more schools, homes, and communities; and users who are inventing applications never dreamed of by the designers.

Since 1995, local communities, states, and the federal government have invested heavily in technology for the nation's schools and classrooms. This first national technology plan challenged America's schools to reach four goals: training for teachers, computers for students, classrooms connected to the Internet, and development of effective software and online learning resources (U.S. Department of Education, 1996). Today virtually all our public K–12 schools and almost three-fourths of classrooms connect to the Internet. The student-to-computer ratio has improved, to a national average of five students per computer. In a small number of schools, every student has a computer. Increasing numbers of new teachers are coming to the classroom with preparation to use these 21st-century tools, and they join colleagues who are also gaining skill and confidence in using technology across the curriculum (U.S. Department of Education, Office of Educational Technology, 2001). But what is missing?

The prevailing view that schools are slow to embrace change notwithstanding, much of the investment in technology has been driven by schools' willingness to experiment and innovate. Parents believe that their children should know how to use modern technology and should become technologically literate, and schools have responded to this demand. Increasingly, however, educators and government officials (from Congress to local school boards) believe that it's time for schools to demonstrate the value of the technology infrastructure. They want to see evidence of effectiveness before they invest further in more computers, more teacher training, and expanded telecommunications capacity.

As director of the Office of Educational Technology, I listened as the calls for research intensified. In March 1997, the report of the President's Committee of Advisors on Science and Technology, Panel on Educational Technology (PCAST) (1997) called for a large-scale program of rigorous, systematic research. Many other reports and policymakers echoed these recommendations. In response, the U.S. Department of Education required Technology Innovation Challenge Grant projects to build evaluation into their 5-year programs and partnerships. Similarly, states and districts began to conduct their own studies of educational technology, including some that encompassed a large number of students and classrooms. A number of these studies have helped identify applications and strategies that work. The Interagency Educational Research Initiative was established with the goal of building on prior work, identifying technology-supported innovations that had some evidence of effectiveness and funding investigations of their effectiveness on a broader scale. (This initiative is supported by the National Science Foundation, U.S. Department of Education, and the National Institutes of Health.)

Nevertheless, I was still convinced that we could do better. Most studies of technology's impact on students looked only at short-term effects and were either small in scale (e.g., a few selected classrooms or schools) or lacking in detailed information about just what technology-supported learning experiences students had had. I believed that researchers could improve the design and collection of data. Just as new technology created new opportunities for learning, it created ways to invent new tools for research and evaluation, particularly ways to track and monitor what, how, and when learning occurred. The question for me was, How to move forward in a practical way? A compelling strategy emerged at a high-level seminar on Technology and Education held by the U.S. Department of Education and the Brazil Ministry of Education, hosted at SRI headquarters on December 2–3, 1998. This meeting in Menlo Park, California, was the Second U.S.–Brazil Binational Education Dialogue, an activity of the U.S.–Brazil Partnership for Education launched in Brasilia on October 14, 1997, with support of President Clinton and President Cardozo.

Participants heard again that many studies pointed to the promising impacts of technology but also learned that, in all too many cases, there were more questions than answers. Classroom access to technology was expanding. All across the country, there were districts that could be rich sources of data and schools that could be laboratories for the development of the next generation of interactive learning resources. Certainly, this was the time to invest in research and evaluation: What would it take to conduct a set of rigorous studies? Was there a base of theory to build on? Were the tools for analysis adequate? Was it possible to conduct serious evaluation in classroom settings where change and revision were continuous? Where to begin?

The researchers around the table suggested a compelling strategy for getting started: Invite experts in research and evaluation from diverse fields to share their knowledge and experience and design new studies that could be undertaken. Soon after the U.S.–Brazil Dialogue, SRI submitted a field-initiated proposal to the Office of Educational Technology; the result is the collection of chapters, dialogues, and analyses presented in these two volumes.

