AFHRL-TP-89-79

AIR FORCE

HUMAN RESOURCES

LABORATORY

AD-A222 090

INTELLIGENT TUTORING SYSTEMS: A TAXONOMY OF EVALUATION ISSUES

Kurt Steuck
J. L. Fleming

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

May 1990
Interim Technical Paper for Period August 1988 - December 1989

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

HENDRICK W. RUCK, Technical Advisor
Training Systems Division

RODGER D. BALLENTINE, Colonel, USAF
Chief, Training Systems Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 1990	Interim - August 1988 - December 1989

4. TITLE AND SUBTITLE
Intelligent Tutoring Systems: A Taxonomy of Evaluation Issues

5. FUNDING NUMBERS
PE - 62205F
PR - 1121
TA - 09
WU - 29

6. AUTHOR(S)
Kurt Steuck
J. L. Fleming

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Training Systems Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

8. PERFORMING ORGANIZATION REPORT NUMBER
AFHRL-TP-89-79

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This paper posits a taxonomy for categorizing issues that arise in the evaluation of Intelligent Tutoring Systems (ITSs). The taxonomy has three dimensions: Life Cycle of Evaluation, Research Issues, and Methodological Issues. The Life Cycle dimension has four levels: pre-experimental, laboratory study, field study, and initial operational test and evaluation. The three levels of the Research Issues dimension--functionality, effectiveness, and cost--are subsequently further divided into several sublevels. The Methodological Issues dimension is discussed in the context of each of the Research Issues levels. A recommendation from this work is that ITS evaluation studies should adopt multi-dimensional, multi-method designs.

14. SUBJECT TERMS
computer-based training
intelligent tutoring systems
training evaluation

15. NUMBER OF PAGES
20

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

AFHRL Technical Paper 89-79 May 1990

INTELLIGENT TUTORING SYSTEMS: A TAXONOMY OF EVALUATION ISSUES

Kurt Steuck
J. L. Fleming

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed and submitted for publication by

Hendrick W. Ruck, Technical Advisor
Training Systems Division

SUMMARY

This paper posits a taxonomy for categorizing issues that arise in the evaluation of Intelligent Tutoring Systems (ITSs). The taxonomy has three dimensions: Life Cycle of Evaluation, Research Issues, and Methodological Issues. The Life Cycle dimension has four levels: pre-experimental, laboratory study, field study, and initial operational test and evaluation. The three levels of the Research Issues dimension--functionality, effectiveness, and cost--are subsequently further divided into several sublevels. The Methodological Issues dimension is discussed in the context of each of the Research Issues levels. A recommendation from this work is that ITS evaluation studies should adopt multi-dimensional, multi-method designs.

PREFACE

The mission of the Intelligent Systems Branch of the Training Systems Division of the Air Force Human Resources Laboratory (AFHRL/IDI) is to design, develop, and evaluate the application of artificial intelligence (AI) technologies to computer-assisted training systems. The current effort was undertaken as part of IDI's research on intelligent tutoring systems (ITSs), ITS development tools, and intelligent computer-assisted training testbeds. The work was accomplished under Work Unit 1121-09-29, Intelligent Training Worlds.

An earlier version of this paper was presented at the Annual Meeting of the American Educational Research Association in San Francisco, CA, 31 March 1989. We would like to thank Drs. Joe Psotka (Army Research Institute) and Valerie Shute (AFHRL/MOE) for their comments on that paper.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTELLIGENT TUTORING SYSTEMS:
A TAXONOMY OF EVALUATION ISSUES

I. INTRODUCTION

Evaluation is the process of applying 'scientific procedures' to collect 'reliable and valid information' to make 'decisions' about an 'educational program' (Berk, 1981, p. 4).

The goal of this paper is to list evaluation issues in an applied setting and to propose a taxonomy which structures those issues. With the recent advances in the field of Intelligent Tutoring Systems (ITSs), it is time to compile the issues that are important to the products of those advances. Most issues described in this paper are substantive or "research" in nature.

In our particular setting, three groups can be involved in the design, development, evaluation, and implementation of an ITS (or tools created and used in the development of an ITS)(see Figure 1). These are a contractor, our research laboratory, and the Air Force (AF) training community. In the building of an ITS, a contractor (many times a university) is responsible for the design and development or coding of the ITS according to the specifications set by laboratory personnel. Upon completion, the contractor delivers or transitions the completed product to the research laboratory. Laboratory personnel are then responsible for evaluating and demonstrating the ITS to the AF training community. The latter must decide whether or not to support the continued development and ultimately the operationalization of the ITS.

