Fowler, F. J. (1984). *Survey research methods*. Beverly Hills, CA: Sage Publications.
Gagne, R. M., Briggs, L. J., & Wager, W. W. (1992). *Principles of instructional design*. New York: Harcourt Brace Jovanovich.
Gardner, H., & Salomon, G. (1986, January). The computer as educator: Lessons from television research. *Educational Researcher*, 13–19.
Heinich, R. (1984). The proper study of instructional technology. *Educational Communication and Technology Journal*, *32*(2), 67–87.
Jamison, D., Suppes, P., & Wells, S. (1974). The effectiveness of alternative instructional media: A survey. *Review of Educational Research*, *44*, 1–68.
Kearsley, G., Hunter, B., & Sidel, R. J. (1983). Two decades of computer based instruction: What have we learned? *T.H.E. Journal*, *10*, 88–96.
Kozma, R. B. (1991). Learning with media. *Review of Educational Research*, *61*(2), 179–211.
Levie, W. H., & Dickie, K. (1973). The analysis and application of media. In R. M. W. Travers (Ed.), *Second handbook of research on teaching* (pp. 858–881). Chicago: Rand McNally.
Levin, H. M. (1983). *Cost effectiveness: A primer*. Beverly Hills, CA: Sage Publications.
Levin, H. M. (1988, May). The economics of computer-assisted instruction. *Peabody Journal of Education*.
Levin, H. M., & Meister, G. R. (1985). *Educational technology and computers: Promises, promises, always promises* (Report No. 85-A13). Stanford, CA: Center for Educational Research at Stanford, School of Education, Stanford University.
Lumsdaine, A. (1963). Instruments and media of instruction. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 583–682). Chicago: Rand McNally.
Merrill, D. M., Jones, M. K., & Li, Z. (1992). Instructional theory: Classes of transactions. *Educational Technology*, 46–52.
Mielke, K. (1968). Questioning the questions of ETV research. *Educational Broadcasting Review*, *2*, 6–15.
O'Neil, H., Jr., Anderson, C. L., & Freeman, J. (1986). Research and teaching in the armed forces. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 971–987). New York: Macmillan.
Reigeluth, C. (1983). *Instructional design: Theories and models*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Reigeluth, C. M. (1987). *Instructional theories in action*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Saettler, P. A. (1968). *A history of instructional technology*. New York: McGraw-Hill.
Salomon, G. (1981). *Communication and education*. Beverly Hills, CA: Sage Publications.
Salomon, G. (1984). Television is "easy" and print is "tough": The differential investment of mental effort in learning as a function of perceptions and attributions. *Journal of Educational Psychology*, *76*(4), 647–658.
Schramm, W. (1977). *Big media, little media*. Beverly Hills, CA: Sage Publications.
Weiner, B. (1992). *Human motivation: Metaphors, theories and research*. Newbury Park, CA: Sage Publications.
Winn, W. (1982). Visualization in learning and instruction. *Educational Communication and Technology Journal*, *30*(1), 3–25.

4 Evaluating Intelligent Tutoring Systems

J. Wesley Regian
Valerie J. Shute
Armstrong Laboratory, Human Resources Directorate, Brooks Air Force Base, Texas

It has long been claimed that automated instruction has the potential for mass delivery of effective and efficient instruction (e.g., Pressey, 1926, 1927; Skinner, 1957; Woolf, 1988). Over the years, a variety of theoretical approaches have been adopted to pursue that potential (e.g., Burton & Brown, 1982; Carroll, 1963; Cohen, J. Kulik, & C. C. Kulik, 1982; Lewis, McArthur, Stasz, & Zmuidzinas, 1990; Sleeman & Brown, 1982; Wenger, 1987). As early as 1926, Pressey described a device that sought to apply then-contemporary learning theory to the task of automated instruction. The mechanical 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 experiment 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)

Pressey's ideas were probably viewed by some as quite promising, given their relationship to then-current learning theory, but they were never applied or even evaluated in any rigorous sense. Today, intelligent tutoring systems (ITS) epitomize the notion of theory-based, individualized, automated instruction. Many of us are just as excited about the potential of ITS as Pressey was about his teaching

machine. Unfortunately, although ITS have been in existence for well over a decade, the degree to which they have been successful remains equivocal due to the relative dearth of controlled evaluations (Baker, 1990; Littman & Soloway, 1988; Shute & Regian, 1993).

