Evaluation of Educational Technology: What do we know, and what can we know

Douglas C. Merrill

DRU-1049-CTI

April 1995

Prepared for Critical Technologies Institute

The RAND unrestricted draft series is intended to transmit preliminary results of RAND research. Unrestricted drafts have not been formally reviewed or edited. The views and conclusions expressed are tentative. A draft should not be cited or quoted without permission of the author, unless the preface grants such permission.

RAND is a nonprofit institution that seeks to improve public policy through research and analysis. RAND's publications and drafts do not necessarily reflect the opinions or policies of its research sponsors.
Evaluation of Educational Technology:
What do we know,
and what can we know

Douglas C. Merrill
Institute on Education and Training
RAND

April 17, 1995
Executive Summary

Education in the United States is widely considered to be in severe trouble. Technology, particularly educational technology (EdTech), has been put forth as a solution to this and other social ills. This belief, and the resulting policies focusing on EdTech, have made the opportunity costs we face extremely high. Policy makers must understand when EdTech works, and for whom, in order to make funding decisions. Answering this question requires evaluation. Evaluating EdTech is particularly difficult, however, because implementing a technology-based curriculum often involves many structural, institutional, and curricular reforms. This multitude of changes make it difficult to recognize what worked and why it did so. In fact, the better the EdTech based reform, the harder it is to evaluate it. One way around this problem has been to employ laboratory-based experimental methods. As is shown in this paper, these evaluations have revealed a great deal about the efficacy of EdTech in specific situations. However, laboratory-based methods do not address all of the difficulties with evaluation. Addressing the other difficulties will require that future research focus less on the particular EdTech itself, and more on the interactions of students, teachers and institutions with it.
Introduction

The education system in the United States has come under attack in the last several years. Standardized test scores, such as the SAT, are declining, students seem to know less and less about the world they inhabit, and employers do not value public school diplomas very greatly. Further, many schools are plagued with low attendance, violence, drugs, and many of the other social ills that face us in these difficult times.

Many solutions have been put forward in this period of retrenching and restructuring. Educational reformers, with the best of intentions, have often been guilty of jumping on the latest fad, such as “schools without walls,” the New Math, and innumerable others. For better or worse, educational technology (EdTech) is a current national obsession.

Educational technology has been hailed as the saving grace of modern education. For example, some advocates have claimed that new technologies will enable life-long learning, ease the school-to-work transition, and so forth. Indeed, in some cases, EdTech can lead to superior learning (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Kulik & Kulik, 1987). These beliefs have resulted, in recent years, in large amounts of money and attention being spent on technological reforms. However, the jury is still out on the efficacy of such reforms in general: When does EdTech work and why?

In today’s society, with money for education and research progressively tightening, the opportunity costs for devoting all, or most, of our resources to EdTech are huge. Even leaving the money aside, if technology is not “the
answer,” then focusing on it has the effect of driving other, potentially more effective, reforms out of the limelight. This single-minded focus — if it is not warranted — might also leave the public thinking “Those so-called experts do not know any more than I do, so why should I listen? After all, these are my children!”

This is a perfectly reasonable question — what do we know about education that parents do not know? The answer involves evaluation. A central part of curricular development, and educational reform in general, is evaluation. Without evaluation, we do not know whether a particular piece of technology worked, what kinds of benefits it produced, or how to apply in other situations.

In this case, the question facing us concerns whether EdTech does in fact work, and for whom. Luckily, we know a great deal about the efficacy of EdTech. For example, technology-based curricula for programming often lead to great reductions in the amount of time required to learn this complex skill, and often enable students who had been avoiding computer courses to learn to program. More results will be described in the body of this paper.

Although the field does know a great deal about the effectiveness of EdTech, the evaluation situation is not all peaches and cream. There are some questions that probably affect evaluation that are currently not answerable. For example, students often respond differently to curricula, a fact which has plagued teachers for years. Currently, we do not have the knowledge to design EdTech that responds to different student styles as well as teachers seem to do. Educational technologists have been using this sort of
issue for the past 30 years to justify additional expenditures for basic research with little consideration of implementation concerns. Given the realities of the world, we probably can not keep money and interest flowing into educational research forever, without actually showing effectiveness — thus, the research endeavor must go on, even though we are unable to answer all questions at this point.

