Jl. of Artificial Intelligence and Education (1993) 4(2/3), 245-271

Principles for Evaluating Intelligent Tutoring Systems

VALERIE J. SHUTE AND J. WESLEY REGIAN
Armstrong Laboratory, Human Resources Directorate
Brooks Air Force Base, TX 78235, USA

There are many and varied intelligent tutoring systems (ITS) described in the literature. Few have been subjected to rigorous, controlled evaluations. This paper addresses what is required to evaluate the efficacy of an ITS. Efficacy, in this paper, refers to assessing if the system teaches what it was intended to teach, to what degree, in comparison to what, and at what cost. Exhaustive and detailed coverage of this topic would not be possible in a short article. Instead, we have attempted to overview the topic and provide illustrations, as well as exemplar references, to direct the interested reader to more detailed coverage. We begin by describing our general approach to research and development of ITS. Then, seven principles are presented that are believed to underlie a good ITS evaluation study:(1) delineate the goals of the tutor, (2) define the goals of the evaluation study, (3) select the appropriate design to meet the defined goals, (4) instantiate the design with appropriate measures, number and type of subjects, and control groups, (5) Make careful logistical preparations for conducting the study, (6) Pilot test the tutor and other aspects of the study, and (7) Plan the primary data analysis as you plan the study. We use these principles as a framework for organizing, discussing, and comparing ITS evaluation studies.

For a long time, researchers have asserted that carefully-designed, individualized tutoring produces the most effective and efficient learning for many people (e.g., Bloom, 1956, 1984; Burton & Brown, 1982; Carroll, 1963; Cohen, Kulik, & Kulik, 1982; Lewis, McArthur, Stasz & Zmuidzinas, 1990; Sleeman & Brown, 1982; Wenger, 1987; Woolf, 1987). Automated instruction has long been seen as a potentially affordable approach to the delivery of individualized tutoring (e.g., Pressey, 1926, 1927; Skinner, 1957). As early as 1926, Pressey described a device which sought to apply contemporary learning theory to the task of automated instruction.

The device, loaded with multiple-choice questions and answers by the teacher, would drill the student on the questions and provide immediate feedback in order to support learning:

> The somewhat astounding way in which the functioning of the apparatus seems to fit in with the so-called 'laws of learning' deserves mention in this connection. The 'law of recency' operates to establish the correct answer in the mind of the subject, since it is always the last answer which is the right one. The 'law of frequency' also cooperates; by chance the right response tends to be made most often, since it is the only response by which the subject can go on to the next question. Further, with the addition of a simple attachment the apparatus will present the subject with a piece of candy or other reward upon his making any given score for which the experimenter may have set the device; that is the 'law of effect' also can be made, automatically, to aid in the establishing of the right answer (Pressey, 1926, p. 375).

Today, intelligent tutoring systems (ITS) epitomize this notion of theory-based, individualized, automated instruction. Unfortunately, although such systems have been in existence for over a decade, the degree to which they have been successful is equivocal because controlled evaluations of ITS are scarce (Baker, 1990; Littman & Soloway, 1988).

There have been some systematic, controlled evaluations of ITS reported in the literature. A few examples include: the LISP tutor (e.g., Anderson, Farrell, & Sauers, 1984) instructing Lisp programming skills; Smithtown (Shute & Glaser, 1990, 1991), a discovery world that teaches scientific inquiry skills in the context of microeconomics; Sherlock (Nichols, Pokorny, Jones, Gott, & Alley, in preparation; Lesgold, Lajoie, Bunzo & Eggan, 1992), a tutor for avionics troubleshooting; and Bridge (Shute, 1991; Shute & Kyllonen, 1990) teaching Pascal programming skills. Results from the evaluations show that these tutors do accelerate learning with, at the very least, no degradation in outcome performance compared to appropriate control groups.

What can we make of these findings? As always, there is a selection bias for publication of unambiguous evidence of successful instructional interventions. We are familiar with other (unpublished) tutor-evaluation studies that were conducted but were "failures." In some cases, we feel these studies may have provided more information given a better experimental design. Because of the interdisciplinary makeup of the ITS community, it is not surprising that among some groups there is a lack of interest in evaluation, or training in experimental design.

Before presenting the seven principles for evaluating intelligent tutoring systems, we briefly present a general approach to research and development of intelligent tutoring systems.

A GENERAL APPROACH TO RESEARCH AND DEVELOPMENT OF ITS

Experimental design involves arranging conditions to promote the validity of an experiment—the measure of effect produced by some independent variable on a dependent variable. If the causal link between independent manipulations and dependent measures is equivocal, the experiment is said to lack internal validity. If the ability to generalize from the experimental sample to the population of interest is equivocal, the experiment is said to lack external validity. In this paper, we argue that internal validity is easier to accomplish in an evaluation conducted within a controlled, laboratory setting (addressing more basic research questions), while external validity is easier to accomplish in an evaluation experiment conducted in "the field" or natural setting (more applied research questions). It is generally rare in research to find a study yielding results which possess high internal validity and high external validity, concurrently. This represents a classic tradeoff between internal and external validity. Figure 1 notionally depicts this relationship: as internal validity decreases, external validity increases. This relationship is particularly true with regard to research on pedagogy. Research in field settings (e.g., high school classrooms) is desirable because all aspects of the target setting are present in the experiment. Many of these aspects, however, are potential confounds to the experiment, making it difficult to relate outcome performance measures to the experimental manipulation. Thus, field research on pedagogy can have high external validity, but correspondingly low internal validity. On the other hand, research in laboratory settings is desirable because of the extreme experimental control that is possible in the laboratory. One can control for prior knowledge, assign subjects to groups, counterbalance for teacher (experimenter, proctor) effects, and so on. But can these results be generalized to others beyond the immediate sample?

Our approach to managing the trade-off between internal and external validity is to begin with laboratory research (high experimental control and internal validity) and slowly increase external validity, ultimately studying the intervention in the target instructional context (field research). We believe that neither laboratory nor field research alone will give a complete and accurate picture of the instructional effectiveness of a particular intervention. We further believe that research on pedagogy should be driven by

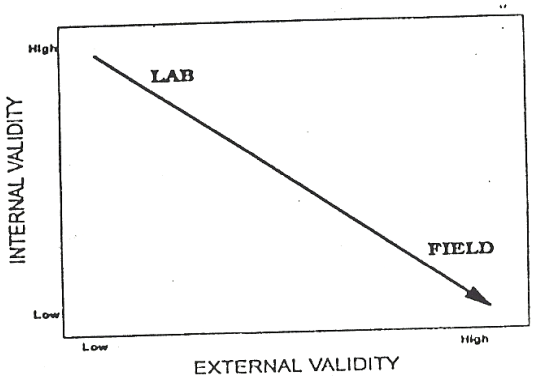

Figure 1. Simple inverse relationship between internal and external validity.

theory and constrained by empirical observation. By theory, we mean a coherent, plausible body of ideas about how people acquire, store, retrieve and apply knowledge and skill. Theory is important in generating potentially fruitful hypotheses about teaching and learning, and in driving generalizations about pedagogical effectiveness across instructional domains. Empirical observation is important to test, often and rigorously, how our ideas fare in the real world. Empirical data about the effectiveness of theory-based instruction provides feedback about how well our implementation works, and also may cause us to rethink our theory. Figure 2 shows this proposed cyclical relationship between theory and data, whereby research commences with theory and becomes progressively modified by empirical findings.

