TECHNOLOGY ASSESSMENT: POLICY AND METHODOLOGICAL ISSUES

CSE Technical Report 299

Eva L. Baker

Center for Technology Assessment
UCLA Center for the Study of Evaluation

September, 1989

The research reported herein was conducted with partial support from the Office of Naval Research, Defense Advanced Research Projects Agency, pursuant to grant number N00014-86-K-0395, and Advanced Design Information. However, the opinions expressed do not necessarily reflect the position or policy of either agency and no official endorsement by either agency should be inferred.

This report will appear as a chapter in *Knowledge Architectures in Intelligent Tutoring Systems* (H. Burns, C. Luckhardt, and J. Parlett, editors), to be published by Lawrence Erlbaum Associates in 1990.

Please address inquiries to: CSE Dissemination Office, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, California, 90024-1521

Technology Assessment

What is technology assessment and what is its purpose? One assesses a situation to make a judgment. That judgment may represent a conclusion (e.g., that was a bad idea), imply a general action (e.g., let's spend more on things like this), or suggest a specific remedy (e.g., whoever thought up that fiasco needs to be punished). Assessment may focus on predictions-and may generate estimates of benefit or risk for planning purposes. Assessment looks forward; it answers the questions "where are we now and where shall we go?" Evaluation, a term with a dictionary definition similar to assessment, looks back at what has been accomplished. Evaluation asks "where have we been and what to we know?" Both functions are decision-oriented. We will return to this distinction later.

Many organizations do assessment and evaluation. Let's consider a few major actors on the the national level in order to explore how technology assessment activities have occurred in the past. Perhaps the most visible actor in the area of technology assessment is the Congressional Office of Technology Assessment (OTA). OTA conducts studies requested by members of Congress after review by a bipartisan committee. These studies explore the need for legislative changes, for policy making rather than for programmatic decisions. OTA conducts its work in the following way: It pools expert knowledge, holds workshops of experts to test ideas and conclusions, conducts field-based observations and interviews, and writes carefully crafted reports.

The OTA does a good job, by all repute, with a task that is at once easy and difficult. An examination of OTA's tasks will serve as a reference point for the discussion as training technology assessment to follow. The OTA has it easy because: 1) technology assessment is the dedicated purpose of an entire agency principally composed of highly skilled staff; 2) reports are prepared for a single, known audience, but are widely disseminated; 3) OTA receives cooperation because they work for an influential and prestigious group; and 4) staff are disinterested in the outcome of the report. Since the quality of the report, not the content of its conclusions, has personal and organizational consequences, the corruptibility of the process is somewhat reduced. OTA's task is hard because: (a) their client is Congress so they need to be circumspect; (b) they must address technology in all its manifestations, from lie detectors, waste management systems, and pacemakers; (c) they have very limited amounts of time to conduct their studies; and (d) the exact uses for studies may not be explicit. OTA helps Congress and the public understand the state of the art of various technologies so that more informed public policy decisions can be made. These decisions are likely to be forward looking, focusing on whether given regulations are appropriate and whether new technologies look promising for solving national problems. Other agencies perform similar functions. For instance, the National Research Council (NRC), which is outside of government control, conducts studies on a broad array of national issues, including those with implications for education, defense policy, and society as a whole. They also review the utility of particular initiatives, such as military performance testing, or may report on general status, such as the social and economic progress of blacks, or make specific recommendations for action, such as how national tests should be revised. The NRC relies on expert judgment to even a greater degree than OTA and therefore takes every precaution to assure objectivity. The findings of both of these agencies are typically widely disseminated and receive strong media attention. Both prepare assessments-that is, they make general estimates of their topic-using a variety of expert sources.

Contrast the work of these two agencies with the responsibilities of the General Accounting Office (GAO), which also reports to Congress. GAO focuses on what has been accomplished, the adequacy of procedures used, and, as its name implies, the prudence and cost of the endeavor. GAO looks at specific programs and

organizations rather than general states. They require information from and make judgments that directly impact on agencies and their staff. For these reasons, their work comes close to incorporating the concepts of evaluation as it is typically used in the social sciences.

In training technology, perhaps because technology products are so concrete, the tradition has been to focus on evaluation-on what has been accomplished. For at least four reasons, the concept of technology assessment should be substituted for evaluation in our thinking. First, evaluation, as it is employed in education and training environments, connotes to many a relatively narrow set of methodological choices. Evaluations often are assumed to have certain features. Evaluations appear, for example, to be empirical in nature and as such, are obligated to: 1) collect data using designs similar to those employed in experiments (i.e., control groups), 2) use quantitative analysis as the basis for inference, and 3) focus on summarizing and reducing data. Viewing evaluation as bound by constrained methodology is a widespread misperception.

A second, hoary belief is that evaluation should be bisected, like hamburger buns or angles into only two pieces. These evaluation sections are known widely as summative and formative evaluation. Believers in this analysis categorize studies as either those whose purpose is to make decisions or those whose purpose is to improve programs-as if such functions were mutually exclusive.

A third and seriously limiting conception of evaluation can be traced to the systems approach underlying most evaluation models: the use of limited criteria for judgment. Such models almost always compare the performance of the intervention exclusively against relatively simple requirements (i.e., desired performance objectives) although we know that many more outcomes are usually affected.