Further, even in areas where we can evaluate the questions, EdTech poses several difficulties for evaluators, largely due to the additional reforms — outside the technological artifact itself — needed to reveal the full promise of technology for education. In fact, the goals of rigorous evaluation and maximally effective use of EdTech are often in conflict. Rigorous evaluation as commonly performed requires careful control, and implies standard experimen
tal design rules, such as “vary only one factor at a time.” However, EdTech often excels in circumstances that are quite different from traditional classrooms and reflect different pedagogical goals, thus varying many things.

Even though evaluating EdTech is extremely difficult, because of these attendant reforms, it is still necessary for policy decisions. Thus we must address these difficulties head on, so that we give policy makers an adequate understanding of when – and if – EdTech is necessary or useful. To this end, alternative evaluation methods must become standard for the evaluation of EdTech.

In this paper, I will describe some of the factors affecting the evaluation of educational technology, and present a representative sample of findings concerning the factors. These factors often lie outside the piece of technology
itself, and thus make it harder to specify what worked, why, and what benefits were produced. After discussing these general dimensions, I will discuss how these factors lead to significant difficulties for evaluators, and finally lay out some guidelines for future evaluations and research.

Many Factors Affect Evaluation

In this section, I describe some of the factors that can change the efficacy of educational technology. The factors fall roughly into three groups: Student, institutional, and task factors. Each of these should influence the metrics chosen by evaluators and the ways those metrics are interpreted. This list of factors is not comprehensive, but serves to highlight the complexity of the questions we are trying to address. These three groups will form the intellectual foundation of a new evaluation scheme that I propose later in the paper.

Student Factors

Naturally, the student’s personal characteristics interact with the educational intervention to produce differing outcomes. For example, there are students who exhibit higher ability than others, regardless of one’s beliefs about the nature of the vague construct “ability.” Higher ability students tend to react differently to feedback than students who are having more difficulty (Reiser, Copen, Ranney, Hamid, & Kimberg, 1994). Furthermore, human tutors, one of the original models for the development of computer-aided instruction (Carbonell, 1970), respond differently to remedial students than to those who
excel at the task (Lepper, Woolverton, Mumme, & Gurtner, 1993; Merrill, Reiser, Merrill, & Landes, 1994). These differences in feedback behavior are often directed at maintaining students’ motivation, another key personal characteristic affecting pedagogical outcomes. Thus, one would expect ability — however defined — and motivation to be important factors to consider when evaluating educational technology.

Students vary on the strategies they know how to employ during learning and problem solving (Perkins, 1985). Some students become expert at developing strategies for acquiring rote information (Hirsch, 1987), while others (or the same student in a different domain) may focus more on conceptual understanding (Wertheimer, 1959). Problem solving can follow a meandering path, which can lead to difficulties learning from the search (Lewis & Anderson, 1985). In contrast, other problem solving instances may be characterized by highly focused, hypothesis driven behavior. Such behavior typically leads to a solution more quickly (McArthur, Lewis, & Bishay, 1994; Merrill, Reiser, Beckelaar, & Hamid, 1992; Schauble, Glaser, Raghavan, & Reiner, 1991), but may also result in more fragile learning, because the students did not encounter as many difficulties to overcome, and hence did not have the opportunity to learn from doing so (Schank, 1990; Schank & Farrell, 1987; Schank & Jona, 1991). So, a student’s strategic knowledge will change the benefits observed by an evaluator — for example, a learning environment that is to support students’ opportunistic learning through relatively unguided search (e.g., Kass, Burke, Blevis, & Williamson, 1992) will probably not be as useful for students who do not employ this more exploratory
problem solving strategy. Does this make the environment bad? No, but it
does require careful interpretation of evaluation measures.

Another issue to consider when designing educational technology eval-
uations is that students, like laboratory animals, learn to adapt to their
circumstances, changing their behavior to optimize their performance. Stu-
dents primary criterion used to prioritize learning may very well be derived
from the testing circumstances rather than some abstract hierarchy of what
they “should” learn. This prioritization leads to the omnipresent question
“Will this be on the test?” However, students’ adaptive strategies for priori-
tizing knowledge presented via lecture may not be optimal for learning from
an exploratory environment. Thus, students may fail to learn as much as
they could have from a technology-based curriculum due to their application
of strategies that are effective in traditional school rather than learning new
ones to use in technology-based curricula. Again, this may suggest that tra-
ditional evaluation metrics – such as posttests to compare student learning
from the EdTech to what is learned from the traditional instruction – do not
measure the efficacy of the technology as it could best be used.