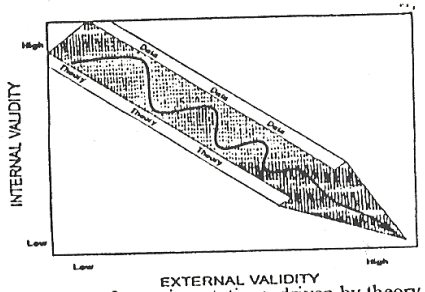

Figure 2. Cyclical process of experimentation: driven by theory and constrained by empirical findings.

SEVEN PRINCIPLES IN ITS EVALUATION

Keeping in mind the general approach to ITS research and development, at one or more points along the big arrow, in Figure 2, you will need to evaluate your ITS. However, the outcome of the evaluation occasionally reflects the goodness (or poorness) of an experimental design, rather than the efficacy of the ITS. In our experience, we have seen evaluation studies fail due to poor experimental design, inadequately operationalized constructs and measures, or deficient logistical planning and implementation. In the following sections, we present seven main principles that may be used to design, plan, and implement an effective ITS evaluation. These principles are:

(1) Delineate the goals of the tutor,
(2) Define the goals of the evaluation study,
(3) Select the appropriate design to meet the defined goals,
(4) Instantiate the design with appropriate measures, number and type of subjects, and control conditions,
(5) Make careful logistical preparations for conducting the study,
(6) Pilot test the tutor and the study, and
(7) Determine the primary data analyses as you plan the study.

Given the plethora of tutors in the world, the relatively few controlled evaluations, and the various pitfalls accompanying evaluation attempts, this paper will describe principles for evaluating ITS. The purpose of the paper is to provide a framework within which to organize, discuss, and compare evaluation studies, codify the process of designing and conducting a competent evaluation study, and maximize the utility of future studies concerned with evaluating ITS. Each of the principles will be addressed, in turn.

PRINCIPLE 1: DELINEATE THE GOALS OF THE ITS

It is a good strategy to review and carefully delineate the goals of the ITS prior to designing the evaluation study. In some cases, it may become apparent that the instructional goals have shifted (subtly or markedly) over the developmental life-cycle of the ITS. In any event, a clear understanding of the ITS is useful at this stage. If the designer of the evaluation study is not intimately familiar with the tutor, then designing a good evaluation is

virtually impossible. We believe the evaluation designer should be knowledgeable about the following critical issues.

What instructional approach underlies the tutor? This question addresses the general and specific instructional approach of the tutor. Is it a tutoring system with pedagogical intelligence, a coached-practice environment, or a more free-form discovery microworld? Is the system supposed to guide learning, or provide a rich environment for the induction of principles, or allow students to practice skills? How does the tutor diagnose performance and select instructional interventions? What information about the student is modeled and how is it used?

What learning theory does it assume? This is an important, but occasionally ignored, part of the ITS-building process. Is there a clear knowledge or skill-acquisition theory in the literature that motivates the instructional approach of the tutor? Interpreting findings from the evaluation of a theoretical, or intuition-based, tutors is difficult. Your ability to generalize findings across instructional domains, or even within domains, is dependent on a theoretical characterization of domain dimensions.

What exactly does it teach? Specific and measurable knowledge or skills should be expressed clearly as the desired learning outcomes. Some possible responses to this question include: The student should be able to apply a mental model of electricity concepts in solving a DC-circuit problem, execute procedural and psychomotor skills associated with an astronaut's job of manipulating a robot arm in a weightless environment, solve high school physics problems, or discover geometry concepts.

What other impacts is it expected to have? After delineating what the tutor is supposed to teach, you need to consider other ways that the tutor is expected to effect the student. In some cases, these effects may be the primary goal of the tutor. Otherwise, they may be considered subsidiary treatment effects. Examples of such effects include enhancing transfer-of-skills to another domain, influencing perceived self-efficacy, or modifying attitudes about computers.

In what context is it supposed to operate? You need to clearly specify the environment in which the tutor is intended to operate because evaluation techniques are differentially suitable for various learning environments. Questions to consider in this section include: Is the system intended to supplement a lecture or laboratory, or provide stand-alone instruction? Does the system teach to individuals or small groups? What prior knowledge is assumed of the students? Is the tutor supposed to be used in an academic setting to support declarative knowledge acquisition, or in an industrial training environment to support acquisition of procedural skills?

PRINCIPLE 2: DEFINE THE GOALS OF THE EVALUATION STUDY

Think carefully about the goals of the study. A clear understanding of what you want to find out provides the basis for selecting an experimental design that will unequivocally address your research goals. You should think about the difficulties involved in achieving the goals, and adjust your goals at the outset to ones that are realizable. Consider the following questions.

What would you like to know after the study is completed? You may simply want to see if the ITS results in students learning something beyond their incoming knowledge and skills. Or, you may want to test the degree to which your tutor affects students' capabilities to perform some task. Another possible question you could pose concerns how the tutor influences learning in relation to classroom instruction on the same material—a comparative study. A different comparative study might address how your tutor affects learning in relation to a colleague's tutor, instructing the same subject matter. This second kind of comparative study would be of interest if there were different design principles motivating the tutors, or different underlying learning theories. You should clearly specify your research questions and hypotheses before you conduct the study and analyze the data.

By what standard will you measure success? Suppose you want students to learn 12 problem-solving strategies. After working with your ITS, students can state 6 of these strategies, on average. Have you succeeded? This question is intractable without some standard of comparison. If you learn that a simple mnemonic allows students to state 11 of the 12 strategies, you may assume your ITS is relatively ineffective. Alternatively, suppose you find that students learning from your ITS can both state and reliably apply 6 of the 12 strategies, whereas students trained with the mnemonic approach can state, but not apply, 11 strategies. Would you now assume that your ITS is effective? This, of course, depends on your tutor's goals. After you've answered the question "What exactly does the ITS teach?" you need to identify ways to measure whatever is being taught (e.g., indices to assess the veracity of knowledge, or successful application of problem-solving skills) and to whom your students will be compared on these measures.