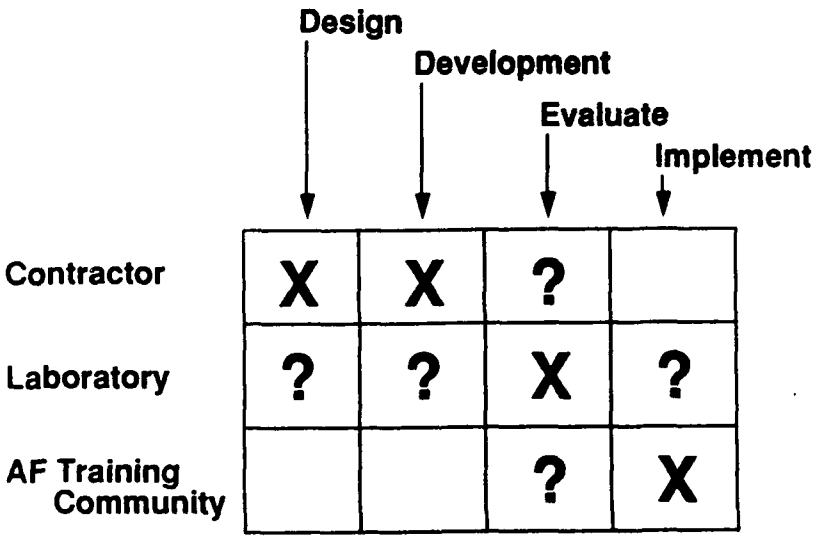

Figure 1. **ITS Development Roles.**

Note: X denotes typical participation
? denotes occasional participation

While the above scenario is generally true, the roles of the three agents vary from project to project. In some situations the laboratory and the contractor share the responsibility for the design and development of the ITS. In addition, the contractor and the AF training community may be involved participate in the evaluation of the tutoring system. Nonetheless, in any of the variations of roles, the contractor delivers the ITS to our research laboratory and then to the AF training community.

One important characteristic of this setting is that the developers, evaluators, and decision-makers may not be the same agents throughout the development and implementation of a tutoring system. In many cases, the contractor plays the role of the developer, the research laboratory personnel are the evaluators and the AF training community makes the decisions concerning adoption of the training systems being developed. This separation requires a more extensive evaluation methodology than when one agency plays all three roles.

Consequently, the evaluators must not only determine the effectiveness of the system (i.e., do the learners actually learn something), but also must evaluate the system on several other dimensions. For instance, the evaluator must assess whether the tutor meets design requirements. This may include assessing the functionality of the system (e.g., does the tutor perform in the manner specified in the design documents), its effectiveness (e.g., do learners learn?), and its efficiency (i.e., is it cost beneficial?).

The evaluation of the tutor must be complete enough for decision-makers to determine its value and relevance to their needs. Demonstrating that students learn within the tutoring environment is not adequate for the AF training community to support advanced development and implementation of the tutor. The evaluators must be able to show that the tutor increases on-the-job performance and is cost efficient. AF training personnel, while interested in gains in performance on the tutor, are much more interested in data that clearly show an improvement in job performance, the ability to learn on the job, or increased motivation to learn more on the job. Simple learning or performance gains within the tutoring environment in a laboratory setting, while necessary, are not sufficient enough for training decision-makers.

As a result of this setting in which the developers, evaluators, and decision-makers are not the same individuals, the nature of evaluation of the system must be more comprehensive than traditional evaluations that address the single, large question, "Is the ITS effective?"

II. TAXONOMY OF ITS EVALUATION

We are proposing a taxonomy for the evaluation of ITSs that has three inter-related dimensions: Phase or Life Cycle of Evaluation, Research Issues, and Methodological Issues (see Figure 2). The Life Cycle dimension refers to a sequence of evaluation from initial pre-experimental studies, to laboratory and field studies, and finally to research on the implementation in the actual training setting. The second dimension, Research Issues, covers the wide range of substantive issues researchers address concerning a system's functionality, effectiveness, and cost. The final dimension, Methodological Issues, includes issues that must be addressed in planning and conducting evaluation studies. Each of these dimensions are elaborated below.