Institutional Factors

Student factors are not all that can affect the metrics used by an evaluation.
The way the technology is integrated within the school setting also can change
the set of appropriate measures. For example, it is often the case that schools
do not place computer-based tools in many classrooms, but rather centralize
them in a computer laboratory, run by a set of “high priests.”
This design does not lend itself to a tight integration of technology into curriculum, because using the machines requires reserving the laboratory, getting the students there, supervising them while they are working, and having the “high priests” around in case something breaks. Many traditional measures used in evaluations of technology use in schools fail to take this design into account. For example, one common early measure of the effect of technology on education was the total time spent using the computers per class. Given a laboratory design, this particular metric is virtually guaranteed to be small — but interpreting this small amount of time on computer tasks as indicative of a failure of technology is probably short-sighted.

The teacher also affects educational technology evaluation. One might suspect that technology does not require much expertise on the part of the teacher — after all, how many people in a school need to know how to replace network drivers? This argument is, however, incomplete: A teacher who does not understand what a piece of technology can do, and is designed to do, can not integrate it into the curriculum in a manner that best employs the capabilities of the technology. Indeed, teachers may even undermine the strengths of the technology by using it in ways that seem appropriate, but have little effect on the learning, such as when the class’ computers are used only as a reward for completing the assigned busy work. Thus, another possible source of interference with evaluation of educational technology is variance in the training of the teachers, and the consequent differences in comfort level and curricular integration.
Task Factors

The nature of the tasks put to students and the knowledge they are to acquire also affect the measures an evaluator should take and the manner of their interpretation. A task could be primarily procedural in nature, requiring the application of a sequence of steps, but not necessarily drawing on many “facts,” known as declarative knowledge. Much of mathematics consists of primarily procedural tasks, with the declarative portion devoted solely to selecting a procedure to apply. In contrast, the study of history involves the recognition of the relationships between actors and events – declarative knowledge. Again, an evaluator must determine which of these two types of knowledge to measure and interpret the results accordingly. It is not obvious, for example, that a multiple-choice test asking for the definitions of various mathematics operators (which therefore assesses largely declarative knowledge) is the most appropriate way to evaluate the efficacy of technology-based curriculum that strives to support students’ problem solving processes.

Evaluations have not addressed these points

Although many educational technology developers recognize the importance of factors outside of the technology itself, evaluations have not commonly taken them into account. Rather, the evaluation processes have focused on one or two measures, using them to describe the “effectiveness” of the technology, as if the technology-based curriculum could be described without attention to student, teacher, and institutional characteristics. This focus on trees (i.e., individual pieces of EdTech) rather than the forest (i.e., EdTech
policy) is understandable, given the lack of incentives for researchers to concern themselves with the broader policy issues.

Even though these evaluations have, by and large, not taken the situational factors into account sufficiently, we have learned a great deal about the possibilities for educational technology. In this section, I will describe a few of the lessons we have learned from “tree level” analyses we have performed. These evaluations have shown that EdTech excels at reducing the time required to reach a criterion of learning, allowing students to see behaviors and perform actions they otherwise could not, and supporting the connection between “school learning” and real-life experience.

Much of the EdTech evaluation has focused on students learning mathematics (typically algebra or geometry), programming (very often LISP, a language used in artificial intelligence research), or science (particularly physics).

John Anderson and his colleagues at Carnegie-Mellon University have created a sequence of Intelligent Tutoring Systems over the past several years, including programming tutors for LISP and other languages (Anderson, Boyle, & Reiser, 1985a; Anderson, Conrad, & Corbett, 1993; Anderson & Reiser, 1985; Lewis, Milson, & Anderson, 1987) and tutors to teach students to generate geometry proofs (Anderson, Boyle, & Yost, 1985b; Koedinger & Anderson, 1990; Koedinger & Anderson, 1993). Many of these tutors use an artificial intelligence technique called model tracing to give students feedback while they are solving the problems. Model tracing tutors have consistently yielded 33% decreases in the time required for students to reach criterion on posttests (Anderson & Corbett, 1993). In other words, Anderson’s tutors
led students to learn the same amount as other students — measured by a posttest — in one-third less time.