Fourth, evaluations normally address individual instances or interventions, such as a single algebra tutor. Such instances often are compared rather facilely with an apparently valid alternative (e.g., trainees taught through lecture). Somewhere purveyors of this research lost the idea of sampling, sampling topics, program designers, and instructors. Because of practical constraints, if not methodological impairment, results reported for studies of this type have no generalizability. (See Leifer, 1976, for an excellent analysis.) The lack of generalizability applies to the technology studied (e.g., intelligent tutoring systems) and for typical comparison conditions (lecturers) in such comparisons.

Training Technologies

What is the range of training technologies to be assessed? We all have different images of existing and ideal training technologies, an assertion that can be verified simply by asking a colleague for a one sentence definition of the term. One definitional problem is to capture the variability of training technology in a useful way. Table 1 (page 3) lists five dimensions along which training technologies vary.

To some, the term technology denotes hardware requirements, but this is a good place to remind ourselves that technology means "systematic treatment" or "applied science" (Miriam-Webster Inc., 1989). For a process to be labeled a technology, then, it only needs to be reproducible, or able to be used in repeatedly with the same consequences. We have come to expect that the consequences of technology use will include both reliable results (e.g., the telephone usually works) and efficiency (e.g., fax is faster and cheaper than overnight mail). To further illustrate this point, we would count as technology certain reproducible procedures that exclude hardware trappings. For instance, research-validated procedures used to organize teams of students for learning tasks is a form of technology, even though

Table 1

Dimensions of Training Technology

Equipment Intensive	<--------->	Equipment Free
Mature	<--------->	Nascent
Comprehensive	<--------->	Adjunct
Systems Targeted	<--------->	Multipurpose
Training Specific	<--------->	Training Adapted

no hardware would be required for this application. Such procedures are sometimes called "soft technology.". On the other hand, many applications may be technology-dependent, involving extensive hardware systems, from videodiscs, television, and computers to elaborate simulators designed to model complete environments such as extraterrestrial systems.

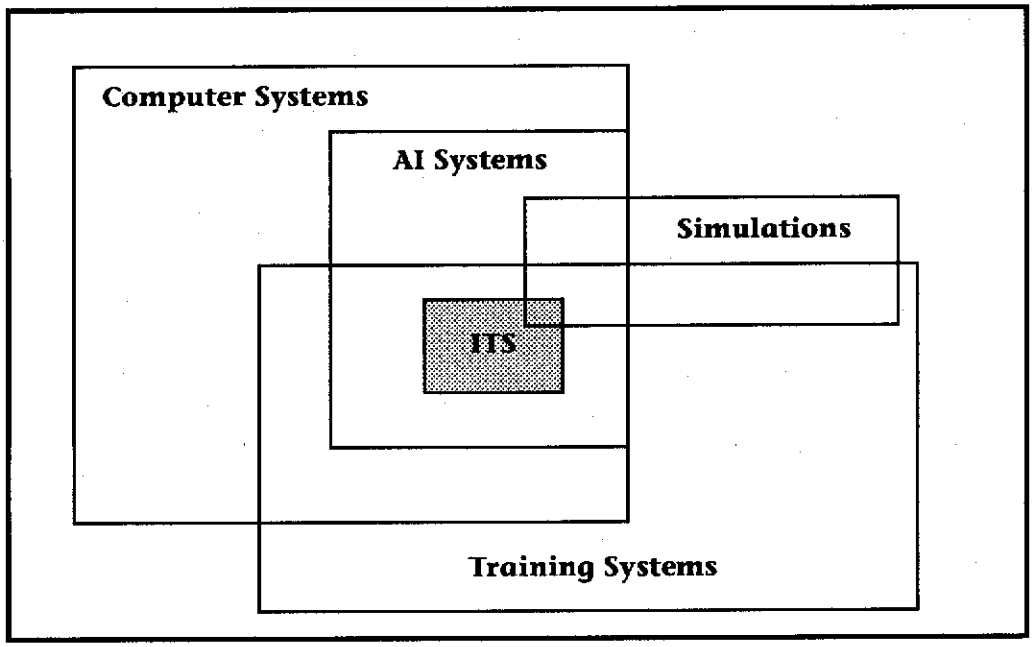
The second dimension shown in Table 1 relates to the technology's level of development, a complex dimension that embraces more factors than those of age. The technology can be classified as nascent or mature. At its outset, new technology has greater risks associated with its exploration and eventual impact: It simply may not pan out. When a technology is mature and a known quantity for certain applications, risk is reduced. However, other liabilities exist. Familiarity may be tinged with contempt and may foster the belief that you only can do what already has been done. This factor may work against the adaptation of a mature technology for new applications. For example, we know well that computer assisted instruction (CAI) improves efficiency of training (Levin, Leitner, & Meister, 1986). Yet, many potential users may continue to resist their use of CAI because early CAI programs were mundane and illustrated only limited understanding of learning psychology. Such users cannot imagine how CAI might be changed to become more interesting.

A third dimension of training technology that is relevant to the consideration of technology assessment relates to the centrality of the technology in the overall training or educational plan. We can imagine computer-assisted training where the system bears exclusive responsibility for teaching the trainee. Contrast that situation with one where the bulk of information and practice are accomplished through lecture and workbench activities and where the computer system presents problems for only the hard-to-train student.