What are potential confounds, and which of these can you control? There are numerous ways in which unwanted influences can contaminate the results of your study. Pinpointing potential confounds before conducting the study makes it easier to control them (beforehand, by altering

the design, or afterwards, statistically). For example, suppose you are interested in testing your ITS in two separate schools. You suspect that the schools may differ in terms of one or more important dimensions (e.g., students' mean IQ, faculty training, per capita income, ethnicity). You may choose to create treatment and control groups at both schools, statistically control for these dimensions, select four schools and counterbalance the design, and so on. All of these are options if you identify the potential confounds, in advance.

Will you use quantitative indices, protocols, or observational data? These three types of data represent the most prevalent means of capturing what a student is learning from the ITS. Because your subjects are working at computers, it is particularly easy to capture on-line, quantitative measures of performance, such as latencies, accuracies, and counts as learning indicators. With some effort and creativity, protocol analyses can also yield a wealth of information about learning that cannot be captured directly by the computer (see, for example, Shute & Glaser, 1991). Trained observers may also ascertain aspects of learning and performance that are impossible to apprehend by other means (Schofield & Evans-Rhodes, 1989).

PRINCIPLE 3: SELECT AN APPROPRIATE DESIGN TO MEET DEFINED GOALS

After you have delineated the goals of the tutor and the goals of the evaluation study, you are ready to choose a design to test your research questions. First, are you conducting a formative or a summative evaluation of your system? The basic distinction between the two is that formative evaluations have an internal control condition, and ask the question: How can we make the system better? In comparison, summative evaluations have an external control condition, and ask the question: How does this system compare to some other systems or approaches? Examples of both types of evaluations are provided.

Formative Evaluations

Formative evaluations are conducted during the developmental phase of an ITS to find weaknesses early enough so that design changes can be implemented. For example, very early on, one should verify that there is support in the literature for the instructional approach and theoretical basis

underlying the tutor design. At a later stage, formative evaluation might involve novices interacting with the ITS and experts commenting on the completeness and accuracy of curricular elements, as well as the validity of the feedback provided by the tutor (where feedback can either be explicitly offered by the computer, or implicitly occurring, as in discovery worlds under "natural feedback" conditions). Results from the formative evaluation can inform the tutor's designer and programmer about bugs in the software, other software-related problems that students encountered, suggestions from students and experts for modifying the interface, and conceptual or procedural problems with the curriculum (e.g., an inaccurate depiction of current flow through a parallel DC-circuit). Two examples of formative evaluations follow, representing our own experiences with evaluating ITS.

Pilot study of the Bridge tutor. Prior to running a large-scale evaluation of the Bridge tutor (Shute, 1991; Shute & Kyllonen, 1990), a pilot study was conducted to determine if the tutor actually worked with "real" subjects (i.e., not laboratory-related personnel). Approximately 200 pilot subjects were run on the tutor, and, to our dismay, we discovered that many of them had significant difficulties learning the programming curriculum because they lacked (or forgot) some prerequisite knowledge presumed by the system (e.g., not knowing what an integer or variable was). Findings from the pilot study highlighted about 10 weak concepts in programming and math: integer, real number, string, data, sum, product, constant, variable, expression, and value assignment. Consequently, we built a "pre tutor," an approximately 2-hour computer-assisted instruction (CAI) module that instructed those 10 concepts. Subjects received on-line definitions of concepts, followed by a series of questions pertaining to the concept (e.g., Is 7.34 an example of an integer?). After each response, feedback was provided on both the accuracy of the response and the item in question (e.g., "No, 7.34 is not an example of an integer because integers are positive or negative whole numbers without decimal points, and 7.34 contains a decimal"). This pre tutor presented items in a learning-by-doing format with a strict mastery learning criterion. That is, the learner had to successfully answer 16 consecutive questions about each concept (two blocks of 8 questions) before it would drop out of the learning cycle. In the large-scale study, once subjects encountered the Bridge tutor, no longer did hands wave and subjects lament, "What is a real number??" The problem was solved, and learning Pascal programming skills was not confounded by inadequate knowledge of necessary concepts.

Protocol analyses with Smithtown. During the course of developing Smithtown, we collected protocol data on effective and less-effective inter-

rogative strategies from individuals interacting with Smithtown (Shute & Glaser, 1991). This pilot data formed the basis for the diagnostician that subsequently monitored learners' inquiry skills. The diagnostician evaluated how a student was interrogating the guided-discovery world of microeconomics, and determined whether he or she was proceeding in a systematic, efficient manner. By collecting think-aloud protocols from a heterogeneous group of subjects as they interacted with Smithtown, we were able to delineate a large group of effective and ineffective inquiry behaviors that the diagnostician should monitor (see Shute & Glaser, 1991 for a complete listing of these behaviors). For example, the first inquiry behavior was Baseline data collection: For an initial exploration of Smithtown (where "initial" refers to time unit 1 in any student-created experiment), were data collected from the market in equilibrium, before any variables were altered? Failing to do this (i.e., changing a variable without first viewing the market in equilibrium) would increment the counter for the corresponding "Baseline data collection critic." If this count exceeded some specified threshold (e.g., three times), then the coach would address the learner about the problem (e.g., "I see that you've been changing too many variables before looking at the market first...").

Summative Evaluations

At the end of ITS development, or at the end of major development stages, summative evaluations serve to assess various aspects of the finished product. The appropriate summative evaluation depends on how you answer the questions underlying Principles 1 and 2, above. This paper outlines five different designs that are suitable for summative evaluation studies:

(1) Within-system Design—How do two or more alternative versions of a single tutor compare to one another?
(2) Between-system Design—How effective is your tutor in relation to another one teaching the same subject matter?
(3) Benchmark Design—How does your tutor fare in relation to some standard instructional approach?
(4) Hybrid Design—A combination of the above options, and
(5) Quasi-experimental Design—How well does your system operate in a real-world setting?

Suppose that you have just completed the development of an intelligent tutoring system (ITS), and you believe it will teach subject matter X effectively and efficiently. You have exhaustively combed the literature and based the tutor's design on valid and general learning and instructional design theories. You have articulated the intentions of the tutor, the goals of the evaluation study, and have identified the target population. It has been developed on a fairly standard computer, and the pilot test has shown it to run bug-free. You may even have two or more versions of the tutor implemented (e.g., one version with extensive feedback, and one with only natural feedback). Now it is time to ascertain its true worth. How do you objectively evaluate the ITS in a controlled experimental setting? Five types of summative evaluation studies are now described.