This speed-up has also been observed in another tutor for LISP programming, GIL (Graphical Instruction in LISP) (Merrill & Reiser, 1993; Merrill & Reiser, 1994a; Merrill & Reiser, 1994b; Merrill, Reiser, Beekelaar, & Hamid, 1992; Reiser, Copen, Ranney, Hamid, & Kimberg, 1994; Reiser, Friedmann, Kimberg, & Ranney, 1988; Reiser, Ranney, Lovett, & Kimberg, 1989; Trafton & Reiser, 1991). However, much of this work has focused on students' behavior during the learning sessions and on students' self-ratings of ability, enjoyment of the task, and so forth, rather than on learning time. The evaluations of GIL have taken place in laboratories, using students with little or no prior programming experience who are matched for measures of aptitude such as Math SAT. This design enables fine-grained comparisons of student performance. For example, low ability students tend to prefer highly structured environments providing a great deal of guidance (i.e., model tracing feedback), whereas higher ability students typically prefer more exploratory environments (Reiser, et al., 1994). Also, leading students to engage in behaviors that they typically would not perform, such as predicting the output of a function they are writing, leads to demonstrably higher performance (Merrill & Reiser, 1994a). The GIL experiments have shown that relatively minor changes in a problem solving environment can have major effects on student outcomes, particularly when one takes aptitude into account.

Another strength of EdTech is that it enables students to perform actions that they otherwise might not be able to do. For example, Alan Lesgold
and others at the University of Pittsburgh developed a tutor called SHERLOCK that is currently in use by the Air Force (Lesgold, Chipman, Brown, & Soloway, 1990; Lesgold, Lajoie, Bunzo, & Eggen, 1992). SHERLOCK trains Air Force personnel to repair the machine that is used to test complex aircraft avionics. This testing machine is a room sized box filled with innumerable wires, switches, and other electronic widgets — not all of which are easily accessible. Simply isolating the location of a problem is a major problem solving task, let alone repairing it. SHERLOCK supports students' learning to identify bugs in the tester, by providing the students tools to employ on a computerized model of the testing machine. This model allows students to test any wires they choose for various sorts of values, such as voltage, resistance, and so forth, without the added challenge of physically finding the wire itself behind panels and masses of other electronics. Furthermore, SHERLOCK uses artificial intelligence to help the students identify mistakes once the solution has been found. This abstracted replay has been found to be extremely useful in supporting future performance.

SHERLOCK and both LISP tutors demonstrate two of the major potentials of EdTech. First, they enable students to perform actions that would be difficult or dangerous in the real world, exemplified by students' ability to measure current across any arbitrary object in the SHERLOCK model, whereas doing so in the real world would have entailed major disassembly of the machine. Similarly, GIL allows students to measure the data output from any function at any point during a function test, and therefore to pinpoint the location of a failed expectation.
Second, EdTech also allows time to be run forward and backward as needed, as is done in SHERLOCK’s abstracted replay. This lets students repeat an action that was incorrect without the attendant side effects that resulted from the mistake, or to watch a process unfold more slowly than it otherwise would. Again, GIL’s extended testing facilities allow students to examine the processes resulting in an output that would normally take place too quickly to understand.

Educational reform of all sorts results in changes in the processes by which students solve problems, both cognitively and institutionally. The previous systems focused on student outcomes. In contrast, some EdTech evaluations have focused on process-oriented outcomes. For example, the Cognition and Technology Group at Vanderbilt University have built and field-tested a series of videodisc based mathematics curricula, called the JASPER series, whose goal is to force students to solve complex real-world problems (Bransford, Goldman, Hasselbring, Heath, Hickey, Pellegrino, et al., 1992; Cognition and Technology Group, 1990). They find that students faced with problems that do not seem arbitrary and repetitive, unlike most mathematics class assignments, are able to solve extremely complex problems with many embedded subgoals. These same problems, expressed in traditional forms, are hard even for older students. Similarly, students whose interests are piqued are able to solve extremely difficult programming problems in elementary school (Harel & Papert, 1990; Kurland & Pea, 1985), and develop understandings of meteorology that require simulations run on supercomputers and are potential topics for dissertations (Pea, 1987; Pea & Gomez, 1992).
These results become less surprising when one remembers that students do not enter school a a blank slate, despite what John Locke would have us believe. Indeed, students’ initial understandings of a domain are extremely valuable hooks for future instruction. For example, students’ real-world experiences teach them that Aristotle was correct about the laws of motion. However, modern science has shown differently, and this “Newtonian” view of motion is presented in classroom physics courses without much reference to students’ prior experiences. This disconnect results in students developing a set of school knowledge that is unrelated to the real world (diSessa, 1988), and is not used in understanding real world events (McCloskey, Caramazza, & Green, 1980). When students are encouraged to use their common sense and all the information they know about the everyday world, they are overcoming this disconnect. This allows them to solve far more difficult problems than when they are asked to ignore everything they know about the world, even though some of what they may “know” about the world is incorrect.