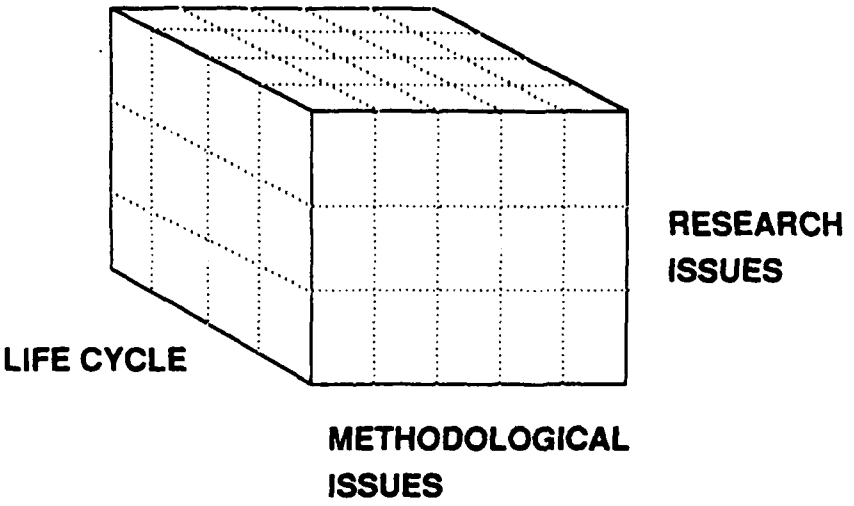

Figure 2. **Taxonomy of Evaluation Issues.**

The phase or life cycle of evaluation consists of levels which cover the flow of evaluation studies from pre-experimental evaluations through a series of studies in which experimental control decreases while operational qualities increase (see Figure 3). At the earliest point of an evaluation cycle, pre-experimental studies determine characteristics of the software, changes in student knowledge and skills at a detailed level, and developmental costs using no or few subjects (i.e., pilot subjects). In some cases, subject matter experts (SMEs) review the software for its accuracy and functionality. Laboratory and field studies determine issues such as the instructional effectiveness of the ITS at a larger scale than the pre-experimental study. While laboratory studies allow a high degree of experimental control of extraneous variables, they may not provide

much information about the application of the ITS in a realistic work place context. Field studies may provide the latter, but they sacrifice experimental control of independent and extraneous variables. These studies also address issues concerning functionality and costs. The final phase is the Initial Operational Test and Evaluation (IOT&E) of the ITS. In this type of study, the system under investigation is actually implemented for a period of time (e.g., several months) in the same manner intended when it is put in an operational environment. This study collects the data used to finally decide on changes to the system before continuing with its use in an operational setting.

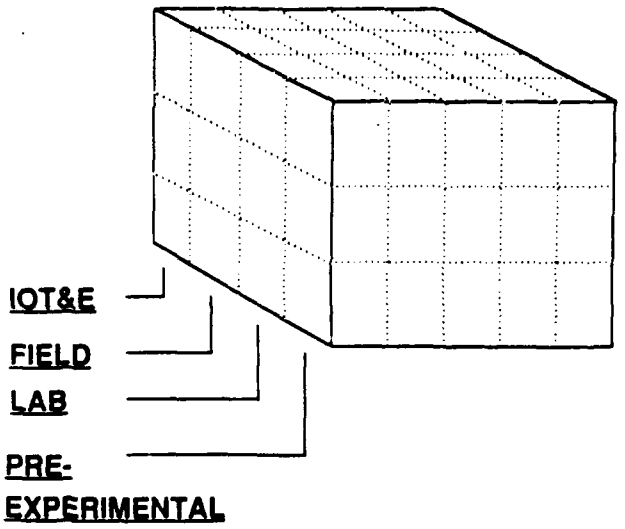

Figure 3. **Life Cycle Dimension.**

The second and third dimensions of the taxonomy, Research Issues and Methodological Issues, directly affect the nature of the evaluation. The second dimension, research issues, consists of the three broad categories of ITS assessment: functionality, effectiveness, and cost (see Figure 4). Standard experimental textbooks address the methodological issues: nature of the subjects, design of the study, independent and dependent variables, instruments, and procedures. The main focus of the remaining portion of this paper is primarily on the Research Issues dimension and secondarily on the Methodological Issues dimension.

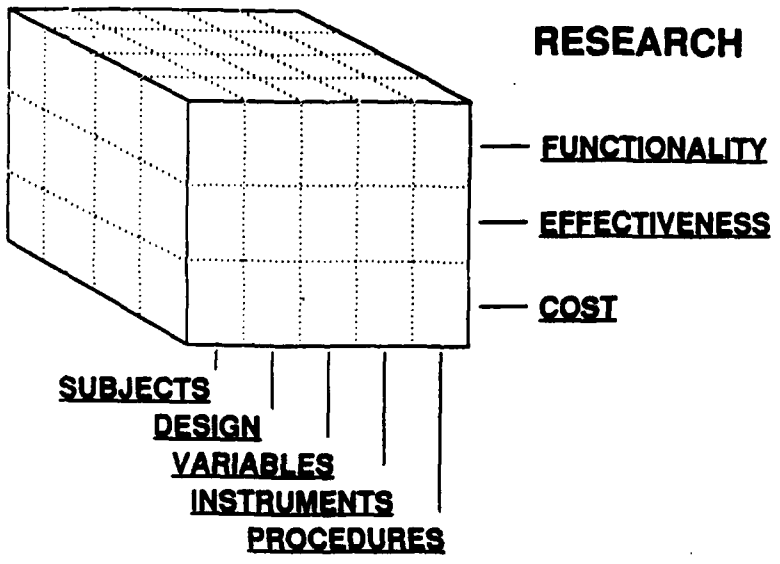

Figure 4. **Research and Methodological Issues.**

attention to the nature of the subjects used in a pre-experimental study of an ITS's functional capabilities while paying more attention to the instruments and dependent variables.