Within-system Designs. This kind of evaluation provides the opportunity to develop alternative versions of one tutor and see (a) which version is, overall, more effective and/or efficient, and (b) if there are differential benefits of the versions, depending on characteristics of the learner (i.e., a classic test of aptitude-treatment interaction, ATI). Thus, it represents a comparison against itself rather than against some other approach. This evaluation is accomplished by manipulating critical aspects of the tutor (e.g., the instructional approach or interface features) to generate separate versions of the tutor. These separate versions are then compared for instructional effectiveness on subjects with diverse aptitudes and backgrounds. One strength of this design is the provision of various treatment conditions that are identical except for the independent variables of interest. A within-system design might well be part of a formative evaluation, allowing the designer to select among promising design alternatives.

To illustrate the within-system approach, an Electricity tutor (Shute, in press-a) was used in an experimental study, instructing basic principles of electricity to about 400 heterogeneous subjects. Two instructional environments were created from the one tutor, differing only in the computer-generated feedback. All other aspects of the tutor were identical. In the rule-application environment, the ITS told the learners what the relevant principles were, and in the rule-induction environment, learners had to induce principles on their own, only given information about what variables were relevant. One learner characteristic that was examined was "exploratory behavior," a quantified measure of on-line tool usage (e.g., taking a meter reading from a circuit). Results showed that learners evidencing a lot of exploratory behaviors learned significantly faster and scored significantly higher on outcome tests if they had been assigned to the inductive environment than the applied environment. On the other hand, less exploratory learners performed significantly better from the more structured, applica-

tion environment compared to the inductive environment. It is interesting to note that there was no significant main effect due to learning environment on any of the many outcome or efficiency measures used in that study, thus there was no clear "winner," overall.

Between-system Designs. This type of evaluation seeks to identify one or more optimal instructional approaches for a given task from a set of distinct ITS. This kind of design may be called for if you wanted to test your chess tutor versus some other chess tutor. The separate systems are compared for instructional effectiveness on subjects with matched (and diverse) aptitudes. Instructional time is fixed, after which performance is measured in a variety of ways. An example of the between-system approach to evaluation is the Defense Advance Research Project Agency (DARPA) Learning Strategies Project (Donchin, 1989). In this project, DARPA funded several groups of learning theorists to develop theory-based, guided practice environments for a standard criterion task. The criterion task was Space Fortress, a complex video game developed at the University of Illinois involving the coordinated application of cognitive, perceptual, and psychomotor skills. Although the Learning Strategies Project was mostly a success (in terms of knowledge gained and publications), there were two shortcomings. First, because data were collected in diverse locations worldwide, populations exposed to the various treatments differed. This is known because baseline data disagrees among the groups (Gopher, Weil & Siegel, 1989). Second, because only one task was administered, and that task was not designed to selectively assess performance types, it is not known how general the derived principles are. But this does typify a between-system approach with a standard criterion task.

Benchmark Designs. This kind of design is the most common in the literature, and jumps to mind when we hear of an ITS evaluation. How did the tutor fare in relation to a classroom or training center? This benchmark design seeks to test the effectiveness of an ITS against conventional or existing instructional approaches, such as a classroom environment with a human teacher and 30 students in a room. In this case, the appropriate "control condition" would be the classroom, and the "treatment condition," the tutor. Of the tutors mentioned earlier that have been evaluated, three of the four used this kind of benchmark design: Smithtown, the LISP tutor, and Sherlock.

To illustrate, Anderson and his colleagues at Carnegie-Mellon University (Anderson, et al., 1984) developed a LISP tutor which provides students with a series of LISP programming exercises and tutorial assistance as needed during the solution process. In one evaluation study, Anderson,

Boyle, and Reiser (1985) reported data from three groups of subjects: human-tutored, computer-tutored (LISP tutor), and traditional instruction (subjects solving problems on their own). This study employed two benchmarks for comparing the efficacy of the LISP tutor: human (one-on-one) tutoring and traditional instruction. The time to complete identical exercises were: 11.4, 15.0, and 26.5 hours, respectively. Furthermore, all groups performed equally well on the outcome tests of LISP knowledge. A second evaluation study (Anderson, Boyle & Reiser, 1985) compared two groups of subjects: Students using the LISP tutor and students completing the exercises on their own. Both received the same lectures and reading materials. Findings showed that it took the group in the benchmark (i.e., traditional instruction) condition 30% longer to finish the exercises than the computer-tutored group. Furthermore, the computer-tutored group scored 43% higher on the final exam than the control group. So, in two different studies, the LISP tutor was successful in promoting faster learning with no degradation in outcome performance compared to traditional instruction.

Hybrid Designs. Given adequate resources, it is, of course, possible to develop hybrid designs. For example, you may want to compare a variety of approaches to teaching electronics troubleshooting (e.g., two versions of an ITS, a desktop drill-and-practice device, and classroom instruction). In fact, we have just started collecting data from an exploratory study employing such a hybrid design. The study (Shute, Gawlick-Grendell, and Young, 1992) compares the following four groups of subjects learning statistics (where the curriculum is Probability):

(a) Stat Lady, an ITS teaching statistics in a graphics-rich experiential learning environment;
(b) a paper and pencil version of Stat Lady consisting of all the same problem sets as the computer version, but all instruction and problem solving is on paper;
(c) a classroom environment consisting of identical curricular elements as contained within Stat Lady, but with a different presentation of materials; and
(d) a no-treatment control group to determine baseline measurements of pretest to post-test changes in learning.

The simple, main effects hypothesis regarding learning outcome is: no-treatment control group $<$ lecture group $\leq$ the paper and pencil version of Stat Lady $\leq$ computerized version of Stat Lady. Further-

more, we plan to investigate the data for possible aptitude-treatment interactions, whereby some individuals may benefit from the computer environment, while others may learn better from the more structured, lecture environment.

Quasi-experimental Designs. A true experimental design in evaluation studies calls for randomly assigning subjects to treatment conditions. Sometimes, however, random assignment in not feasible (i.e., impractical or impossible), such as when subjects self-select or when intact classrooms are studied. In these instances, the internal validity of the evaluation may be suspect. This means that there is a reduced ability to affirm that any observed outcome differences are causally linked to the treatment manipulations. On the other hand, this reduction in internal validity with field studies can sometimes buy you an increase in external validity. This would mean an increased ability to affirm that the treatment works in the field (as opposed to the lab) because it was evaluated in the field.

It is possible to conduct "true" experimental evaluations in the field if you are able to randomly assign subjects (or groups, under certain circumstances) to treatments. In cases where quasi-experimental designs are called for, there are several issues (or threats) relating to internal validity that must be addressed. Some of these include: history (external events that influence outcomes), maturation (changes in subjects that occur over time), testing (changes caused by your measurement procedures), mortality (biased loss of subjects over time), and selection (biased assignment of subjects to treatments). Campbell and Stanley (1968) provide an excellent discussion of these and other threats to internal validity, along with a presentation of quasi-experimental designs to minimize such threats.