A related issue is whether the training technology is designed exclusively for a particular system or planned to serve many training purposes. Designing a system that trains an individual to troubleshoot a particular system, for instance, to repair a named radar, or to fly a particular airplane has different conceptual demands and consequences than designing a technology that will be used for a particular first purpose, and simultaneously, maintain a view of a larger set of applications. (O'Neil, et al, discuss an example of this latter sort of technology in their chapter.) An extension of this contrast in design foci occurs in the case where the technique or system is developed specifically for training environments or compared with an application adapted from other contexts.

The interrelationships among these dimensions are obvious, and wherever any particular technology falls among them, the juncture has clear implications for technology assessment. Figure 1 depicts some of the major sets of training technology elements. The shaded portion, ITS, displays where intelligent tutoring systems, the topic of this book, falls within the training technology array.

Figure 1
Training Technologies



Using the dimensions presented earlier in Table 1, we would classify ITS as a technology that was hardware intensive, nascent, conceived typically as a training adjunct, and usually targeted for a particular system. Whether ITS is classified as training specific or as an adaptation of the other artificial intelligence (AI) technologies depends upon one's viewpoint. Probably nothing new for the AI field will come from the development of ITS, so from that perspective ITS is simply an adaptation. If you believe, on the other hand, that tutors require synthesis from a range of disciplines that include psychology, education, and content domains, then ITSs may be viewed as training specific creatures. Either way, they are comparatively expensive to design and implement on a large scale basis.

Characteristics of Technology Assessment

Technology assessment should supplant evaluation if only to avoid the enumerated liabilities of an older term. But exploring the concept of technology assessment is appropriate to our present discussion for a number of important, additional reasons. Consider first that training technology itself-what is being assessed-differs fundamentally from other instructional interventions. Technology is interactive, dynamic, and develops rapidly, often in astounding leaps and surprising directions. Paradoxically, the power of technology continues to expand, as its cost, with relatively small bubbles, continues to drop. Thus, to think of technology as simply another delivery system, comparable to lecture-discussion, is to miss the conceptual boat. Decision makers should not focus only on short-lived races between one instructional delivery system and another. When new technology first gets built and evaluated, it usually fares poorly in comparison to well-established practical alternatives, such as lectures and books. Thus, the initial effects of the technology are almost always underestimated. Studies of technology must be especially sensitive to the notion of technology-push (Glennan, 1967; United States Air Force, 1986)-the idea that technology bumps up against the usual requirements-driven programs in odd and unexpected ways-for technology is almost guaranteed to

generate, by its very existence, outcomes and applications that were not previously considered by the training system, nor imagined by the technology designer. These new uses may be described mistakenly as side effects, when, in fact, they may be the delayed but central outcomes of the innovation. A critical element in technology assessment, therefore, is identifying when these options represent powerful, useful approaches, goals, or recombination or redefinitions of prior goals.

As a corollary, new technology, more than other type of innovation, should not be shut off because its superiority on existing goals cannot be demonstrated immediately. For example, one effect of designing tutors may be the development of technology to create new kinds of human performance measures (Baker & Linn, 1985; Lesgold, Bonar, & Ivill, 1988; Collins, 1987) and new ways of conceiving performance tasks (Means & Gott, 1988). It is possible that such practical and conceptual outcomes may be more important than the adaptive wonders of instruction that particular intelligent systems are purported to create. If we are to develop clear traces of the broad utility of technology to meet training needs, studies must involve analyses that range far beyond what the technology designer or any given set of trainees believes or experiences. Policy makers need to be involved early and actively to determine what options should be highlighted, tracked, and ultimately ratified as bonafide new goals and functions.

The detection of the unforeseen has fundamental requirements that policy makers should consider. These requirements involve changing expectations about the purpose of new development. At minimum, policy makers must accept a period of suspended disbelief and a planned commitment to the conduct and the analysis of a network of studies of individual cases of technology. Because it takes time to execute such studies, they cannot be the sole initiative of an individual who is committed to only a limited period of assignment. Some larger, longer-term policy must be put in place. To reiterate, the purposes of such investigations focus on not only the differential impact of particular instances-tutor A versus option B-for particular tasks, but the larger and more important task of forecasting the utility of a class of technology. Thus, the explicit goals of technology assessment are dual: the case, usually against a specific training requirement; and the class, forecast for known and uncertain future requirements.

Technology Indicators

Because the view of technology assessment is more global than that of product evaluation, so is its methodology. Although a product evaluation might be interested in relatively well-specified conditions, technology assessment attempts to determine the full range of use. To do so, the model underlying technology assessment should seek to represent thoroughly the conditions that contribute to its impact as well as to take an expansive view of impact itself.