These impressive cognitive outcomes also affect the students’ interpersonal behaviors and the classroom structure as a whole. For example, students learning geometry from one of Anderson’s intelligent tutoring systems are more willing to try potentially incorrect solutions, and hence risk being wrong, than when they were taught in a traditional classroom structure (Schofield, Eurich-Fulcer, & Britt, 1994; Schofield & Evans-Rhodes, 1989). Further, students tend to ask each other for help and to offer help in a more positive way than often seen in traditional classrooms (Harel & Papert, 1990). The performance differences between males and females on mathe-
matics and science learning may also lessen under the new classroom order imposed by technology (Linn, 1994; Linn & Hyde, 1989). It has been hypothesized that these differences are reduced because female students are socialized to avoid the conflict that so often permeates our view of science, and hence science learning (Tobias, 1990). This hypothesis suggests a connection between these two, because taking risks, and hence making errors, is often thought to be a major contributor to learning (Schank & Abelson, 1977). If a technology-based environment allows students to get the advantages of this sort of learning without the social costs, then a broader group of people will be able to enter science. Both of these classroom outcomes have large potential policy implications, if indeed they do lead to a more gender-balanced representation of scientists.

**Summary of evaluation findings**

As this brief review has shown, we do know a great deal about the areas of efficacy of EdTech. These areas include increasing efficiency of learning (measured by time to reach criterion), enhancing student abilities by giving them access to processes that would otherwise be invisible and by supporting their use of real-world knowledge. Students also have affective responses to educational technology, but these responses vary depending on students' ability. Finally, classes that incorporate technology into the curricula have very different dynamics, with students interacting with each other more and being less afraid of errors.

Even though we do know so much about EdTech, we may not yet have
the complete story to lead policy makers to decisions. For example, all of the
evaluations cited here are in mathematics or science domains. Should policy
makers encourage the development of EdTech for social science domains?
There have been technological curricula for humanities and social science
domains (Bransford, Goin, Goldman, Hasselbring, Rieth, Sharp, et al., 1992),
but they have not yet been evaluated as carefully, because social sciences
curricula pose other problems for the evaluator. The systems described in
this section all, to one degree or another, use performance on problems solved
to judge students’ learning. In LISP programming, a “problem” is a clear
construct: Writing a program or function to accomplish some goal. However,
in history or english, problem has no such clear definition. Assessment in
these fields has often focused on facts, such as the date of a particular invasion
(Hirsch, 1987), rather than on the critical thinking skills that are central to
success in these fields (Voss, Blais, Means, Greene, & Ahwesh, 1989). Thus,
policy decisions regarding EdTech for the social sciences are not sufficiently
well-informed — but must be made nonetheless.

Another central potential of EdTech in general is to allow measurement of
more complex entities such as higher order thinking skills (Gitomer, Lajoie,
& Lesgold, 1989; Lesgold, Lajoie, Logan, & Eggan, 1990). This suggests that
EdTech may have great potential in domains where higher cognitive skills are
the goal, such as social science domains. However, by and large, evaluations
— even in science and mathematics domains, where evaluation is simpler —
have not often focused on these higher order thinking skills. The challenge
that we face as a field is to determine what we should be assessing and how to
go about measuring it. In the next sections, I will propose an alternative view of evaluation of EdTech that capitalizes on its abilities to measure complex skills. The alternative view, however, may require a reconceptualization of how we view evaluation as a process.

Reformulating Technology Evaluation

As can be seen from the previous section, many evaluations have employed traditional metrics, such as comparisons of posttest performance of students in a classroom to others using the EdTech, or have used less traditional metrics, such as on-line behavioral measures, but in settings that are not directly applicable to classrooms.

This latter category, which includes my own work, tends to emphasize rigorous laboratory-based experimental comparisons of different learning environments. This approach enables the researcher to disentangle various factors about the technological artifact itself. However, as argued earlier, the technological factors may very well be the least important of all influences on the success or failure of educational technology.