The chapters in this volume and the analysis and discussion of the chapters in the companion volume, *What Research Should Tell Us About Using Technology in Schools*, provide both theoretical constructs and pragmatic designs that address different uses of technology within different settings for different purposes. A reading of the papers from the experts makes clear that no one study will answer all the critical questions. Furthermore, the process won't be easy, given the many different purposes for which various technologies are used and the complexity of fully integrating technology into teaching and learning. The research and evaluation designs also make clear that it will be essential to develop new assessment tools to tap into the kinds of deep understanding and complex skills that technology-based innovations are trying to foster.

These volumes will be an invaluable resource for the academic community and those who are engaged in evaluation of projects and initiatives in the United States and other countries. Examples of better assessment will help state and local educational decision makers plan new evaluation efforts using measures that are more sensitive than standardized tests at detecting technology's effects. These new assessments can not only provide evidence of technology's effects on knowledge and skills in subject areas but also reveal the degree to which students have acquired technology skills that can be used to support their schoolwork and other activities. The volumes also present examples of more rigorous research and evaluation designs that can guide the collection of evidence in order to confirm or refute causal claims about the efficacy of technology in educational settings. Such multivariate designs illustrate the need for studies that attend to the many influences that can moderate technology's effects.

I am hopeful that the challenges to be faced, along with the chance to shape the development of the next generation of technology for learning, will capture the interest and imagination of a new cadre of researchers, those not yet immersed in educational technology but already involved in the study of cognition and learning, and those who have gained their expertise in other related fields.

These volumes are timely. The Elementary and Secondary Education Act of 2002 calls for rigorous evaluations of programs, particularly those funded under Title II, Part D, Technology. Congressional leaders want to see evidence of the impact of the technology programs on student academic achievement, including the technological literacy of all students. The legislation also calls on states and districts to evaluate the effectiveness of their technology efforts, particularly with regard to integrating technology into curricula and instruction, increasing the ability of teachers to teach, and enabling students to meet challenging state academic content standards. Finally, the new education bill requires the secretary of education to conduct an independent, long-term study of educational technology.

There is much to be accomplished. I am confident that this collection of chapters and new thinking on the design of rigorous evaluations of technology and learning and the assessments that can best document their effects will make important contributions to the field in these areas: (1) gaining the attention of researchers as well as bringing experts from other disciplines into the field; (2) improving the evaluations of technology-supported innovations; (3) stimulating the development of technology-based data-collection tools and analysis; (4) shaping the federal education research agenda; and, most importantly, (5) expanding our theory and knowledge.

The work in this book and the companion volume also lends support to policies and practice that focus on the integration of technology into teaching and learning. In the process of asking better questions and improving the tools for analysis, we can enhance our ability to get the most solid contributions to students' learning from our investment in technology.

REFERENCES

Grove Consultants International and the Institute for the Future. (2000). *The educational technology horizon map and user guidebook*. San Francisco: Author.

President's Committee of Advisors on Science and Technology, Panel on Educational Technology. (1997). *Report to the president on the use of technology to strengthen K–12 education in the U.S.* Washington, DC: Author.

U.S. Department of Education, Office of Educational Technology. (2001). *E-learning: Putting world-class education at the fingertips of all children*. Washington, DC: Author.

U.S. Department of Education, Office of Educational Technology. (1996). *Getting America's students ready for the 21st century: Meeting the technology literacy challenge*. Washington, DC: Author.

Introduction to Part I

Evaluating the Effects of Learning Technologies

Barbara Means
Geneva D. Haertel
Lincoln Moses

At a time when educational research in general is often characterized as weak and inconclusive, both in national policy forums (Brookings Institution, 1999) and in congressional testimony (Bennett et al., 2000), few research areas have been more controversial than studies of the impact of technology on student learning (Oppenheimer, 1997; Stoll, 1995). With the dramatic increase in computer and network technology in U.S. schools (National Center for Education Statistics [NCES], 2000), many are looking for the fruits of the investment in technology infrastructures. While technology proponents hold forth a vision of students and teachers participating in technology-supported learning environments featuring individualized instruction, interactive simulations, and tools for knowledge representation and organization, policy makers look for positive effects on large-scale assessments linked to content standards. Some argue that traditional approaches to schooling have resisted any significant reshaping in response to the availability of technology, just as they proved impervious to the influence of television, teaching machines, and radio (Cuban, 2000). Other critics argue that technology does indeed affect schools but that its impacts are largely negative—diverting resources from more worthy pursuits (such as art, music, or basic skills) and wasting students' and teachers' time with the intellectually trivial mechanisms of technology use (Healy, 1999).