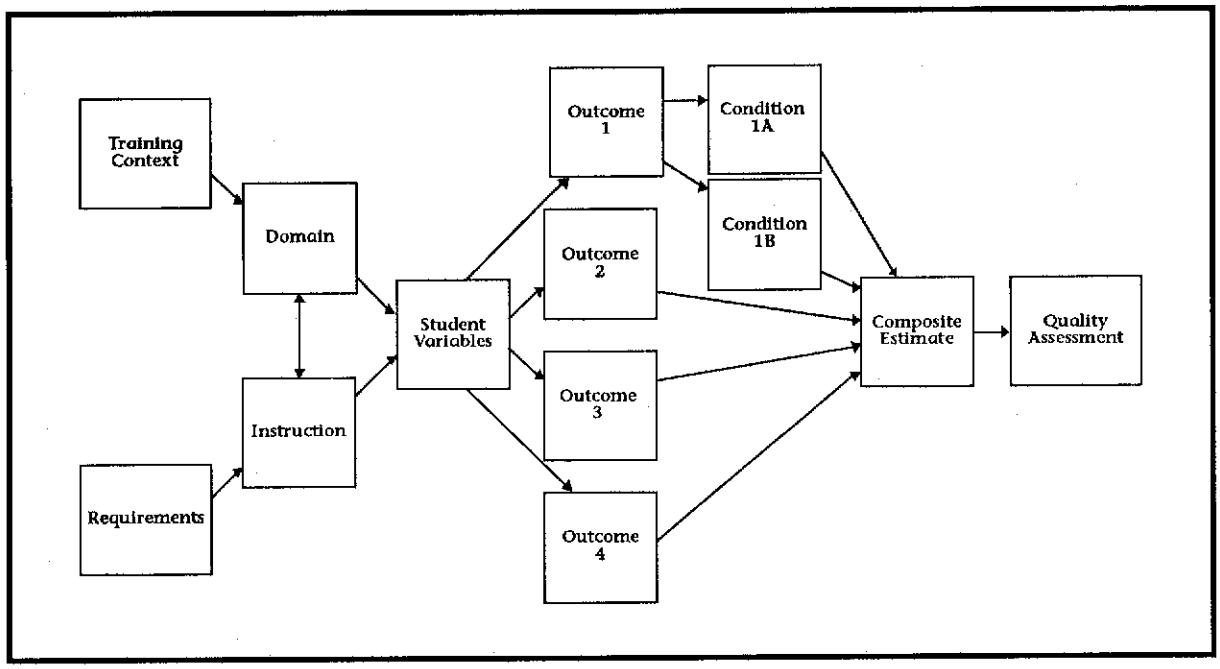
Two requirements flow from such a model: (a) the system studied must address estimates of input, context, and conditions of implementation as well as the outcomes specified above; and (b) multiple measures of major dimensions must be employed. Such an approach can be called a training technology indicator system.

Educational policy makers at the federal and state levels have adopted the indicators metaphor as a way to assure that they collect data on aspects of policy under their control (e.g., the requirements for teacher training) in addition to those measures they affect only indirectly (e.g., students' standardized test scores). Such an approach is especially suitable when the object of study is technology, in the light of its potential portability and flexibility. Considered as a system, input (such as requirements and trainee populations) and context and conditions (such as the tasks required, time availability, financing, criticality, and outcomes, including requirement-driven outcomes and unforeseen effects) present a firmer basis for

decision-making. Contrast this approach with the more usual evaluation study, which has limited scope. Generalizations about the use of instructional tutors that are based on one carefully done laboratory study with fixed input (the trainees and instructors), limited conditions (a four-day course on some topic), and one or two outcome measures that can be quantitatively scored produce decision confidence only when the cost of being wrong is small. As in the field of economics, absolute meaning in evaluation does not inhere in any measure, but in the behavior of indicators under variation. What becomes important are the interrelationships and the changes in indicator value over time; sophisticated statistical analyses are available to support these causal inferences.

The second component of a technology indicators approach that makes it especially suitable for studying technology is its reliance on multiple measurement of single variables (e.g., achievement). Instead of a single check list, achievement outcomes might be composed of a number of single measures, such as student problem solving, problem identification, efficiency, attitude, and instructor or commander estimate of proficiency. Creating composites of these single measures allows an overall estimate of quality to be determined. Varying the parameters or weightings of individual measures permits a decision maker to examine and make explicit value structures and, at the same time, to study patterns of relationships over time. Such a systematic approach also permits the longitudinal review of policy decisions. Figure 2 depicts an indicator model, with boxes identifying elements that would be measured. Notice that multiple measures are depicted only for Outcome 1.

Figure 2
Indicator Model



The goal of this model is to develop training quality indicators that provide composite estimates of variables and the relationships among them, much as economic indicators provide composite descriptions/forecasts. This indicator assessment perspective, by the way, is apparently useful to state policy makers, legislators, governors, and educational boards and superintendents as they try to

determine systematically and longitudinally the consequences of policy changes intended to improve the quality of precollegiate educational services (see, for instance, U.S. Department of Education Office of Educational Research and Improvement, 1988). Note that multiple indicator development is a natural opportunity for collaboration among branches and services that assess technology. Yet, full-blown quality indicator systems for technology are still a long way off. Many specific problems must be addressed first .