III. CATEGORIES OF RESEARCH ISSUES

In our applied setting, we raise three categories of issues: functionality (What does the ITS do? Does the ITS do what it should do?), effectiveness (Is the ITS effective?), and cost/benefit (Is the ITS efficient?). To explore each research issue within these broad categories methodological issues concerning (a) design of the evaluation study, (b) types of data to collect, and (c) methods of data collection must be addressed. The following sections of the paper will briefly describe several further research issues within each of these categories.

Functionality Issues

Completeness of Code

The issue here is whether the developer delivered all source code, compiled code, and associated programming documentation. This is necessary to ensure that the tutor is executable, designed appropriately, well-documented, and modifiable or extendible for further development. As with all software, bugs should be expected and the only way to fix them is with acces to the source code and proper documentation. This is not always trivial.

Requirements

This issue is whether the developed system meets two types of requirements set forth at the beginning of the ITS development. One is functional specifications and the other is performance. Functional specifications describe how the tutor should behave. They are described at the beginning of the research effort and can include requirements for, among other characteristics, the system's human-computer interaction, instructional approach, and hardware. Evaluation of this type, then, involves comparing the design of the ITS to the functional specifications prescribed prior before development. The performance issue is whether or not the implemented system as a whole meets those requirements. For instance, the functional specifications and design may include a context-sensitive help facility. Performance evaluation would assess the degree to which a context-sensitive help facility was indeed implemented. This issue must be addressed for contract evaluation, in prototyping new systems, and for subsequent enhancement of the system.

Relation to Taxonomy of Learning Environments

ITSs vary greatly in how they interact with students, the structure of the curriculum, the types of knowledge students will be learning, and so on. Kyllonen and Shute (1988) proposed a taxonomy of learning environments for describing and classifying ITSs. The four proposed dimensions are: knowledge type, instructional environment, domain, and learning style. The knowledge type dimension includes declarative knowledge (knowing that), procedural knowledge (knowing how to perform a task), and mental model (knowledge of the causal relations within a domain). The instructional environment dimension is classifies the instructional approach embodied in the ITS. Examples include Learning by Analogy, Learning from Instruction, and Inductive Learning. Domain, the third dimension described by Kyllonen and Shute (1988), represents dimensions underlying domain-specific learning. Domains vary in the degree to which technical, quantitative, qualitative, and verbal ability play a role in competent performance. The final dimension of the learning taxonomy is Learning Style. It covers the learner's characteristics that influence instructional activities and in turn can be modified through instruction. While we direct the reader to Kyllonen and Shute (1988) for further details, suffice it to say that a taxonomy such as the one described here would have

activities and in turn can be modified through instruction. While we direct the reader to Kyllonen and Shute (1988) for further details, suffice it to say that a taxonomy such as the one described here would have implications for evaluations of ITSs.

Adopting a taxonomy and evaluating a tutor relative to it has both scientific and practical benefits (Kyllonen & Shute, 1988). On the practical side, a taxonomy might provide information to decision makers about the applicability of a specific tutor to the problem at hand. A taxonomy might also suggest ways to describe the nature of the material taught in the domain (e.g., predominantly procedural). A third way a taxonomy might be helpful is in specifying an appropriate instructional approach given the domain. To realize these benefits, evaluators must plan and conduct studies which collect data pertinent to the dimensions and their levels of the adopted taxonomy.

Furthermore, as the ITS field grows and matures, review (e.g., meta-analysis) studies will compare results and synthesize conclusions. These reviews will be highly dependent on the completeness of and accuracy in the descriptions of the student samples, educational treatments, data collection procedures and so on (Abrami, Cohen, & d'Apollonia, 1988). Adopting a taxonomy will then facilitate the accuracy and subsequent usefulness of those reports not only for decision makers in the training community, but also for the scientific community.

The scientific community would also benefit from a learning environment taxonomy in forming research hypotheses and guiding evaluative research. Research questions could be raised about each cell of the proposed taxonomy. Questions such as: "Is an expository approach appropriate for students with strong perceptions of self-competence in the domain of air traffic control?" Furthermore, a taxonomy could generate research questions concerned with the efficacy of different design approaches of ITS components (see below) given the cells of the taxonomy. For instance, how rich or intensive does the student modelling approach need to be given a particular instructional approach, domain, and type of knowledge taught?