The President's Committee of Advisors on Science and Technology (PCAST), in its 1997 *Report to the President on the Use of Technology to Strengthen K–12 Education in the United States*, pointed out that given the stress on current educational reform efforts, policymakers are looking to the research base on the effectiveness of, and critical implementation factors for, applying technology to the improvement of student learning. Unfortunately, PCAST found that policymakers had little convincing evidence to turn to regarding technology's effects: "We are not yet able to answer this question . . . with the degree of certainty that would be desirable from a public policy viewpoint" (p. 87). PCAST called for a program of "rigorous, well-controlled, peer reviewed, large-scale empirical studies to determine which [technology-supported] educational approaches are in fact most effective in practice" (p. 10).

In the 5 years since the PCAST report was published, the research base on technology has grown, but it is still without large-scale experimental studies or even very many studies with carefully matched comparison groups (Mislevy et al., 2002).

Premise of Multiple Studies and Approaches

No single study, or even set of studies, could investigate all the ways in which educational technology affects student learning and technological literacy, let alone the conditions under which such effects occur and the practices and conditions that enable teachers to successfully integrate technology with academic instruction. The breadth of computer technology applications now being applied to education is staggering—and continues to grow. We have electronic analogs to virtually every educational resource traditionally employed for teaching and learning—not only books, encyclopedias, chalkboards, pencils, and typewriters but also drafting boards, three-dimensional models, slide rules, and dissecting tables. In addition, computers expand the range of skills and understandings that schools even *attempt* to impart. The high school curriculum in particular now contains topics and competencies made relevant and teachable by the availability of technology supports for working in areas such as chemistry, statistics, and design.

This volume is based on the premise that a serious investigation of the impacts of educational technology on student learning will require multiple studies and more than one methodological approach. The nationally recognized experts in research methodology who were invited to contribute chapters viewed the issues in evaluating technology's effects through the lenses of their particular experience and methodological expertise. We have grouped the chapter according to theme. The chapters in a given thematic section do not necessarily propound the same design approach, but they do touch on the same issues and challenges to doing rigorous, usable research on learning technologies.

Debate over Experimental Designs

This first section deals with what has emerged as the great methodological debate over whether experiments in which students, classrooms, schools, or districts are assigned at random to either experimental (technology use) or control (no technology use) conditions should be the method of choice in evaluating technology's effects. This debate is particularly pertinent in light of the recent passage of the No Child Left Behind Act, with its call for the secretary of education to "conduct a rigorous, independent, long-term evaluation of the impact of educational technology on student achievement using scientifically based research methods and control conditions." (See Section 2421(a), Part D, of Title II of the enabling legislation.)

The essential features of a random-assignment experiment have been set forth by Moses (2000):

- The treatments being compared are actively imposed on the experimental units, in contrast to the observation of treatments or innovations where they happen to occur.
- All treatments (or the experimental treatment and the control, or no-treatment, condition) are applied within the same time period.
- After a group of eligible study subjects is defined, each receives one of the treatments or is assigned to the control condition by random choice.
- The whole enterprise is organized and conducted in accordance with a written experimental protocol.
- After random assignment, all measures and processes other than the treatment being tested that may affect the data are symmetrically applied to all study subjects.
- The unit of treatment application and randomization (whether students, classes, schools, or districts) is the unit of a statistical analysis and defines the sample size.