Some of the more familiar ITS evaluations reported in the literature 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 (Lesgold, Lajoie, Bunzo, & Eggan, 1992), a tutor for avionics troubleshooting; and Bridge (Shute, 1991; Shute & Kyllonen, 1990) teaching Pascal programming skills. Results from these evaluations show that the tutors can accelerate learning with, at the very least, no degradation in outcome performance compared to appropriate control groups.

How much can we make of these findings? As always, there is a selection bias for publication of unambiguous evidence for successful instructional interventions. Thus, we do not know how many studies found disappointing results. We are personally familiar with other (unpublished) tutor-evaluation studies that were conducted but were "failures." We are also aware of a great many ITS that have been built but never evaluated. We believe that a consistently applied, systematic, and rigorous approach to evaluation would speed the emergence of ITS into applied settings. The primary goals of this chapter are to outline a systematic approach to research and development of intelligent tutoring systems, and to present a set of steps to organize the design of evaluations for these systems.

RESEARCH AND DEVELOPMENT OF ITS: A GENERAL APPROACH

Our general approach has two main thrusts. First, in order to manage the tradeoff between internal and external validity, we believe ITS research and development should progress from laboratory studies of pedagogy in artificial tasks toward field studies of fully implemented ITS for real-world tasks. In other words, begin by identifying powerful instructional manipulations in controlled settings, and then work up to evaluating those manipulations in applied settings. Second, to maximize the efficiency of the research as well as the generality of the results, we believe ITS research and development should be driven by learning theory and constrained by evaluation data. In other words, if it should work, try it; if it doesn't work, change it.

Managing Experimental Validity

Experimental design involves arranging conditions to promote the validity of an experiment. 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 pedagogical research, increases in external validity are generally accompanied by decreases in internal validity. As you increase your ability to generalize pedagogical findings to applied settings, you lose the level of experimental control afforded within the laboratory. We believe the solution to this problem is to initially develop and test pedagogical principles in a laboratory setting with careful attention to experimental control. Promising approaches should then be tested in increasing fieldlike settings, and ultimately in applied settings with careful attention to external validity.

Figure 4.1 depicts the posited tradeoff between internal and external validity; as internal validity decreases, external validity increases. This relationship is particularly true with regard to research on pedagogy. Research in laboratory settings is desirable because of the experiment control that is possible in the laboratory. One can control for prior knowledge, assign subjects to groups, counterbalance for teacher (experimenter) effects, and so on. On the other hand, 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. Field research on pedagogy, if done well, can have high external validity, but often at the expense of internal validity.

Our approach to managing the tradeoff between internal and external validity is to begin with laboratory research (high experimental control and internal validity) using carefully designed laboratory tasks. As we find instructional manipulations that are powerful, we attempt to replicate the effect with more realistic tasks, again within the laboratory. Eventually, we study the intervention

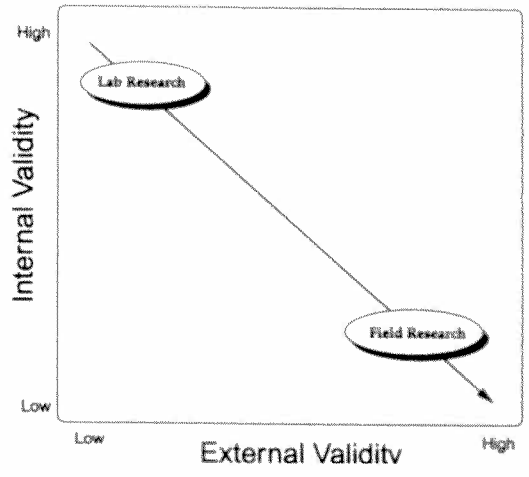

FIG. 4.1. Simple inverse relationship between internal and external validity.