ITS Components

Several research issues surround the evaluation of the functionality and design of ITS components. It is not sufficient enough to only ask if the ITS as a whole is effective or whether ITS components are effective. We must also address whether ITS components function in accordance with design specifications or human performance in the real-world setting. Each component and a few issues are presented.

One component of an ITS is the Expert Model. It represents the domain related knowledge that experts possess about tasks that are performed, problem solving techniques and strategies, equipment, and experts' reasoning (Anderson, 1988). An important question that must be addressed is whether the representation of the knowledge and reasoning skills are appropriate, accurate, and complete given the domain. Relatedly, how to verify the veracity of the representation is also critical (a methodological issue).

Another component of an ITS is the Student Model. It represents the characteristics of each student and is dynamically updated based on a student's performance during the course of instruction (VanLehn, 1988). Questions addressed here include: Is the representation of the student employed in the student model detailed and complete enough for capturing a student's strengths and weaknesses relative to the domain? How intensive does the student model need to be given the nature of the domain? How should we evaluate the appropriateness of the representation of the student in the student model? Its evaluation is critical because it plays an important role in diagnosing a student's needs and individualizing instruction.

A third component of an ITS is the Instructional Model. It is responsible for comparing a student's performance relative to expert performance, developing an instructional plan based on the student's needs and abilities, and delivering that instruction (Halff, 1988). This capacity to dynamically adapt instruction to the individual is one of the greatest advantages of ITSs over more traditional computer-based training (CBT) or classroom training where the student-teacher ratio is high. Evaluation studies must be able to assess to what degree the system under evaluation actually accomplishes this. ITSs have the potential to individualize

instruction through context-dependent explanations, remediate after failure, coach a floundering student, present instruction based on the student's learning style, monitor the amount of time remaining for the lesson, and respond to student requests. Evaluation studies need to further explicate the issues centering around individuallzation and how to assess the extent to which an ITS individualizes instruction.

Studies should also evaluate the relationship of the instructional approach embodied in the tutoring system to theories and principles of learning and instruction. ITSs intentionally or by default adopt a particular instructional approach or approaches in tutoring students. They also vary in the degree to which they adopt those teaching strategies or techniques. Evaluation studies could then determine whether an ITS follows a well-founded instructional theory and to what extent. Furthermore, this kind of analysis could provide information for improving instructional theory and practice underlying ITSs.

Evaluation studies of the fourth component of an ITS, the interface, have traditionally addressed issues of user acceptance. In an actual training environment, the "user" could be the student or the instructor/administrator. Studies of student acceptance have investigated the computer-to-student flow of information (e.g., the student's ability to understand directions and explanations) and the student-to-computer flow (e.g., menuing). For example, Williams, Hamel, and Shrestha (1987) have constructed a checklist for evaluating computer-assisted instruction (CAI) interfaces.

Interfaces also must be acceptable to instructors/ administrators. Evaluation studies should determine if instructors consider the tutor (a) easy to use, (b) easy to learn, and (c) easy to teach to students. If tools, such as authoring or management, are available, evaluators must also determine how easy each is to learn, understand, and use.

Evaluating an interface must go beyond assessing traditional "acceptance." Frye, Littman, and Soloway (1988) found that inexperienced users (in this case children) had more problems operating the programs than older, more experienced children. Difficulties in using the programs reduced the students' access to the educational content, thereby reducing the overall instructional effectiveness of the tutor. Frye, Littman, and Soloway (1988) pointed out that not only was contact with the instructional content reduced, but also that the interface directly interfered with students' understanding of the content. This example points to the need for evaluation of interfaces beyond that of surveys or ratings of user acceptance.

Instructional Context

Implementing an ITS in an actual classroom context can have profound impacts on that context. Not only may student and teacher roles change, but also the student-teacher interaction may change (Zimmerman, Smith, Bastone, & Friend, 1989). In addition to adjusting to changing roles, the teacher must be able to integrate the ITS into the existing curriculum and daily and weekly schedules. There may also be physical characteristics of the instructional context that must be taken into account, such as hardware and the arrangement of the room. It is especially important to measure these potential changes in the instructional context during the infancy of ITS implementations. Early findings could lead to subsequent research addressing instructional context characteristics that facilitate or hinder the final implementation of tutoring systems.

Methodological Issues

After deciding what features of an ITS to evaluate, researchers must address several methodological issues. One set involves the design of the study. Some research questions lend themselves to experimental comparison; others require interviews with domain experts. The design of the study is not only affected by the research question, but also limitations in resources allocated to the evaluation. For instance, since domain experts are scarce and in large demand, it is not feasible to have 20 domain experts review the representation of expert reasoning.