Reiterating, we prefer theory-driven laboratory evaluations with true experimental designs, followed by data-constrained field evaluations. The laboratory offers a level of experimental control difficult to achieve in a field study, while the field offers a level of external validity not possible in a laboratory. Figure 3 depicts the complete representation of these relationships among: internal and external validity, formative and summative evaluations, different design types, and true vs. quasi-experiments. Finally, the area in the lower-left quadrant of the graph (i.e., low internal validity and low external validity) can be likened to a "research wasteland" as there is nothing that can come out of those studies. In contrast, studies possessing high internal validity and high external validity (i.e., studies conducted in the field but with a true experimental design) are the kind of studies we dream about. However, given the constraint of reality, these results may alternatively be obtained at different times, across progressively more finely-

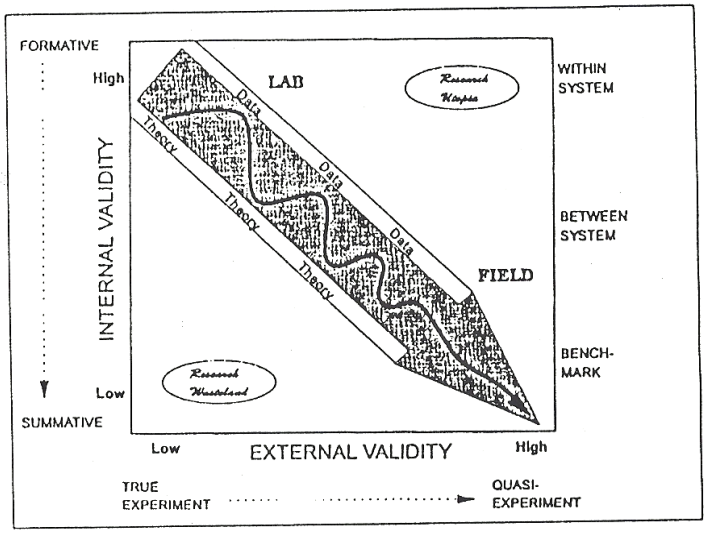

Figure 3. Relationships among internal and external validity, iterative experimentation, and design types.

tuned research studies, and again: driven by theory and modified by empirical findings.

PRINCIPLE 4: INSTANTIATE DESIGN WITH APPROPRIATE MEASURES, NUMBER & TYPE OF SUBJECTS, AND CONTROL GROUPS

While the preceding principles may be likened to the "skeleton" of an evaluation study, Principle 4 constitutes the "viscera." You've now selected one of the several design types that suits your needs and research questions. The next step is to carefully plan the details of the design by considering and instantiating the dependent (outcome) and independent measures, the number and type of subjects needed in your experiment, and the appropriate control group(s).

Learning Outcomes (or Dependent Measures)

A very common problem in ITS evaluations has to do with the inadequate selection, or poor implementation, of criterion tasks and other dependent measures assessing knowledge and skill acquisition (collected before, during, and following instruction). It is extremely important to give careful consideration to the dependent measures you will use in your evaluation. Ask yourself how the dependent measures reflect the goals of the ITS, then relate the measures specifically to the issues mentioned in the goals of the study. If, for example, the goal of your tutor is to teach particular knowledge and skills associated with introductory statistics, and the goal of your evaluation study is to test the efficacy of your tutor in relation to traditional classroom instruction, then your learning criteria should measure the acquisition of the specified knowledge and skills associated with introductory statistics. To illustrate, we are currently evaluating an ITS for introductory statistics, Stat Lady (Shute, Gawlick-Grendell, & Young, 1992) that includes multiple instruments and measures to assess understanding (and enjoyment) of the course material. Some of these measures include:

(a) Open-ended questions (e.g., Why is random sampling from some population so important?),
(b) Survey (e.g. How well do you feel you learned statistics?),
(c) Final "research project" (Students pose a research question, use the on-line "number scoop" to collect computer-generated data, compute appropriate data analyses, summarize results, and type a paragraph about what it all means),
(d) Concepts post-test (e.g., The "variance accounted for" in regression analysis is represented by ___ and refers to ____?),
(e) Skills post-test (e.g., From the following data, compute the correlation between number of hours a person sleeps and their income. Is this significant? What does it mean?)
(f) Final statistics "exam" (Students, in pairs, compose an on-line exam, allegedly to administer to other students to measure their understanding of statistics.

Students are encouraged to include in their "test" a variety, and sufficient number, of questions covering a range of material. In addition, they need to include an answer key for all items). These measures, collectively, will really reflect the degree to which students learned the course material.

There are at least two good reasons why you should use multiple dependent measures. First, because ITS instruction is done on computers, you

have the option to capture as much data, of whatever kind, you choose. You should err on the side of gathering too much data. You do not have to analyze it all at once, but if you don't collect it, you will not have the option of going back to look at something you may later realize is important. Second, it is the nature of learning and instructional research that the effectiveness of an intervention will depend, in part, on the aspects of performance you are trying to teach, and how you measure these indicators of performance.

In summary, multiple dependent measures are better than fewer measures, and they should reflect the goals of the tutor and the evaluation study. Some possible learning or performance measures that can be collected in an ITS evaluation study include: performance latency, performance accuracy, declarative knowledge, procedural knowledge, procedural skill, automatic skill, secondary task performance, higher-order knowledge, as well as measures of near transfer, far transfer, and skill retention (decay). Teaching one thing and measuring another is bound to result in a failed study.

Independent Measures (Individual Differences)

In addition to collecting learning and performance measures, you should also consider collecting individual differences measures. Individuals come to any new learning task with differing profiles of knowledge, skills, and traits (i.e., individual difference dimensions). Some common individual differences measures include general intelligence, GPA, and SAT scores. In addition, you may want to collect data on cognitive process measures (e.g., working memory capacity, information processing speed), personality measures (e.g., impulsivity, aggression, introversion), and demographic information (gender, age, years of school, experience with computers). If you don't collect these data and then discover some potential confound in your experimental design (e.g., two schools with different mean IQs for enrolled students), you can't easily correct it. For instance, having IQ data from the students would allow you to statistically control for this variable in the equation, and then test for differences between the two sites with IQ parceled out. Another problem associated with not collecting these data arises when you conduct a treatment-effects study and find, overall, no differences between two approaches across a variety of outcome measures. You would probably conclude that there are no differences between the approaches (e.g., Sleeman, Kelly, Martinak, Ward, & Moore, 1989). However, when aptitudes are considered in the equation, then you may find in-

stances of aptitude-treatment interactions, where one kind of person performs better in one environment, and another performs better in a different environment (e.g., Shute, 1992, in press-a, in press-b). Because ITS are intended to adapt instruction to the learner, it makes sense to understand the learner as thoroughly as possible. Thus, aptitude data can be as theoretically important in analyzing learning from an ITS as traditional, lower-level performance data used in cognitive diagnosis and remediation.