Proponents of the use of random-assignment experiments within education argue that no other approach can tell us what would have happened to the treatment group if they had not received the particular treatment (innovation) under study. Drawing on experimental research traditions in other fields—including medicine and agriculture—these methodologists argue that the experiment provides the most definitive, efficient test of a hypothesis of the form *A causes B*.

Moses illustrates this power of random-assignment experiments with examples from medicine. He recalls that in the 1950s, there was a decades-old controversy about the proper surgical treatment of breast cancer. Stage

I and stage II disease were treated preferentially with the Halsted radical mastectomy in the United States, but less mutilating procedures were preferred in Europe. Each school of thought was "known" by adherents to be better. Ambiguities about how to assess the initial severity of disease and patient selection clouded the interpretation of historical data. In the 1970s, thousands of women annually were receiving surgical treatment that others were being denied—or protected from. The inability of observational data to convince allowed this dispute over treatment to persist for decades. Then, in 1977 McPherson and Fox presented the results of eight random-assignment experiments (generally called randomized clinical trials in medicine), each pointing toward the equivalence of survival with the more and less radical surgical procedures for breast cancer, and toward greater surgical trauma with the more radical surgery. Now, another quarter-century later, the proper place of less radical procedures for treating breast cancer is well established.

Other medical examples illustrate the efficiency of random-assignment experiments compared to other research methods. Early papers concerning lung cancer and smoking appeared around 1950, but it was much later, in 1964, that the Surgeon General's Report on Smoking and Health (Office of the Surgeon General, 1964) summarized many scores of studies and asserted that cigarette smoking was a major cause of lung cancer. Truth by experience (in the form of correlational data) won out eventually. But contrast that hard-won, slow success with another study. In the summer of 1954, the Salk vaccine trial (Meier, 1989) compared polio rates in a carefully designed randomized comparison of 201,000 vaccinated children with 201,000 placebo controls in the same schools. In the vaccinated group there were 57 cases of polio, and in the placebo group 142 cases; the corresponding rates, 28 per 100,000 and 71 per 100,000, differed importantly and with high statistical significance ($p < .000001$). The experiment established the effectiveness of the vaccine *in that summer*. The firm conclusion about lung cancer and cigarettes—a much larger effect—took more than 10 years to become established. The relative efficiency of the experimental approach is clear.

While examples of large-scale random-assignment experiments in education are few and far between, they are not totally absent. The most influential current example may well be the experimental test of the effects of class size conducted in Tennessee. The state of Tennessee decided to carry out a substantial experiment dealing with the class size question for 4 years starting in 1985, beginning with kindergarten and continuing through the early grades. The classes were to be of three types: small (13–17 pupils), regular (22–25 pupils), and regular with a teacher aide. Within a school grade, pupils and teachers were assigned to classes at random. The state funded the extra teachers and aides required for the experiment.

The first graders took two standardized tests in reading, the Stanford Achievement Test (SAT) for word study skills and reading, and the Tennessee Basic Skills First (BSF) tests for reading, a curriculum-based test. In mathematics first graders took one SAT (standardized) and one BSF (curriculum based) test.

The difference in performance between groups is often given as an "effect size," here defined as the difference between the group means divided by the standard deviation for individuals in the regular classes. For the standardized test (SAT), both reading and math showed a benefit of about one-quarter of a standard deviation. For the curriculum-based tests (BSF), smaller classes produced an advantage in reading of about one-fifth of a standard deviation and in math of one-eighth of a standard deviation. The addition of an aide also produced better achievement, but the advantages were smaller, averaging about one-twelfth of a standard deviation.

Special interest attaches to the effect of class size on the minorities as compared with Whites. In small classes compared to regular-sized or regular-sized with aide, the effect size for the minority students was just about double that for Whites, averaging over the four tests.

Thus, in a very large study (Finn & Achilles, 1990), at least in the first grade, class size matters and teaching support does also, though not as much. This experiment was carried out in kindergarten, first, second, and third grade. Then all children reverted to regular-sized classes. Over the years the improvement for the smaller classes held up well through the grades. In a follow-up survey for grades 4, 5, and 6, after students returned to regular-sized classes, the students who were in the smaller classes during kindergarten and grades 1, 2, and 3 continued to perform better in reading and mathematics than the students who started out in regular-sized classes.