Data Collection. Another set of methodological issues involves the type of data to collect. Steinberg (1984) gives an excellent enumeration of data important to this issue of functionality. In it she states they should revolve around the accuracy and completeness of the content, an expert instructor's opinion of the method of presentation, technical flaws, flow of the lesson, time required to complete a session, and students' attitude toward the tutor. She recommends keeping a computer file of students' keystrokes for analysis of such questions as:

What proportion of keystrokes or clicks are erroneous?

How long do students spend on each part of the tutor?

How many times do students press the help key?

What is the number and nature of unanticipated keystrokes?

Another computer file she recommends is one that allows students to enter immediately comments or recommendations about the tutor.

Data collected by unobtrusive observers is also significant. During student trials, they can determine such information as the keyboard/mouse manipulation requirements imposed by the tutor, the readability of screens, clarity of instructions and the tutor's technical correctness.

Steinberg (1984) also recommends observers interview tutored students to collect valuable feedback data. In these, students are asked their overall opinion, what they consider the best and worst parts, recommendations, and clarification of any notes the observer makes during the session. Furthermore, Kyllonen and Shute (1988) list 29 indicators of student's progress in an ITS relating to activity level and exploratory behaviors, data recording, use of embedded tools, effective generalizations of principles, and effective experimental behaviors.

Instruments. Issues concerning how to collect the data are directly tied to decisions concerning what data to collect. Instruments include, but are not limited to, the ITS itself, verbal protocols, video taping, checklists, ratings scales, technical analysis of code, and interviews with students, experts, instructors, and administrators.

Effectiveness Issues

The most important information that researchers can collect, summarize, and present to decision-makers concerns the effectiveness of the tutoring system. Other issues, such as the functionality of ITS components, are not of consequence if evaluations of a tutor show that the ITS is not effective in producing student gains. Although evaluating functionality is relatively straight forward, determining the effectiveness of a tutoring system is not. Decisions must be made about the design of the study (e.g., experimental, longitudinal), what constitutes an appropriate control group, controlling or measuring access to the tutoring environment, measuring access to the curriculum, measuring the effects on student performance and motivation to continue learning and performing.

Access to Learning Environment

Since one goal of an evaluation may be to determine what ITS component or components affect student performance, access to the learning environment must be controlled or measured and analyzed. This type of access can be thought of in terms of total time allocated for training or in terms of the quality of that time (e.g., uninterrupted blocks of time). If the ITS group receives more learning opportunities than the control group (e.g., On-the-Job Training (OJT)), then the comparison is not one of effectiveness of an ITS compared

Access to the Curriculum

Training systems, in general, vary in the representation of domain-relevant information, such as heuristics, algorithms, concepts, and devices. Training systems also vary in the ways students access that information. For example, in traditional classrooms students access the curriculum through books and lectures. ITSs now provide unique ways for students to come into contact with the domain due to their ability to present information in time-compressed methods and through their ability to represent the knowledge and skills from several individuals who could not be present for training purposes (e.g., domain experts). Evaluation studies should address access to the curriculum as part of a training system.

ITSs can present more of a curriculum to each student due to its ability to deliver time-condensed training. An example is in a microworld named Orbital Mechanics (OM). The goal of OM is to develop in the student an understanding of the relationship between several numerical parameters and the visualization of the ground trace of a satellite. It takes a student about 2 to 5 hours to perform all the equations underlying the orbit by hand. In OM, it takes about 5 to 10 seconds, because the equations are embedded in the microworld. Thus, a student can "access" more of the curriculum due to the time compressed delivery capabilities.

Another way in which access to the curriculum may be increased is through the representation of an expert in the ITS. In the work place environment, experts are not plentiful and do not have time to train novices how to solve domain problems. By embedding expert knowledge and reasoning in the ITS, more novices can have access to expert thinking. As a result, more students can be trained without allocating expert resources to the training process beyond initial development of the tutor. This increased access to expert reasoning, and hence the curriculum, could account for the differences between ITS applications and alternate training approaches.

One advantage of measuring access to the curriculum is that it gives decision makers additional information about the potential of ITSs in general. The empirical demonstration that ITSs can deliver more instruction in the same amount of time or the same instruction in less time provides additional important information about the benefits of adopting an ITS approach.

Learning Indicators

By far, the single most significant finding of all evaluation is whether or not ITSs increase a student's knowledge, skills, and strategies in the target domain. Studies designed to answer this "grand" question must, therefore, gather valid indicators of learning. These indicators can be collected prior to a student entering instruction, during instruction, and after instruction. Not only can we collect data on changes in domain related knowledge, skills, and strategies, but also in other more subtle indicators revealed in the dynamics of the student's interaction with the learning environment. For instance, changes in the pattern of student's help requests may indicate growing cognitive structures. Other indicators include such measures as latencies in interacting with the learning environment, menu selection, and responses to tutor advice or directives. Kyllonen and Shute (1988) described the impact of an ITS on students' learning by collecting data in 29 learning indicators within three broad groups--activity and exploratory behaviors, data management skills, and thinking and planning skills.