Technical Issues

Although it is easy to describe an ideal technology assessment system, it is unfortunately true that a litany of technical issues must be confronted in order to implement a technology assessment approach to training innovations. The assessment design must consider the number of alternative outcome measures for any given set of variables (e.g., instructional context, the timing, valuing and weighting procedures of such variables) as well as the choices among statistical models for analysis. Each of these issues is extraordinarily complex, but in some way, each also presupposes that the outcome measures of interest have been assessed adequately. No one would invest seriously in causal modeling with a fallible criterion. Even the most limited assessment needs high quality outcome data to be credible and to estimate present and predict future impact. One obvious source for identifying outcome measures resides in the goals adopted by the designer during the creation of the new technology. In the case of ITS, we would look at the assembled empirical base and make a judgment about what should be assessed. Unfortunately, existing ITS literature is a barren source for good examples of outcome measurement. Very few studies address the problem in any systematic or explicit way that exploits the potential power of intelligent systems. This criticism holds in part because designers have been so focused on the intricacies of making systems work, or even designing pieces of systems such as an expert problem solver, that whole tutors rarely have been produced, let alone empirically validated. Even when outcomes are measured, the techniques used rarely approach the state of the art in other aspects of achievement measurement. Imagine a not-imaginary example where the designer accommodates outcome assessment by slapping on a standardized measure of learning, or looks at job performance by throwing in a simple check list of correct procedures. Making decisions based on such data is equivalent to listening to a symphony on scratchy LPs when compact discs are available.

How have tutor outcomes been measured in the past? In a series of empirical studies undertaken at UCLA, we (Baker et al., 1985) were given the task to evaluate three different tutors that were nominally available or in development at that time. We could find only two such systems that were reasonably amenable to the task: WEST, a program designed to teach number facts in a game strategy (Brown & Burton, 1978) and the program analyzer section of PROUST (Johnson & Soloway, 1987). We hoped to develop a complex set of dependent measures to assess the full range of outcomes for these systems-outcomes that were claimed by the designers and outcomes that could be inferred from system operation. We developed attitude measures and domain-referenced achievement tests based on the designers' articulated goals and, in addition, collected other aptitude measures, such as Scholastic Aptitude Test scores or verbal, spatial ,and mathematics reasoning scores. We developed measures that we thought captured important goals of each system: in WEST, arithmetic skills and the strategy used to play the game; in PROUST, the quality of the Pascal program generated by students. WEST students did not show much improvement when compared to controls. The particular student computational goals and prerequisites articulated by the designers sadly missed the target. That is, students who passed the pretest also passed the posttest without instruction, whereas students who did not possess prerequisites never

learned enough to profit from the WEST experience. Strategy, when measured as the solution to particular WEST board problems, was not affected by the WEST practice sessions. In the case of PROUST, the developers ultimately did not permit the use of the program writing measure, preferring an option where students analyzed bugs in Pascal programs (ironically, this was also what the computer did). Our technology assessment experience was not one to lead to great confidence in designer-generated measures. A more positive finding was obtained by O'Neil and Nizamuddin (1989), where they found effects while looking at new outcome measures for treatment variations of Sleeman's Algebra Tutor. (This study was particularly interesting in that, in addition to measures of students academic ability, it systematically assessed students' anxiety reactions.)

So it is very possible to develop outcome measures that are sensitive to technology concerns. Why doesn't this happen more often? One jaded view is that effects aren't measured carefully because it is better not to do so. A second reason may rest in the nature of the interests and expertise of the experimenter. It is our view that outcome measurement of complex training is so important that it cannot be left to the designer alone to accomplish. Designers may lack the expertise to create good measures. As noted earlier, important training outcomes need multiple measurement across time and conditions to consider effects beyond the short-term achievement of the training goal. These dimensions include retention, robustness of performance across field conditions, transfer, and assessment of underlying constructs to facilitate cross training.

New developments in performance assessment lean heavily on trainee-generated performance, or constructed responses, using constructs from cognitive learning theory to derive scoring attributes (Wittrock & Baker, in press). The primary message of this development is that measures must map back to characteristics of learning (i.e., elaboration, schema, and problem detection; see Hayes, 1989, and Marshall, 1989). Such approaches are partially validated by using expert-novice distinctions (Chi & Glaser, 1980; Baker & Clayton, 1989). This concern for the close relationship between learning and measurement conditions contrasts strongly with the majority of current outcome testing practice, in which convenient test formats strongly limit what we are able to say about student performance.

A second direction in performance assessment (Baker, O'Neil, & Linn, 1989) involves improved, more sensitive ways to select and train judges of performance. No longer is simple designation as a subject matter expert-or, for that matter, a tutor designer-sufficient to assure reliable and valid rating of performance.

There will be some resistance to the dictum of multiple indicators of outcomes, particularly from those stuck in a frozen view of evaluation. Remember that the term evaluation still calls up for many the specter of a single, monolithic methodology, largely derived from social science, experimental, and quantitative in nature. That view may have accurately characterized the majority of social science research twenty years ago, but accounts only for a limited proportion of current effort. It also is true that the field of evaluation was fractionated into methodological camps a dozen years ago, with lines drawn between quantitative experimentalists and qualitative interpreters. However, at present, a more balanced blend of methodology is common, and desired. For example, relatively objective forms of performance assessment are often mixed with qualitative analyses of protocols of trainee thinking (Feifer, 1989). Intensive descriptions of processes-for instance, knowledge engineering (Baker, Novak, & Slawson, 1989)-and understanding queries in natural language systems (Baker & Lindheim, 1988) can be combined with surveys and more traditional test forms to provide a more complete explanation for findings.

Criteria for Technology Assessment: Choosing What Gets Assessed

We have discussed methods for collecting information including the use of multiple sources of information. What guidelines can be used for selection and design? Table 2 summarizes these criteria for technology assessments.