In summary, collecting individual differences measures in your experimental study is a good idea, for several reasons. First, you can be sure that your treatment effects are real, and not simply an artifact of differential learner traits. Second, if you have collected aptitude data and find that they do impact the treatment condition, if necessary, you may then statistically control for those data that affect the treatment condition. Finally, if you don't have any aptitude data, then you cannot investigate aptitude-treatment interactions.

Control Conditions

One of the biggest problems in designing "rigorously controlled" evaluation studies is identifying suitable control conditions. The choice of treatment condition, as well as the proper control condition(s), must be principled, based on a theoretical approach to performance. Historically (see Cronbach & Snow, 1977), one of the main reasons that treatment effects were difficult to find was because of various uncontrolled conditions and unanticipated interactions across settings (e.g., different instructional materials, classroom dynamics, and teachers' personalities) (see also Shute, 1992). Lessons learned from these older studies conducted with classroom environments yield certain conventions that may be adopted to eliminate control-condition problems in ITS evaluation research:

(a) Use tutors that are based on a theoretically principled approach to learning and instruction;
(b Give preference to data collected at a single site (rather than multiple sites) with standard procedures and measures;
(c) Obtain a range of demographic and aptitude measures from subjects; and
(d) Pre-specify a standard criterion task along with multiple dependent measures to be taken at various intervals during the course of learning.

We are moving in the direction of designing and validating standardized criterion tasks (see Shebilske, Regian, Arthur, & Jordan, 1992).

Another problem concerning control conditions for ITS research is the creation of Hawthorne effects. Similar to placebo effects, Hawthorne effects are treatment differences due only to the fact that one group (i.e., the group receiving instruction on the tutor) is receiving special attention and consideration. These effects are easily obtained, and thus, must be carefully avoided. One should be wary of treatment and control groups that are treated differently from baseline, or if the control group is primarily a no-treatment control.

A third important decision concerns time. Two main options include: (a) Should time be allowed to *vary* across subjects, as with self-paced and mastery learning (e.g., Levin, 1974), or (b) Should time be held *constant*, allowing achievement to vary (e.g., Shebilske, et al., 1992)? If the end-user of the ITS evaluation research is an adherent to lock-step instruction, one recommendation is to fix time and see how the system reduces individual differences in outcome performance. This decision is less critical for criterion-based instruction because the learner or trainee becomes employable as soon as he or she reaches criterion on task performance.

Finally, suppose you are interested in the relative costs and benefits of the ITS versus alternative approaches. Your control conditions (or historical data) must similarly be based on cost/benefit data from the alternative approaches.

Subjects

Another problem area in ITS evaluation has to do with obtaining the right *type* and *number* of subjects for the experiment. First, it is imperative to identify the target population for whom the tutor is intended (e.g., university students taking an introductory Astronomy course, jet-engine mechanics, or students beginning 8th grade algebra). Then, be sure that the sample you are testing matches your target population. That is, if the purpose of your ITS is to teach university students a certain curriculum, and your test subjects come from a different population, you won't be able to accurately assess the effectiveness of your tutor. Different populations of subjects vary across an assortment of dimensions. For example, Smithtown was created as a guided-discovery microworld for college students to induce principles of microeconomics (Shute & Glaser, 1990). We conducted a series of studies using Smithtown, first with a group of university students (N=10), and then with 530 U. S. Air Force recruits on their 6th day of basic training. For the most part, the university students enjoyed the

freedom of the environment to conduct experiments in Smithtown, collect and record data, and make generalizations about their findings in inducing microeconomic laws and principles. In comparison, the majority of recruits did not enjoy the unstructuredness of the system, and their relative performance and acceptance of the system was considerably less positive.

Second, you need to figure out how many subjects are needed for the study. As a rule-of-thumb, ITS evaluations should have at least 30 subjects per condition for simple treatment comparisons. Ideally, if one condition takes place in a group setting (e.g., classroom environment for the control condition), then the treatment condition should similarly take place within a group setting. Often, however, this is not possible, due to a shortage of computers on which to run the treatment (ITS) condition all at the same time. In this case, analyses of outcome should take into consideration the fact that the two groups differed in testing/learning environment, and test for differences in variability between groups on select measures. Usually, this is not a problem in a well-controlled experiment.

Third, aptitude-treatment interaction (ATI) studies, using individual difference measures as independent variables, should use about 100 subjects per treatment (Cronbach & Snow, 1977). This rule-of-thumb can be relaxed somewhat for sufficiently powerful designs involving extreme groups, or matched cases. Most investigators in the ATI tradition before 1980 used 40 or fewer subjects per treatment, and may have lacked the power to pick up even moderate effects. Keep in mind the relationship between sample size and power. The ability to pick up a given treatment effect goes up as sample size increases. It is thus possible to estimate the required sample size for picking up a treatment effect of some hypothesized magnitude. Also keep in mind the weak relationship between statistical significance and real-world importance. With enough power (large enough sample), you can pick up tiny, but reliable, treatment effects, even though the effect size may be too small to be of practical importance.

Finally, random assignment of subjects to conditions is critically important and should be achieved whenever possible. If subjects are not randomly assigned to treatment or control conditions, then any ensuing treatment effects may be attributable to a host of confounds. Possible confounds include self selection, site differences, experimenter bias, and so forth. If it is not possible to randomly assign subjects, then it is important to think carefully about potential confounds, and then measure and control them, as necessary.

PRINCIPLE 5: MAKE NECESSARY LOGISTICAL PREPARATIONS FOR CONDUCTING THE STUDY

Many potentially good studies are ruined because of poor logistical preparations. Some examples include: subjects failing to provide a critical piece of data (and omissions weren't caught in time to rectify), a proctor not executing an important step (given an ambiguous or incomplete "script" by the researcher), sufficient materials were unavailable at the data collection site, or other reasons, unrelated to design. These kinds of disasters can render expensive data useless. They can be avoided with careful planning, training, and general preparation. Be sure you have qualified, trained personnel to implement the study. Provide the proctors with clear, complete scripts and procedural checklists. Ensure that subjects are being treated the same at all locations.

You should also consider, in advance, possible "worst-case" scenarios, such as what you would do in case your hardware or software fails. Specifically, do you require computer analysts or technicians at each data collection site? What plans do you have for rescheduling subjects in the event of some computer-related problem? It is a good idea to have extra computers at the site as backup systems.