This view of educational technology as a treatment in the classic experimental psychological sense is thus inherently limited. While it is necessary to use experimental methods in the evaluation of educational technology, it is not sufficient. Researchers must also account for the institutional and professional reforms required by the introduction of technology into schools.

This sword cuts both ways, however. It is also not enough to assess educational technology that has been forced to fit into the archaic view of
educational practice, as embodied by traditional school organizations. Forcing the EdTech square peg into the traditional round educational hole avoids many of the hard questions facing us — how do institutional structures have to change to employ EdTech effectively. It is terribly easy to fall prey to this problem, with the best intentions, as exemplified by ANGLE (Koedinger, 1991; Koedinger & Anderson, 1990), a geometry tutor developed to fit into a particular piece of local high school curriculum.

ANGLE is a novel paradigm for creating geometry proofs that results in significant learning and understanding advantages for students in comparison to traditional teaching methods. However, very few people ever create proofs after leaving geometry, and the proof structure does not match the way people make arguments, so geometry training probably will not improve people's argumentation skills. Thus, geometry is probably not the "domain of tomorrow" that will revolutionize the educational enterprise. Recall that one of the major promises of educational technology is precisely this transformation — allowing students to learn new domains, such as meteorology, that affect their everyday lives but were previously beyond their means. Anderson's evaluations do not give us the information that we need in order to evaluate this promise of educational technology. Integration of educational technology into standard classroom domains, such as Anderson's ANGLE geometry tutor, helps to build credibility of educational technology as a teaching aid, and provides extremely fine-grained data about human information processing that can inform subsequent curricular design. However, it is only a part, and one that can not be the primary mode of educational technology imple-
This leaves us in a quandary: If laboratory studies are not sufficient, and assessment within the bounds of traditional subjects (without the accompanying changes) both fail to tell us about the sources of success and failure of educational technology, what should the field do? The answer, not surprisingly, is "do both and more." The research endeavor must include controlled experimental studies, theory-based institutional reform, and philosophical examinations of the nature of the knowledge students should gain from the particular instruction.

Researchers performing experimental studies must focus not on the technology per se, but rather on the interaction of the student and technology. Focusing on the interaction will require researchers to address some of the student factors listed above, and enable better credit and blame assignment. Many of commonly used measures, such as fact-based posttests, fail to get at the more complex student factors. We must broaden the metrics gathered to include measures of higher order thinking skills. Specifying individual measures of such skills is extremely difficult, so often a portfolio approach must be used (Gitomer, Lajoie, & Lesgold, 1989; Hawkins, Frederiksen, Collins, Bennett, & Collins, 1993; Lajoie, Lawless, Lavigne, & Munsie, 1993; Lane, Liu, Stone, & Ankenmann, 1993; Lesgold, Lajoie, Logan, & Eggen, 1990), in which several different sorts of products are measured. These products could be artistic works, as in the common use of portfolios, or could be different types of homework, such as reports, pictures, simulations, and the like (diSessa & Abelson, 1986; Guzdial, 1993; Guzdial, Weingrad, Boyle, &
Portofolio assessment is an extremely difficult and time-consuming enterprise, but technology can support it. For example, a rich source of data about students’ learning is to be found in the process by which they reach a solution. Experts solve problems differently from novices, and some novices employ superior learning strategies than others (Chi, Bassok, Lewis, Reimann, & Glaser, 1989). In previous work, I argued that a goal of instruction is to lead *all* novices to engage in these more profitable strategies (Merrill & Reiser, 1993; Merrill, Reiser, Beekelaar, & Hamid, 1992). EdTech allows us to measure these more profitable strategies far more easily than otherwise would be possible, such as by automatically counting the number of hypotheses a student creates, and marking them so that a human can easily find them and rate their quality.