Table 2

Criteria for Technology Assessment

Selection Criteria

I. Risk-Balanced Portfolio

II. Potential Benefit to Participants

Design Criteria

I. Impact

- individuals
- units
- training systems

II. Costs

Risk-Balanced Portfolios

Whether empirical or expert-based, a technology assessment efforts should include a balanced portfolio of cases. To provide a fair and responsible assessment, these must include efforts that represent various levels of risk. Risk can be assessed in terms of the emergent or mature status of the technology, its place in the R&D cycle, its scientific knowledge base, its payoff, the length of its development cycle, the difficulty of the goals undertaken, and its potential for generalizable or targeted use.

Benefits to Participants

Attention should clearly be given to technologies that have higher potential benefit to significant users in any system. How one decides user significance is a matter of prior policy development. Criteria such as the potential number of users, the criticality of their roles, and existing performance deficit, among others, will come into play to make this decision. Important social values should also be considered, including issues of equity and self-worth .

Issues for Review in Technology Assessment

Impact. In addition to the review of benefits to individuals as a selection criterion for the review of technology, impact on individuals should be considered as a main criterion in the design of the assessment. Factors that should be weighed include who will feel the impact, to what degree, and with what anticipated outcome in terms of learning, competency, adaptability, readiness, etc. Related to this issue is the impact of the technology on the unit or collectivity to which the individual is assigned. A connected concern is the potential impact on the existing training organization. What changes will need to occur in the training requirements

because of technology use? What changes in the training approach or sequence itself are anticipated? How will staffing be affected? Are adjustments anticipated to be major or marginal? How might they be phased?

Costs. Costs of various levels of implementation of the technology need to be projected. This must be computed with normal concerns for life-cycle costs, including start up and maintenance. With technology, cost estimates may be difficult to compute. Consider that the overall cost of computational and visual power will drop over time. But technology creates appetites for better, faster, glitzier implementations that shorten the interval for technology obsolescence and the requirement for reinvestment. The impact of a CAI program developed ten years ago is surely less than it might be today because of the rapid advances in graphics and distribution of computational support. Thus, the changing context and expectations for technology performance contributes inevitably to shorten the life of any particular implementation. A related concern is how the costs are distributed to agencies, industries, and other funders of technology development work.

Certainly, any assessment will include a set of tradeoffs among criteria such as risk, impact, organizational effects, and costs. The manner in which such tradeoffs are treated probably distinguishes the overall quality of the assessment. But in addition to criteria, organizational realities rear their often ugly heads.

Procedures for Technology Assessment

The goal of technology assessment is to provide a larger view of the utility and potential impact of a class of technology for a set of potential uses. Thus far, we have described how technology assessment is conducted on behalf of the public by national agencies. We have pointed out some shortcomings of the few existing efforts, more evaluation of individual cases in the area of intelligent tutoring, and desirable characteristics for future efforts. But the discussion has been abstract. In the operational conduct of technology assessment, what should be achieved? First of all, a given technology should be assessed systematically as a class. Two sorts of analysis are appropriate: empirical analysis and expert analysis (see Table 2).

Table 3

Procedures for Technology Assessment

Empirical Analysis

- Preordinate
- Natural variation
- Post hoc

Expert Analysis

- Panels
- Surveys

Empirical Analyses

An ideal approach to empirical analysis is a design strategy, where plans are made to coordinate the investigations of a technology across organizations or agencies. In that way, a range of tasks, contexts, developers, outcomes, and users (or

other variables of interest) could be analyzed in parallel natural experiments. This preordinate, empirical approach runs into a variety of difficulties, including funding coordination, competition among agencies, and so on. Certainly it also would be unusual for the full range of variables of interest to be represented in the planned studies. A second alternative is post hoc analyses of the sort conducted by Levin et al. (1986), writing on the utility of computer-based instruction. These studies are based on the available literature, usually on existing evaluations of individual systems developed for a unique purpose. Post hoc analyses attempt to generalize to the class of technology. But their utility is limited also. Available studies will differ in quality, in reporting detail, and in nature of comparisons, so inferences drawn from meta-analytical approaches must be interpreted with extreme caution. Because these studies were not designed with the class of applications of interest to the policy maker in mind, significant gaps in knowledge may exist. Yet, some sort of empirical, how-does-it-work evidence is essential to an appropriate assessment.

Expert Analyses

Expert analyses can be conducted through the use of panels of knowledgeable individuals where judgment of utility, future impact, range of applications, limitations, barriers to implementation, and like topics are considered and some kind of consensus is reached. Expert analyses are helpful if the experts have credibility to the policy audiences they are ultimately addressing and if they can be kept to some general set of policy issues and recommendations. Structure for such groups is essential and a policy statement or other product is clearly desirable. Another form of expert analysis can be developed through the use of survey and interview techniques. In these cases, broader samples of experts may be consulted, and the judgment reached will be an aggregation of individual views rather than the group consensus, although this outcome can be altered by use of Delphi approaches to survey consensus. In addition to losing the creative interaction groups can achieve, survey or interviews are many times conducted in a relatively decontextualized way, without sufficient preparation to get the expert up to speed on the particular set of relevant issues. The trade-off is probably breadth versus depth.

It is our recommendation that both empirical and expert approaches are used in technology assessments. The ideal would be to use preordinate and post hoc empirical studies as well as group and individual expert analyses.