PRINCIPLE 6: PILOT TEST THE TUTOR AND THE STUDY

This principle is very important. There may be a very large gap between the way you believe students will interact with your system and the way they actually interact. It is important to find these things out before committing to the expense and trouble of a full evaluation. Pilot testing of the tutor is simply a form of formative evaluation, and "an ounce of prevention is worth a pound of cure."

Some things to look for during the pilot test of the tutor include:

(a) Is the tutor running bug-free?
(b) Do subjects know what they should be doing at all times?
(c) Are subjects learning anything?
(d) Do subjects indicate that they like the system?
(e) Did you estimate the learning time appropriately, or do subjects take longer (or less time) than you had anticipated? and
(f) Were all subjects able to complete the tutor?

It is also a good idea to run pilot subjects through each condition in the study. During this "dress rehearsal," watch the subjects within each phase of the study, interview them afterward, and take a look at the data. Some things to look for during the piloting of the study include: (a) Are your time estimates accurate? (b) Are your manipulations having the desired effect? (c) Are the data recorded and stored as you intended?

PRINCIPLE 7: PLAN YOUR PRIMARY DATA ANALYSIS AS YOU PLAN THE STUDY

As you design your study, you should concurrently be considering how you will analyze the data. This section briefly outlines possible statistical techniques, organized by experimental design for illustration purposes only (e.g., types of confirmatory data analyses for conducting a benchmark evaluation study). Each ITS evaluation study will have its own unique requirements for data analysis. Some kinds of statistical analyses are better suited to certain classes of design types than others. Although we recommend you include an appropriately trained individual in your design team, some rules-of-thumb for common analyses follow.

Confirmatory Data Analyses. When you have a specific hypothesis you want to test, you should employ a confirmatory data analysis technique. Data arising from some of the evaluation designs discussed in this paper, such as a benchmark evaluations, could be analyzed by t-tests, correlations, Chi-square, confirmatory factor analysis, or analysis of variance, depending on the focus of the research questions. Some statistics books even offer decision trees for choosing appropriate analyses (e.g., Howell, 1982).

Exploratory Data Analyses. Too often, researchers fall into the trap of equating "data analysis" with confirmatory analysis—seeking answers to questions like, "Do my data confirm the hypothesis that Treatment A is better than Treatment B?" But that position closes the door on exploring alternative patterns that may exist in your data. Exploratory data analysis is interactive and iterative with no fixed procedure to analyze the data. In terms of the experimentation cycle depicted in Figure 3, exploratory analyses (theoretically-driven) would occur at the starting point of the big arrow, progressing toward more confirmatory analyses (empirically-constrained). Thus, exploratory analyses tend to suggest rather than confirm hypotheses, and both confirmatory and exploratory techniques have important parts to play in ITS evaluation studies. As an historical illustration of the importance of exploratory methods, consider the evolution of our current theories

of intelligence and abilities (e.g., Ackerman, 1988; Anderson, 1987; Kyllonen & Christal, 1989). They began as results from exploratory factor analyses (e.g., Thurstone, 1938; Guilford, 1967), and thus contributed to the *development* of these theories.

So, for studies that are exploratory in nature, (e.g., pilot studies, formative within-system evaluations), exploratory techniques are suitable. Some relevant data analysis methods include: exploratory factor analysis, cluster analysis, multiple regression analysis, and structural equation modeling (e.g., Shute, 1992, in press-a, Shute & Kyllonen, 1990).

Cost-Benefit Analyses. If you are interested in evaluating other aspects of the ITS, there are many standard methods documented to estimate both cost and utility of systems (e.g., Stone, Turner, Fast, Curry, Looper, & Engquist, 1992). For instance, you may want to assess whether the programming costs associated with implementing various bells and whistles are justified in terms of increased learning outcomes. One study (Shute, in press-a) found that differential usage of some computerized exploratory tools in an electricity tutor (e.g., tools allowing subjects to change circuit component values, meter on parts of the system, view the hypertext dictionary of electricity concepts) did *not* enhance learning (in fact, there was an overall negative correlation between tool usage and outcome). However, use of the on-line tools was optional, thus most learners elected not to use them. One suggestion would be to make on-line tool usage mandatory during the early stages of learning (where there was a positive relationship between tools and outcome), then gradually reduce that requirement. Other aspects of the ITS, such as the cost of the hardware and software should be entered into the cost-benefit equation, as well. Finally, we recommend that you automate the moving, recoding, and formatting of data as much as possible, and carefully check your automated procedures. Try to keep human recoding to a minimum to reduce errors. It is possible to be very efficient managing data that is initially collected on the computer. Think ahead.

CONCLUDING REMARKS

This paper was motivated by the fact that many ITS presently exist, but there are correspondingly few controlled evaluations. We have been involved with a number of large-scale ITS evaluations, and know about results from many more evaluation studies, some successes and some failures. A lot of the failures could have been avoided, we believe, had the experimental design been better planned, operationalized, and implemented. As an attempt to reduce future flawed ITS studies, we have proposed a list

of seven main principles that should be considered when conducting an evaluation study:

(1) Delineate the goals of the tutor,
(2) Define the goals of the evaluation study,
(3) Select the appropriate design to meet the defined goals,
(4) Instantiate the design with appropriate measures, number and type of subjects, and control groups,
(5) Make necessary logistical preparations for conducting the study, (6) Pilot test the tutor, and
(7) Plan your primary data analysis as you design the study. For each of the principles, we defined a set of issues to contemplate.

In addition, we included a collection of experimental designs that serve different research purposes. This paper also describes various rules-of-thumb for selecting the appropriate design for your particular research questions, determining the appropriate dependent and independent measures, and deciding on the correct sample size and control conditions for your experiment.

There are a few "take home" messages from this paper (in addition to the seven principles). First, it is not possible to be over-prepared for a major study. Second, for your dependent and independent variables, you should use validated instruments, if they are available. If you have thoroughly read the literature, you will know what is available. Talk to experts who are more knowledgeable in certain areas and domains than you are. In addition, be creative. It may be better to adapt something to suit your particular needs rather than to make something up, from scratch. However, given an intimate knowledge of the tutor's curriculum, as well as what you want the learners to walk away with, you can design appropriate tutor-specific outcome measures, provided they adhere to basic psychometric tenets (e.g., enough problems in the tests/surveys to determine reliability).

Building a tutor and not evaluating it is like building a boat and not taking it in the water. We find the evaluation as exciting as the process of developing the ITS. Often, the results are surprising, and sometimes they are humbling. With careful experimental design, they will always be informative.

References

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General, 117*, 288-318.

Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem solutions. *Psychological Review, 94*, 192-210. Anderson, J.R., Boyle, C. and Reiser, B. (1985). Intelligent tutoring systems. *Science, 228*, 456-462.

Anderson, J.R., Farrell, R., and Sauers, R. (1984). Learning to program in LISP. *Cognitive Science, 8*, 87-129.