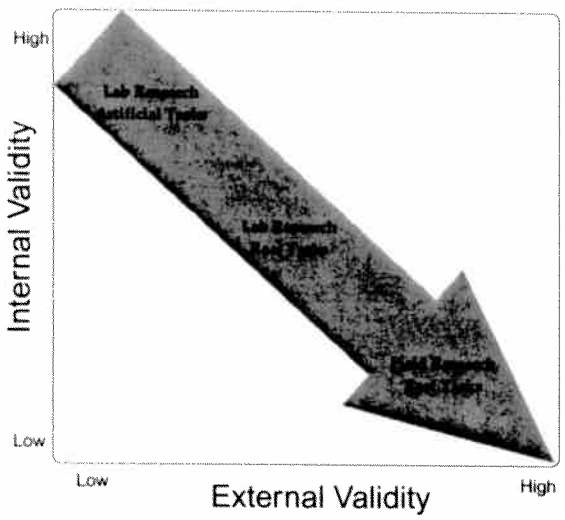

FIG. 4.2. Categories of pedagogy studies along a validity vector.

in an applied instructional context (field setting) with real-world tasks. Figure 4.2 depicts this flow, and places these stages of research on the validity vector from Fig. 4.1. We believe that neither laboratory nor field research alone will give a complete and accurate picture of the instructional effectiveness of a particular intervention. Further, we believe the choice of target tasks used to study instructional interventions at the various stages deserves careful consideration.

Managing Experimenter Bias

We believe that ITS research and development can advance more rapidly if the process is both driven by learning theory and constrained by empirical data. By *learning theory,* we mean a coherent, plausible body of ideas about how people acquire, store, retrieve, and apply knowledge and skill. We find that instructional prescriptions (or "theories" of instruction) that fail to address mechanisms of human knowledge acquisition and representation, are inadequate for our purposes. However, even if the design of an ITS is carefully linked to well-established theory, its value can only be ascertained from empirical testing. Theory is important in generating hypotheses about teaching and learning, and in driving generalizations about pedagogical effectiveness across instructional domains, but empirical testing is critical in order to judge how these ideas fare in reality. Only empirical data, with appropriate control conditions, can provide convincing documentation of the effectiveness of an intervention. Figure 4.3 shows this proposed cyclical relationship between theory and data, where

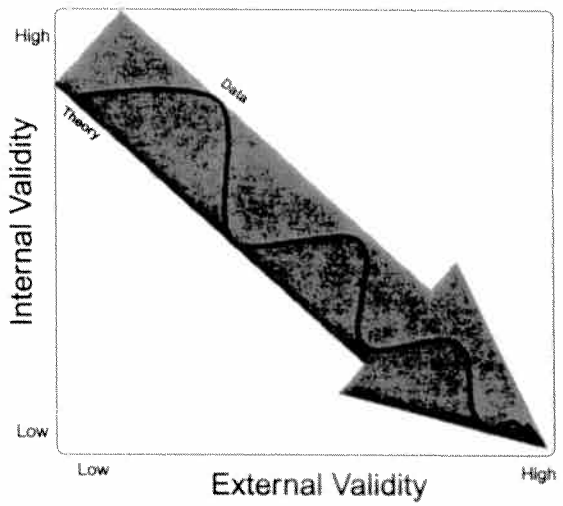

FIG. 4.3. Iterative process of experimentation—driven by learning theory and constrained by evaluation data.

research begins with theory that is progressively modified by empirical results.

In summary, empirical data about the effectiveness of theory-based instruction provides feedback about how well our implementation works, and may also lead to a revision of the original theory. We believe it is extremely important to evaluate systems rigorously and often in order to make progress in ITS effectiveness.

In the next section, we present examples of research that reside at each of three anchor points along the validity vector in Fig. 4.2. Our intention is to give a sense of the kinds of studies that are possible, and to indicate how the results of such studies can influence subsequent ITS work. It is important to keep in mind, however, that we view these three stages of research as relatively arbitrary points along a continuum, rather than truly distinct categories. We present examples of (a) laboratory research employing artificial tasks, (b) laboratory research employing real-world tasks, and (c) field research using real-world tasks.