Organizational Issues

Two major types of organizational issues potentially impede the use of more comprehensive approaches to technology assessment. One is the perceived social impediments surrounding assessment. A second is the legitimate organizational boundaries that make difficult such tasks. Let's consider the social interpretation first.

One liability that hangs on from the era of evaluation is the inferred political impetus of any decision to assess anything. Most efforts at assessment or evaluation provoke some level of resistance, resentment, or defensiveness. No one really believes the slogan "we're here to help," and they are often correct. The person who is evaluated may believe evaluation is an instrument of aggression (for evaluation only occurs when someone has a problem). Program managers, on the other hand, may use evaluation defensively. They primarily may be interested in it as a defense against future assaults, rather than for the information it provides about innovation. Over and above usual paranoia, additional issues deserve comment, since much assessment is a social as well as a technical enterprise. One issue is: Who does the assessment for what ostensible purpose? If the designer of the technology

is responsible for assessment, one is not only limited to a particular vision, but also inevitably confronted with self-interest. Furthermore, designers are more committed to the task of creating systems than to creating systems that result in demonstrable trainee outcomes. Particularly in computer-based technologies, the trick is to make a system run according to prediction. The importance of process is highlighted in an article on AI "Evaluation" by Cohen and Howe (1988). As they describe evaluation, it is limited to an expert review of the quality or process of research efforts. This article was especially heartening for me because it confirmed an earlier conclusion about the distinctions among expectations of high technology researchers, program managers, and evaluation and assessment professionals, and the resulting social complexity of getting the job done.

Societal constraints are a matter of semantics and marketing as well. Researchers may propose to create a training system, program managers may think that's what they bought, and those charged with assessment may assume that training outcomes should be measured. In fact, the likelihood of strong outcome effects increases only with the maturity of the technology. With a new technology, the designer's claims and focus on a training system simply may provide necessary limits for what he perceives as a research problem; the training focus is only a means to conduct research. The researcher may say, and believe, that the chosen task is to develop an intelligent tutor to teach specific outcomes, but what the researcher means is that research will be conducted on an interesting part of the problem of developing a tutor. Researchers are not the same as training system developers, and this fact is demonstrated recurrently by the woeful number of partial systems in the intelligent tutoring community: tutors without student models, tutors with wonderful diagnostic capabilities, tutors without pedagogical modules. Awareness and understanding of the various contexts and subtexts of communication among researchers, managers, and assessors may allow some form of collaborative assessment to work. Program managers can benefit from an understanding of underlying messages, particularly when they may have obtained priority for funding a particular technology program by promising a product for an actual training system.

A more obvious difficulty, especially within newly emerging fields, is the insider/outsider problem. Expertise and expectations differ and suspicions abound, not only between measurement specialists and AI researchers, but among linguists, psychologists, and AIers, and within the AI community, between devotees of one or another approach. We have experienced this phenomenon in our DARPA project on assessing AI systems. We have tried numbers of options to bridge the communication and knowledge gaps, including hiring AI people, providing incentives, using consecutive translation, and throwing ourselves on the mercies of friends. The trade-off is objectivity and detachment for credibility and insight. A solution, of course, is to train people who become proponents and experts in the assessment of technology, but that will happen only when the technology has a surplus of researchers-a self-contradictory state when the focus is new technology. Yet, collaborative teams are at work. How they forge successes should be an interesting story that most probably will unfold sometime in the future.

The second organization issue in technology assessment concerns bureaucratic reality and organizational boundaries. Technology assessment is recommended because it will give program managers a better estimate of where to invest resources. The problems described in this paper identify requirements for technology assessment: 1) taking a long-range view; 2) focusing on a class of technology rather than single copies; 3) multiply measuring an integrative set of variables in an indicator system; and 4) infusing assessment expertise into a social situation already made complex by promises and suspicions. These requirements can be met in a situation that assures growth in resources and some spirit of cooperation among decision makers who have agreed to share a vision. The current reality for many technology R&D agencies, private or public, is such that rich resources are a

dim memory, fading fast. Bureaucracies also inhibit most forms of cooperation. In the military, the long-standing competition and the vastly different cultural norms among services discourage such interaction. It is possible, however, that cooperation may be the only way to accomplish much at all in a time of declining resources. Risk is shared and relatively low; benefit may be high.

Reporting

A final, and overlooked, area is concerned with the nature of reporting useful information from assessment for various levels of program and policy decisions. The identification of the full range of audiences is a critical point, as is the understanding that any data or conclusions can be used or misused against you. The challenge is to find ways of communication that will contextualize results appropriately, without endless qualification, reams of tables, or micromud descriptions that put off all but the most devoted reader. One area of general interest focuses on identifying the report users' mental models and their options for making effective decisions. If we could apply what we know from cognitive psychology to assist sophisticated decision makers' process and integrate assessment findings, we would develop clues related to what information was most relevant. Furthermore, one might expect that such report readers might themselves need a modicum of training in order to assure that more than one reader would reach one set of conclusions given similar findings.