Posttest data is the primary data of interest when one is evaluating an ITS as a whole. This can be done by comparing it to some other educational system (e.g., traditional education) or by assessing changes in the level of knowledge, skills, and strategies of individual students. However, posttest and concurrent data, which can be collected unobtrusively, can be used when assessing ITS component effectiveness. This latter form of evaluation can be done by comparing the ITS in question to another computerized instructional system or by analyzing changes in students' performance profiles.

Job Performance

One issue that training system decision makers want addressed is the training system's ability to improve on-the-job learning and performance. While researchers may get excited about gains in student performance within a tutoring environment, the operational training community needs evidence that ITSs improve actual job performance. This requires that data outside of the ITS environment (e.g., troubleshooting ability) be collected. The requirement to gauge the impact of training with ITSs on job performance is essentially one of measuring transfer of training. Cormier and Hagman (1987) give an excellent treatment of this issue. In Chapter Nine, "Measuring Transfer in Military Settings," Boldovici (1987) notes training with devices can have positive, negative and neutral effects on job performance.

The usual experimental design for measuring transfer is to first train two groups with different methods and then measure differences in their job performance. To reduce sources of error frequently encountered in this design, Boldovici suggests one in which job performance of three groups is measured after three training intervals. The groups are: (a) the device group which receives training with the device (ITS in our case), (b) the conventional group which receives conventional training without the ITS, and (c) the control group which receives no training. Tests are given to each group at three equal intervals during the period that is usual for conventional training. Table 1 depicts this design.

Table 1. **Schedule for a Transfer Experiment**

	Weeks						
	1,2,3,4			5,6,7,8		9,10,11,12	
Device group	Test Task B	Train Task A	Test Task B	Train Task A	Test Task B	Train Task A	Test Task B
Conventional group	Test Task B	Train Task A	Test Task B	Train Task A	Test Task B	Train Task A	Test Task B
Control group	Test Task B		Test Task B		Test Task B		Test Task B

Boldovici's design has several advantages. It separates the amount of training from the effects of training media by making it an investigated effect. It also allows for the inspection of reliability of Task B measurements. Furthermore, causal linkages between learning Task A and performing Task B can be made.

Motivation to Learn and Perform

Not only do new training systems affect cognitive variables (e.g., domain knowledge), but they influence students' motivation and attitudes. According to Bandura (1982) and Schunk (1984), positive experiences in a learning situation lead to the development of positive self-efficacy. This in turn leads to increases in willingness to learn more, willingness to take risks, and willingness to persist in the face of failure. Given this perspective, studies of ITSs should evaluate the effects on motivation not only to perform, but also tc learn more.

ITS Components

In early stages of development of ITS technology evaluative studies should assess the effectiveness of various approaches to ITS components. ITSs have the ability to record data about each student's activities and performance, the instructional events that occur, and the relationship between the two. ITSs are data-rich environments--they can assess not only the effectiveness of components in isolation, but also in complex interactions. For instance, in one domain it may not be necessary to have an elaborate, intensive model of student knowledge and skills, but in a different domain a detailed representation of the student's abilities, misconceptions, and performance may be required for the ITS to be effective.

Methodological Issues

As described under the section on functionality issues, evaluators of ITSs must address methodological issues concerning subjects, design, variables, instruments, and procedures before conducting evaluative studies. Discussion of three important methodological issues follow.

Comparison group. The choice of a comparison group is tied directly to the specific question addressed. If the goal is to make conclusions about a tutoring system's effectiveness relative to an extant educational system, then that extant group can serve as the comparison group. In contrast, if the goal is to determine what components or functions make a tutor effective then the extant group can serve as a comparison group only if controls are placed in the extant learning environment that limit the influence of extraneous variables. For instance, if the goal is to determine whether an ITS is effective due to the individualization of instruction, then control of other variables, such as the quantity and quality of the curriculum, must occur to guarantee the equivalence of the two groups. Table 2 presents the dissimilarities of three training environments in the Air Force. Because of the vast differences, extant educational systems should be used for comparison only when the goal is to show differences in educational systems at the system level. Other approaches are needed, such as monitoring changes in the student model as a result of instructional events, if the goal of the evaluation is to determine the effectiveness of specific ITS component approaches.

Table 2. **Stylized Description of Three Educational Systems**

Educational system	Curriculum	Instructional materials	Agent of delivery
Technical School	Structured	Texts, Notes, Lectures	Instructor
On-the-Job Training	Incomplete, Fragmented	Manuals, Actual Equipment	Expert
ITS	Structured	Problems Text	Interface Module

Instruments. Several techniques have been used to collect data in the evaluation of instructional systems. The most prominent is to have student performance data collected by the computer or via external measures such as paper and pencil. Others have used measures which reflect actual job related performance, verbal protocols (both concurrent and retrospective), interviews, surveys, audio and video recordings, and direct observations.