Laboratory Research With Artificial Tasks

Artificial, or laboratory, tasks do not exist (in their exact form) in the real world. Examples of artificial tasks that have been used over the years in psychological research include memorization of nonsense syllables, cryptarithmatic, forced-choice reaction time tasks, and cursor-tracking tasks. Such tasks are devised to allow uncontaminated study of various phenomena of interest to the experimen-

ter. Artificial tasks are useful in pedagogical research because they allow us to eliminate or reduce the effects of prior knowledge on learning, isolate very specific components of real tasks of interest, and study knowledge/skill acquisition in shorter time frames than would be required for real tasks.

Space Fortress *Research on Small-Group Pedagogy.* *Space Fortress* is an example of an artificial task that has a rich record as a research tool for studying issues of training and skill acquisition (see *Acta Psychologica,* Vol. 71, 1989). This gamelike dynamic control task was originally developed at the University of Illinois under funding from the Defense Advanced Research Project Agency (DARPA) as part of the Learning Strategies project (Donchin, 1989). We have developed an updated version of the program to run in our laboratory, with more flexibility than the original for our research purposes. Shebilske and Regian (1992), and Shebilske, Regian, Arthur, and Jordan (1992) report some of the research conducted in our laboratory using *Space Fortress.* For instance, we have found that it is possible to train up to four subjects on a single computer while achieving individual performance levels equivalent to those attained by subjects trained in the same amount of time on four separate computers. This is achieved by using a training protocol that involves a combination of whole-task practice, shared part-task practice, and observational learning among trainees. These findings are leading us to design prototype instructional systems that operate from the perspective of small-group pedagogy rather than individual pedagogy. In general, automated instruction is more cost-effective and sometimes more instructionally effective when designed for small groups rather than for individuals. When teaching multiple students simultaneously at a single computer, the hardware investment is reduced, as are demands on human instructor time. Also, students in small groups tend to teach one another, benefiting both the provider and the recipient of the instruction. Because students diagnose and remediate each other's performance, the difficult problems of automated diagnosis and natural language processing are avoided.

Space Fortress *Research on Gender-Related Performance Differences.* Another research area that we have been examining with *Space Fortress* involves the significant spatial component of the task. Typically, spatially loaded laboratory tasks yield robust gender effects with males outperforming females. This effect occurs on static spatial tasks (e.g., mental rotation, mental paper folding, form boards) as well as dynamic ones (e.g., collision estimation, dynamic control of moving figures). There has been a long-running argument over whether this gender difference in spatial performance is due to sex-related biological differences or differential developmental experiences. There is clear evidence that at least some of the difference is related to differences in testosterone levels (e.g., Kimura, 1992; Shute, Pellegrino, Hubert, & Reynolds, 1983). Recently, we have been examining instructional interventions that may overcome the hypothesized experiential deficits. We were surprised to find a very simple intervention that seems to go a long way in this direction; namely, placing women in discussion groups with men. We had female subjects participate in brief but regularly occurring (and specifically structured) discussion groups with male subjects, to talk about *Space Fortress* strategy and tactics. This simple intervention dramatically increased women's *Space Fortress* performance. Short discussion groups following practice sessions produced a small positive effect on men's skill acquisition, but a significant positive effect on women's. The women under this treatment performed nearly as well as the men. This very inexpensive intervention may help reduce gender differences in skill acquisition for other kinds of tasks as well.

Loader *Research on Mental Models During Training.* A second example of an artificial task from our laboratory is *Loader*—a complex procedural task (Farquhar, 1992). It requires subjects to execute long sequences of console-operation actions (e.g., button presses, switch actuations, dial rotations) to accomplish specific goals. The task is based on a computer-simulated console that controls railroad cars, tracks, and cranes in a fictitious railroad yard. The task is designed to be a laboratory analog of procedural console operations and process control tasks, which are common in industrial and defense settings.