Stated more generally, EdTech enables educators to watch the novice-expert shift taking place, and assess the process by which it is occurring. This, in turns, enables researchers to model the students’ learning in an artificial intelligence procedure. This model “knows” what the student knows and can therefore solve similar problems. This model represents the student’s knowledge state, and hence can be “graded” as a proxy for the student, perhaps on more problems than the student would solve, or on isomorphic problems that would bore the student (Martin & VanLehn, 1993). Similarly, on-line evaluation of students’ progress — evaluating the student during the learning process — enables tailoring of the length of an instructional period that is difficult, if not impossible, under the current educational regime.
Another branch of EdTech research must address the curricular and institutional changes that accompany the development of educational technology. For example, the traditional 50 minute class length arose not from concerns about learning per se, but rather because students tend to get bored and stop paying attention after that period. This is not always the case in EdTech curricula — in fact, some researchers have found that traditional class lengths are so short as to interfere with students’ learning from well-designed EdTech (McArthur, Robyn, Lewis, & Bishay, 1992; Schofield, Evans-Rhodes, & Huber, 1990). Further, the curricula that are best suited to technology are often broad-ranging and complex, but do not fit very well into traditional curricular boundaries. For example, the Collaborative Visualization (CoVis) project (Pea, 1993) teaches students meteorology by giving them access to physics, mathematics, and weather experts on the Internet, tools to solve equations and visualize data, historical records of storms in a variety of formats, and multimedia recording tools to use in developing presentation-quality final papers. Where should such a curricula fall? Computer science — it does involve some things that look like programming. Mathematics? There is certainly a great deal of mathematical modeling of behavior involved. Communications, given the project’s focus on students creating products that are communicated to other students and their use of expert’s knowledge via the Internet?

All of these, and more, are plausible choices, and indeed are probably valuable places to locate such a curriculum. However, this response begs the
question: If a single curriculum can address all of these different domains so effectively, why do we view the domains as different? Perhaps for much of school learning, the knowledge should be viewed as integrative, in the Collaborative visualization model (Pea, 1993), where mathematics is not separable from science from English, rather than the current model, where each domain is a kingdom unto itself. Indeed, a branch of cognitive science has been arguing for years that the epistemology underlying a view of knowledge as independent pieces is incorrect, to be replaced by a view of knowledge as interconnected skills and experiences (Lave, 1988; Saxe, 1992). This sort of semi-philosophical question must be addressed in order for EdTech to be employed as effectively as possible. More than simply academic, this question might suggest radically new policies about educational structure.

Similar questions arise about institutional constraints. For example, does the ease by which students can gain information from the Internet suggest that we should spend less money on school libraries — or rather less money on books in school libraries? In a slightly different vein, if we argue that the current domain divisions are arbitrary, what is to become of Education schools, and their concentrations in mathematics education, science education, and the like? The policy implications of this are staggering, and are currently being resolved without adequate understanding. Examining the school’s pocketbook is also a needed research topic — for example, what proportion of school funds should go to teacher training on the technology? Should teachers have their own computers, or should they share? These sorts of questions highlight the institutional issues that make up the school and
affect the ways EdTech can be employed and its potential success.

In this section, I argued that the view of evaluation of EdTech as a simple matter of scoring a posttest is hopelessly short-sighted. The three factors presented earlier form the basis of a new way of looking at EdTech — not in terms of individual artifacts, but as a system that includes students’ characteristics, institutional characteristics, and task or domain characteristics. This is a far more difficult research challenge, particularly since many of its proposals do not lie within the traditional academic boundaries that circumscribe researchers’ legitimate areas of inquiry.

**Summary and Future Directions**

Earlier in the paper, I presented three types of factors that an effective EdTech evaluation should take into account. These three groups provide a way of thinking about evaluation writ large, which is how we need to view EdTech evaluations.

Future evaluations of educational technology should not focus so single-mindedly on simple outcome measures, such as posttests, without also attending to more complex metrics describing the learning process, such as cognitive modeling. Furthermore, the research must also take the institutional factors into account, such as where a curriculum will fit and what changes must accompany its introduction. Similarly, each different task proposed as part of a new curriculum could require different sorts of evaluations.

The basic point is that educational technology researchers have long had the luxury of focusing solely on the relatively simple part of evaluation, be-
cause we have been concerned with developing basic principles rather than implementing solutions. However, after almost 30 years, the time has come for EdTech to demonstrate its potential, or admit its status as just another fad.

This does not mean that basic research should stop, but it does need to be focused and have additional layers of complexity added on top. The additional complexity will also require an adaptation in the traditional nature of educational research. It is no longer sufficient to leave educational research to psychologists and educators — computer scientists, epistemologists, economists, and policy makers must become integrally involved.
Reference


tion (AI-ED '93), (pp. 9-16). Edinburgh, Scotland: The University of Edinburgh.


ference on Intelligent Tutoring Systems, (pp. 222-229). Montreal:


Tobias, S. (1990). They're not dumb, they're different: Stalking the second