Baker, E. L. (1990). Technology assessment: Policy and methodological issues. In H. L. Burns, J. Parlett, and C. Luckhardt (Eds.), *Intelligent tutoring systems: evolutions in design*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bloom, B.S. (1956). Taxonomy of educational objectives: The classification of educational goals. In B. S. Bloom (Ed.), *Cognitive domain, handbook 1*. New York: McKay.

Bloom, B.S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4-16.

Burton, R.R., and Brown, J.S. (1982). An investigation of computer coaching for informal learning activities. In D. Sleeman and J.S. Brown (Eds.), *Intelligent tutoring systems*. London: Academic Press.

Campbell, D.T., and Stanley, J.C. (1963). Experimental and quasi-experimental designs for research. Chicago: Rand McNally.

Carroll, J. (1963). A model of school learning. *Teachers College Record, 64*, 723-733.

Cohen, P.A., Kulik, J. and Kulik, C.C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19*(2), 237-248.

Cronbach, L.J. & Snow, R.E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.

Donchin, E. (1989). The learning strategies project, *Acta Psychologica*, 71, 1-15.

Gopher, D, Weil, M., & Siegel, D. (1989). Practice under changing priorities: An approach to the training of complex skills, *Acta Psychologica*, 71, 147-177.

Guilford, J. P. (1967). *The nature of human intelligence*. New York: McGraw-Hill.

Howell, D. C. (1982). *Statistical methods for psychology*. Duxbury Press, Boston, MA.

Kyllonen, P.C. & Christal, R.E. (1989). Cognitive modeling of learning abilities: A status report of LAMP. In R. Dillon & J.W. Pellegrino (Eds.), *Testing: Theoretical and applied issues*. San Francisco: Freeman.

Lesgold, A., Lajoie, S.P., Bunzo, M., and Eggan, G. (1992). A coached practice environment for an electronics troubleshooting job. In J. Larkin, R. Chabay and C. Sheftic (Eds.), *Computer-assisted instruction and intelligent tutoring systems: Establishing communication and collaboration*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Levin, H.M. (1974). The economic implications of mastery learning. In J.H. Block (Ed.), *Schools, society, and mastery learning*. New York: Holt, Rinehart, & Winston.

Lewis, M.W., McArthur, D., Stasz, C., & Zmuidzinas, M. (1990). Discovery-based tutoring in mathematics. *AAAI Spring Symposium Series*. Stanford University, Stanford, CA.

Littman, D. and Soloway, E. (1988). Evaluating ITSs: The cognitive science perspective. In M.C. Polson and J.J. Richardson(Eds.), *Foundations of intelligent tutoring systems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Nichols, P., Pokorny, R., Jones, G., Gott, S.P., and Alley, W.E. (in preparation). Evaluation of an avionics troubleshooting tutoring system. Technical Report, Armstrong Laboratory, Human Resources Directorate, Brooks AFB, TX.

Pressey, S.L. (1926). A simple apparatus which gives tests and scores-and-teaches. *School and Society, 23*; 373-376.

Pressey, S.L. (1927). A machine for automatic teaching of drill material. *School and Society, 25*; 549-552.

Schofield, F. W., & Evans-Rhodes, D. (1989). Artificial intelligence in the classroom: The impact of a computer-based tutor on teachers and students. In D. Bierman, J. Brueker, & J. Sandberg (Eds.), *Artificial intelligence and education: Synthesis and reflection*. Amsterdam, the Netherlands: IOS.

Shebilske, W. L., Regian, J. W., Arthur, W., & Jordan, J. (1992). A dyadic protocol for training complex skills. *Human Factors, 34*, 369-374.

Shute, V.J. (1991). Who is likely to acquire programming skills? *Journal of Educational Computing Research, 7*(1), 1-24.

Shute, V.J. (1992). Aptitude-treatment interactions and cognitive skill diagnosis. In J.W. Regian and V.J. Shute (Eds.), *Cognitive approaches to automated instruction*. Hillsdale, NJ: Erlbaum.

Shute, V.J. (in press-a). A Comparison of learning environments: All that glitters... In S.P. Lajoie and S.J. Derry (Eds.), *Computers as cognitive tools*. Hillsdale, NJ: Erlbaum.

Shute, V.J. (in press-b). A macroadaptive approach to tutoring. To appear in the *Journal of Artificial Intelligence and Education*.

Shute, V.J. and Glaser, R. (1990). A large-scale evaluation of an intelligent discovery world: Smithtown. *Interactive Learning Environments, 1*, 51-77.

Shute, V.J. and Glaser, R. (1991). An intelligent tutoring system for exploring principles of economics. In R. E. Snow & D. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Shute, V.J. & Kyllonen, P.C. (1990). *Modeling programming skill acquisition* (Report No. AFHRL-TP-90-76). Air Force Systems Command, Brooks Air Force Base, TX.

Shute, V. J., Gawlick-Grendell, L. A., and Young, R. (1992). Stat Lady: An experientially-situated ITS for introductory statistics. Paper to be presented at the American Educational Research Association, Atlanta, GA.

Skinner, B.F. (1957). *Verbal behavior*. Englewood Cliffs, N.J.: Prentice-Hall.

Sleeman, D., & Brown, J.S. (1982). *Intelligent tutoring systems*. London: Academic Press.

Sleeman, D., Kelly, A.E., Martinak, R., Ward, R.D., & Moore, J.L. (1989). Studies of diagnosis and remediation with high school algebra students. *Cognitive Science, 13*(4), 551-568.

Stone, B. M., Turner, K. L., Fast, J. C., Curry, G. L., Looper, L. T., and Engquist, S. K. (1992). *A computer simulation modeling approach to estimating utility in several Air Force specialties* (Report No. AL-TR-1992-0006). Air Force Systems Command, BrooksAir Force Base, TX.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.

Wenger, E. (1987). *Artificial intelligence and tutoring systems*. Los Altos: Morgan Kaufmann Publishers.

Woolf, B. P. (1988). Intelligent tutoring systems: A survey, in Sehrobe H., and the American Association of Artificial Intelligence (eds.), Exploring Artificial Intelligence, Morgan Kaufmann, Palo Alto, Ca. pp 1-44.

Notes

We would like to sincerely thank Lisa Gawlick-Grendell, Pat Kyllonen, and Bill Tirre for their valuable comments and suggestions on this paper. The research reported in this paper was conducted by personnel of the Armstrong Laboratory, Human Resources Directorate, Brooks Air Force Base, Texas. The opinions expressed in this article are those of the authors and do not necessarily reflect those of the Air Force.

Correspondence concerning this paper should be addressed to Valerie J. Shute, Armstrong Laboratory, Human Resources Directorate, Brooks Air Force Base, Texas, 78235.