We hypothesized that acquisition of *Loader* performance skill would be supported by the development of a dynamic mental model linking console actions to events in the "railroad yard." That is, in the process of learning to carry out a specific sequence of actions to accomplish a goal, the operator would come to imagine the corresponding events in the railroad yard, even if she could not actually see the yard while operating the console. We therefore conducted the following experiment (Farquhar, Shebilske, & Regian, 1992). Forty subjects were shown a simple static diagram representing a bird's-eye view of the railroad yard. The diagram depicted the layout of tracks, the initial location of cars, and the locations of bins and the crane. Subjects were told they would learn to operate a console that would enable them to move the cars around on the tracks, and to use the crane to move canisters between bins and cars. Subjects were randomly divided into two groups (i.e., dynamic model vs. no model). During training, both groups received identical text-based instruction in an instructional-window above the *Loader* interface. One group, however, additionally saw a *dynamic* version of the bird's-eye view of the railroad yard. After training, both groups were tested under identical conditions. They were asked to perform the complete procedure without guidance and without access to either type of railroad yard representation. The results were striking. Rather than becoming dependent on the animated rail yard model, subjects in the dynamic model condition apparently internalized the model, as evidenced by their performance after the model was removed. Posttraining performance was 33% faster and 50% more accurate for subjects trained with a dynamic graphical model compared to the no-

model condition, even though the graphical model was not present during testing. This is an example of a very simple graphical aid that can be added to simulation-based ITS to produce significant and enduring performance enhancement.

Laboratory Research With Real Tasks

By real tasks, we mean tasks that constitute all or part of actual tasks that are performed in the world outside of the laboratory. In addition to having counterparts in the real world, they typically are more complex in structure than artificial tasks. Because real tasks tend to be more complex, they may take longer to acquire and performance may be more error-prone than artificial tasks.

Bridge Tutor. "Bridge" is the name of an ITS that teaches a subset of the Pascal programming language, which fits our criteria for being a *real* task (Bonar, Cunningham, Beatty, & Weil, 1988). We conducted a study using this tutor prior to conducting a full field evaluation of Bridge (Shute, 1991; Shute & Kyllonen, 1990). Approximately 200 subjects participated in the initial laboratory study. Many of them, however, had significant problems learning the programming curriculum because they lacked or had forgotten prerequisite knowledge presumed by the system (e.g., not knowing what an integer or variable was). Findings from this study highlighted about 10 weak concepts in programming and math: integer, real number, string, data, sum, product, constant, variable, expression, and value assignment. As a result of this laboratory investigation, we built a "pretutor,"an approximately 2-hr 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 5.24 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, 5.24 is not an example of an integer because integers are positive or negative whole numbers without decimal points, and 5.24 contains a decimal"). This pretutor presented items in a learning-by-doing format with a strict mastery learning criterion. In the subsequent field study, once subjects encountered the Bridge tutor, they no longer had to ask, "What's a variable?" or "What's an integer?" The problem was solved, and learning Pascal programming skills was not confounded by inadequate knowledge of necessary concepts.

Electricity Tutor. MHO is a tutor that teaches basic principles of electricity (Lesgold, Bonar, Ivill, & Bowen, 1989). In one laboratory study (Shute, in press-a), we tested about 400 subjects using two instructional environments created as slightly different versions of MHO. These two environments differed 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 from this laboratory study showed that learners with more 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, application 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, neither of the instructional approaches was a clear "winner," overall. Instead, the study gave us critical information about how subsequent versions of the tutor should adapt to student behavior.

Field Research With Real Tasks

This section discusses studies that have employed real tasks tested in the field as opposed to in the laboratory. As noted earlier, the controls that are possible within laboratory environments may be more difficult, or impossible, to achieve during the conduct of field studies. For example, it may not be possible to randomly assign subjects to treatment conditions in a field study. However, with field research, the ability to generalize to the actual instructional context of interest is enhanced, increasing the study's external validity.