Cost Issues

The third level under the Research Issues dimension of our proposed taxonomy of evaluation is cost. To evaluate fully an ITS for potential implementation, data needs to be collected not only on the cost of development, but also on cost of evaluation of the ITS, initial implementation in the operational environment, and maintenance and updating once the ITS is operational. Development costs include the time and dollars spent on and by knowledge engineers, subject-matter experts, instructional developers, and computer programmers. Evaluation costs are not trivial when dealing in an applied setting and might easily be overlooked. Evaluation costs could include travel expenses, subject-matter expert time, student time, and evaluator time. Decision makers in the training community need estimates of implementation costs for accurate planning and budgeting. Implementation costs include those related to course instructor training, hardware requirements, and software needs (e.g., knowledge engineering and authoring tools). Decision makers also need to know potential maintenance costs for the hardware and software once the instructional system is in place.

IV. CONCLUSIONS AND RECOMMENDATIONS

As with much in the worlds of education and computers, developers are hurriedly building ITSs for real world applications. We see this as an expensive but positive step. To reduce costs and facilitate the proliferation of this technology, researchers must increase discoveries about the effectiveness of ITSs and use those findings to produce better tutors. This can best be done with an organized approach by the research community to evaluate these systems by constructing and applying experimental paradigms which address the issues mentioned in this paper.

We offer several recommendations:

1. Adopt a taxonomy of learning environments for an efficient, comprehensive description of ITS functionality.

2. Adopt multi-method evaluation methodologies.

3. Describe evaluation studies fully.

4. Find out in as much detail as possible what potential users of ITSs require--get to know the users in more than a clinical sense.

5. For evaluating effectiveness, especially for simulation based ITSs, consider the experimental design proposed by Boldovici.

6. Create a taxonomy of effective designs of ITS components for different domains.

REFERENCES

Abrami, P.C., Cohen, P.A., & d'Apollonia, S. (1988). Implementation problems in meta-analysis. *Review of Educational Research, 58* (2), 151-179.

Anderson, J.R. (1988). The expert model. In M.C. Polson & J.J. Richardson (Eds.). *Foundations of Intelligent Tutoring Systems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist, 37*, 122-147.

Berk, R.A. (1981). Introduction. In R.A. Berk (Ed.). *Educational evaluation methodology: The state of the art.* Baltimore, MD: Johns Hopkins University Press.

Boldovici, J.A. (1987). Measuring transfer in military settings. In S.M. Cormier & J.D. Hagman (Eds.). *Transfer of learning*. San Diego, CA: Academic Press, Inc.

Cormier, S.M., & Hagman, J.D. (1987). *Transfer of Learning*. San Diego, CA: Academic Press, Inc.

Frye, D., Littman, D.C., & Soloway, E. (1988). The next wave of problems in ITSs: Confronting the "user issues" of interface design and system evaluation. In J. Psotka, L.D. Massey, & S.A. Mutter (Eds.). *Intelligent Tutoring Systems: Lessons learned*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Halff, H.M. (1988). Curriculum and instruction in automated tutors. In M.C. Polson & J.J. Richardson (Eds.). *Foundations of Intelligent Tutoring Systems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kyllonen, P.C., & Shute, V.J. (1988). Taxonomy of learning skills (AFHRL-TP-87-39, AD-A190 669). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Schunk, D.H. (1984). Self-efficacy perspective on achievement behavior. *Educational Psychologist, 19*, 48-58.

Steinberg, E.R. (1984). *Teaching computers to teach*. Hillsdale, NJ: Lawrence Erlbaum Associates.

VanLehn, K. (1988). Student Modeling. In M.C. Polson & J.J. Richardson (Eds.). *Foundations of Intelligent Tutoring Systems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Williams, K.E., Hamel, C.J., & Shrestha, L.B. (1987). *CAI evaluation handbook: Guidelines for user interface design for computer-assisted instruction* (NTSC-TR-87-033). Orlando, FL: Naval Training System Center.

Willson, V.L. (1988). Evaluation of learning strategies research methods and techniques. In C.E. Weinstein, E.T. Goetz, & P.A. Alexander (Eds.). *Learning and study strategies: Issues in assessment, instruction, and evaluation* (pp. 263-274). San Diego, CA: Academic Press.

Zimmerman, B.J., Smith, C.P., Bastone, L., & Friend, R. (March, 1989). *Social processes in microcomputer learning: A social cognitive view*. Paper presented at the annual meeting of the American Educational Research Association in San Francisco, CA.