Because this experiment was so large and so well controlled (randomized teachers; randomized students; scores of schools; thousands of students, including those from urban, inner-city, rural, and suburban areas), the results were compelling. The state of Tennessee introduced the smaller classes in the 17 districts in which students are most at risk of falling behind in their school progress. Other states, including California, have since moved to reduce class size in the early grades (albeit not always with the same positive results).

Alternatives to Random-Assignment Experiments

This part contains four chapters. In the first, Thomas D. Cook of Northwestern University, writing with collaborators from SRI's Center for Technology in Learning, makes a case for the use of random-assignment experiments in education and illustrates how this design could be applied to testing the effectiveness of a use of educational technology. The next three chapters advocate a greater emphasis on understanding the contexts within which innovations unfold and use of nonexperimental approaches to evaluating learning technology. These chapters are written by authors who have been actively engaged in working with schools and school districts implementing technology: Alan Lesgold of the University of Pittsburgh; Katie McMillan Culp, Margaret Honey, and Robert Spielvogel from the Center for Children and Technology; and Eva Baker and Joan L. Herman from the University of California, Los Angeles.

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 as well as practitioner and policy audiences.

Both Lesgold and Culp, Honey, and Spielvogel assert that large-scale and summative evaluations have not traditionally been expected to answer questions about why an outcome occurred. From their perspective, knowing "why" something happened entails knowing about the processes through which an intervention had its effects, in contrast to the experimentalists' emphasis on knowing whether the intervention caused the effects. In the view of these authors, 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. In addition, the approaches taken by Lesgold and Culp, Honey, and Spielvogel will provide specific information on the conditions in which effects are produced. Thus, their emphasis on specifying contexts contributes to stakeholders' understanding of what features of the educational setting are associated with what outcomes and for which groups of students and teachers.

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 designed to serve 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.

The question is whether such studies can lead to the accumulation of knowledge in ways that can inform practice. Moses's review of the history of various issues in medicine suggests that individual nonexperimental studies are likely to produce conflicting results and to be hard to aggregate. Lesgold, Baker and Herman, and Culp and colleagues all maintain that such aggregation of results from quasi- and nonexperiments can be done, provided that there is careful documentation of context and implementation variables and that a system for standardizing measures and aggregating data is put in place. Lesgold provides a detailed description of how jointly conducted studies, each with its own emphasis but all using an agreed-upon set of context measures or "maturity scales," might do so.

ISSUES COMMON TO BOTH METHODOLOGICAL APPROACHES

The great many different technology tools used in classrooms, and the nonfeasibility of evaluating all of them, was mentioned above. Beyond their sheer numbers, however, is a deeper layer of complexity in defining what it is that we wish to evaluate.

Defining the Innovation

No one believes that merely pulling up to a school building with a truckload of computers is going to improve student learning. Clearly, it is the teaching and learning experiences supported or mediated by technology that will or will not have the desired effects. In reviewing studies of the effectiveness of computer-assisted instruction, Clark (1985) argued that to measure technology's impact, we need to compare two sets of teachers (or, ideally, the same teachers with two sets of classes) using the same method with the same material, in one case with the material and pedagogy presented by computer and in the other, by more conventional media. Today, most developers and researchers working in the learning technology field reject this design guideline (cf. Means, Blando, Olson, Middleton, Morocco, Remz, & Zorfass, 1993). The point of using technology, developers and proponents argue, is not to do what we have always done electronically but rather to provide kinds of learning experiences that are impossible to provide by any other means. When students interact with computer-generated dynamic three-dimensional representations of molecules in equilibrium, for example, they are having experiences that cannot be reproduced with static textbook diagrams or toothpicks and Styrofoam balls. To limit technology-based experiences to interactions that can be provided offline would be self-defeating. In practice, then, pedagogy and content are usually confounded with the use of technology, and when we do comparative studies we are testing the differential effects of the package, rather than of technology by itself.