Research and Development Implications

Short Term R & D

The issue enumeration above leads directly to some recommendations for R&D activities to advance the field. First, in the general area of technology assessment of intelligent tutoring systems, it will be important to categorize systematically the existing and developing defense-supported tutors by attribute, technical approach, and training task. UCLA has undertaken this task for DARPA in the area of natural language (NL) understanding systems and has created a sourcebook of the problem types that natural language systems address. A second short-term project involves the creation of advisory or assessment authoring systems particularly suited to technology assessment problems. A prototype system has been developed at UCLA on the narrow problem of reliability for criterion referenced tests, and costs for a library of such aids are relatively small. A third activity might be a case study analysis of an attempt at class-oriented technology assessment of ITSs, using naturally occurring and planned assessments. Fourth, research on decision maker's mental models could be conducted to provide a better understanding of assessment and reporting requirements.

Long-term Studies

The design of seriously planned embedded assessment systems that includes the full range of input, process, and outcome data, such as individual differences, process, trainee performance, retention, and transfer data could be undertaken in a long-term study.

References

Baker, E.L., & Clayton, S. (1989, June). *The relationship of text anxiety and measures of deep comprehension in history.* Paper presented at the Conference of the Society for Test Anxiety Research, Amsterdam, The Netherlands.

Baker, E. L., Aschbacher, P., Feifer, R. G., Bradley, C., & Herman, J. (1985). *Intelligent computer-assisted instruction study* (JPL Contract #956881). Los Angeles: UCLA Center for the Study of Evaluation.

Baker, E. L., & Lindheim, E.L. (1988). *A contrast between computer and human language understanding.* Los Angeles: UCLA Center for the Study of Evaluation.

Baker, E. L., & Linn, R. (1985). *Institutional assessing and improving educational quality* (Proposal to the National Institute of Education for the Center on Student Testing, Evaluation and Standards). Los Angeles: UCLA Center for the Study of Evaluation.

Baker, E. L., Novak, J., & Slawson, D. (1989). Feasibility study of an AI testing advisor (Report to OERI). Los Angeles: UCLA Center for the Study of Evaluation.

Baker, E.L., O'Neil, H.F., Jr. , & Linn, R.L. (in press). Performance assessment framework. In S.J. Andriole (Ed.), *Advanced technologies for command and control systems engineering.* Fairfax, VA: AFCEA International Press.

Brown, J.S., & Burton, R.R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2,* 155-192.

Chi, M.T.H., & Glaser, R. (1980). The measurement of expertise: Analysis of the development of knowledge and skill as a basis for assessing achievement. In E.L. Baker & E.S. Quellmalz (Eds.), Educational testing and evaluation: Design, analysis, and policy. Beverly Hills: Sage Publications, Inc.

Cohen, P., & Howe, A. (1988). How evaluation guides AI research. *AI Magazine, 9*(4), pp. 35-43.

Collins, A. (1987). *Reformulating testing to measure thinking and learning* (Report No. 6869). Cambridge, MA: BBN Systems and Technologies Corporation.

Feifer, R.G. (1989). *An intelligent tutoring system for graphic mapping strategies* (Technical Report UCLA-AI-89-04). Los Angeles: UCLA Computer Science Department.

Glennan, T. K., Jr. (1967). Issues in the choice of development policies. In T. Manschak, T. K. Glennan, Jr., & R. Summers (Eds.), *Strategies for research and development.* New York: Springer-Velaz.

Hayes, J.R. (1989). The complete problem solver (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Johnson. W.L., & Soloway, E. (1987). PROUST: An automatic debugger for Pascal programs. In G.P. Kearsely (Ed.), *Artificial intelligence: Applications and methodology.* Redding, MA: Addison-Wesley.

Leifer, A. (1976). Psychology of new teaching methods. In 1977 NSSE Yearbook. Washington, DC: NSSE.

Lesgold, A., Bonar, J., & Ivill, J. (1987). *Toward intelligent systems for testing* (LRDC Technical Report ONR/LSP-1). Pittsburgh: University of Pittsburgh Learning Research and Development Center.

Levin, H., Leitner, D., & Meister, G. (1986). Cost effectiveness of alternative approaches to computer assisted instruction (CERAS Report). Stanford, CA: Stanford University, CERAS.

Marshall, S.P. (in press). Mathematics: What cognitive skills do parents offer children? In T. Sticht (Ed.), The intergenerational transfer of cognitive skills. Norwood, NJ: Ablex.

Meriam-Webster Inc. (1989). Webster's ninth new collegiate dictionary. Springfield, MA: Author

Means, B., & Gott, S.P. (1988). Cognitive task analysis as a basis for tutor development: Articulating abstract knowledge representations. In J. Psotka, L.D. Massey, & S.A. Mutter (Eds.), *Intelligent tutoring systems: Lessons learned.* Hillsdale, NJ: Lawrence Erlbaum.

National Science Board. (1987). Science and engineering indicators. Washington, DC: Author

O'Neil, H.F., Jr., & Nizamuddin, K.G. (1989, July). *Effect of anxiety in intelligent computer-assisted instruction.* Paper presented at the 10th International Conference of the Society for Test Anxiety Research, Amsterdam, The Netherlands.

United States Department of Education Office of Educational Research and Improvement. (1988). *Creating responsible and responsive accountability systems: Report of the OERI State Accountability Study Group.* Washington, DC: Author.

United States Air Force (1986). *Project Forecast II: Executive summary.* Washington, DC: Author.

Wittrock, M.C., & Baker, E.L. (Eds.) (1989). *Testing and cognition.* Englewood Cliffs, NJ: Prentice-Hall.