Lesgold (Chapter 2) takes this idea one step further, "[T]echnology is generally not a direct cause of change but rather a facilitator or amplifier of various educational practices." Thus Lesgold is explicit in denying that technology is the primary independent variable under investigation. Misunderstandings arise because the computers or software are often the most novel or striking feature of the intervention, and it is easy for everyone to slip into talking about the impact of *technology* rather than the impact of a specific *technology-supported* intervention.

Regardless of whether or not technology is involved, defining the program or intervention being evaluated has always been a challenge with instructional innovations. Teachers, who generally regard themselves as professionals with independent authority over instructional practice, essentially determine the fidelity of treatment implementation through the countless decisions they make every day (Cohen & Ball, 1999). Their personal views about how to make learning occur and what priority to give different curriculum objectives compete with external theories of how teachers should use educational resources. This competition typically produces variability in the way an intervention is delivered, particularly in the case of more open-ended or "concept-based" instructional strategies.

Early developers of technology envisioned that it would create greater uniformity in practice; some even predicted replacing teachers altogether. Clearly this has not come to pass, and many qualitative studies document the variability in teacher practice with the same technology (Means, Penuel, & Padilla, 2001; Shear & Penuel, 2002). Even given an order to have students spend a specified number of minutes per week with a specific piece of software, different teachers might provide their students with very different experiences. The way that a given teacher embeds the software or Internet use within classroom time, the subtle cues he or she gives to students, and the students' responses to both explicit direction and implicit cues all become variables that influence the enactment of any externally designed instructional practice, including uses of educational technology.

The extent of variation in teacher practices is likely to differ, depending on the category of technology use. The most common technology uses in schools (word processing and Internet research) involve application of general-purpose software tools that leave the instructional activity wide open for teacher decision making.

SPECIFYING THE RANGE OF IMPLEMENTING CONDITIONS TO BE STUDIED

School environmental factors—most notably, technology infrastructure and teacher technology expertise—and the school's student population are all likely to affect study findings concerning technology's effects on student learning. Minimum conditions enabling implementation must be specified with respect to (1) computer hardware, software, and networking infrastructure, and (2) teacher competency in using the technology under study. Inadequate access to technology and inadequate time to learn how to use it lead to teacher misgivings about using computers for instruction (Adelman et al., 2002; Cuban, 2001; Sheingold & Hadley, 1990). Evaluators need to avoid judging the value of a technology based on student outcomes in settings where the necessary resources are absent.

On the other hand, evaluators must recognize that demonstration or pilot programs are often implemented in schools with atypical infrastructures and technical support. When scale-up occurs, the initial findings may not be replicated with new and more representative classrooms. An analogous situation exists in medical research and is explicitly recognized by that research community. When investigating the effectiveness of a new treatment or drug, researchers first conduct what they call "efficacy" trials, in which every effort is made to provide conditions that support an "ideal" implementation of the treatment (for example, hospital personnel might administer the drug to ensure that dosages are correct and none are missed). After such experiments have established evidence for the positive impact of a treatment or drug under such controlled conditions, researchers then undertake additional experimentation to test the effectiveness in a wider range of contexts and under representative, rather than ideal, conditions (e.g., patients are responsible for administering their own medications). Where practice varies between biomedical and educational research is not the conduct of studies under carefully controlled versus naturalistic conditions, but rather the rarer use of experimental designs in education research in either type of setting.

Beyond setting minimum conditions for infrastructure and teacher training in sites to be studied, there is the question of how to deal with variation in these conditions. If the number of teachers and schools participating in a study were large enough, implementation factors hypothesized to affect the impact of technology on student learning could simply be incorporated into the design as independent variables, or covariates and interaction effects tested through statistical analysis. But because many research studies involve only a small number of implementation sites, they cannot incorporate all potentially important implementation variables into their designs, and it is usu-