Smithtown. Shute and Glaser (1991) developed an ITS designed to improve an individual's scientific inquiry skills as well as provide a microworld environment for learning principles of basic microeconomics. Both of these foci constitute real knowledge and skills as they are applied outside of the laboratory in the real world. Shute, Glaser, and Raghavan (1989) reported that results from a field study comparing three groups of subjects: a group interacting with Smithtown, an introductory economics classroom, and a control group. The curriculum was identical in both treatment groups (i.e., laws of supply and demand). Results showed that whereas all three groups performed equivalently on the pretest battery (around 50% correct), the classroom and the Smithtown groups showed the same gains from pretest to posttest (26.4% and 25.2%, respectively), significantly outperforming the control group. Although the classroom group received more than twice as much exposure to the subject matter as did the Smithtown group (11 vs. 5 hrs, respectively), the groups did not differ on their posttest scores. These findings are particularly interesting because the instructional focus of Smithtown was not on economic knowledge, but rather on general scientific inquiry skills, such as hypothesis testing.

LISP Tutor. Another example of an ITS field was conducted by Anderson and his colleagues at Carnegie-Mellon University (Anderson et al., 1984). They

developed a LISP tutor that provides students with a series of LISP programming exercises and tutorial assistance as needed during the solution process. In one study, Anderson, Boyle, and Reiser (1985) reported data from three groups of subjects: human-tutored, computer-tutored (LISP tutor), and traditional college instruction (subjects solving problems on their own). The time to complete identical exercises were 11.4, 15.0, and 26.5 hrs, respectively. Furthermore, all groups performed equally well on the outcome tests of LISP knowledge. A second evaluation study (Anderson et al., 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 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.

In this section, we have provided examples of the types of studies we believe are useful at various stages in the development of ITS. We hoped to give a sense of the kinds of studies that are possible, and to indicate how the results of such studies can influence subsequent ITS work. In the following section, we turn to the goal of designing evaluation studies. We describe a set of steps that may be used to organize the design of ITS evaluation studies.

STEPS IN ITS EVALUATION

Outcomes of evaluation studies occasionally reflect the quality 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, and even deficient logistical planning and implementation. In the following sections, we present some general steps that may be followed to implement an effective ITS evaluation (see Shute & Regian, 1993, for a fuller discussion on this topic): (a) Clearly delineate the goals and methods of the tutor; (b) clearly define the goals of the evaluation study; (c) select the appropriate design to meet the defined goals; and (d) instantiate the design with appropriate measures, subjects, and controls.

Step 1: Clearly Delineate the Goals and Methods of the ITS

A careful review of the ITS goals and methods should be undertaken prior to designing an evaluation study. Occasionally, the instructional goals or methods may have shifted over the developmental life cycle of the ITS. In any event, if

the designer of the evaluation study is unfamiliar with the tutor's goals and methods, then designing a good evaluation study is almost impossible. We believe the evaluation designer should be very clear about the following critical issues.

What Instructional Approach Underlies the Tutor? How, generally and specifically, does the system accomplish instruction? Is instruction guided or unguided, student directed or tutor directed? Is knowledge explicitly presented by the system or induced by the student? To what degree will all students have seen the same information or experienced the same interactions? Do students "complete" the tutor after a fixed time period or after reaching some performance criterion?

What Learning Theory Does It Assume? What knowledge or skill-acquisition theory motivates the instructional approach of the tutor? Which aspects of the tutor are directly theory-driven and which are arbitrary? How closely linked are the instructional approach and the learning theory? It is important to distinguish among a learning theory, a general instructional approach, and a specific instantiation of that instructional approach. Failure to do so can lead to over generalizations about evaluation results (e.g., Sleeman et al., 1989).

What Exactly Does It Teach? It is important to be very clear about what students are expected to learn as a result of interacting with the tutor. First, in a concrete sense, what exactly will they know or be able to do after tutoring that they did not know or could not do before? Specific and measurable knowledge or skills should be clearly delineated as the expected learning outcomes. For example, one might hope that students will be able to solve differential equations, list the bones in the human hand, or diagnose faults in a specific electromechanical system. It is also useful to characterize the goals of instruction in a more abstract manner. For example, one might hope to teach procedural skills, declarative knowledge retrieval, or logical problem solving. The ability to generalize findings across or within instructional domains is dependent on some type of theoretical characterization of domain dimensions.

What Other Impacts Is It Expected to Have? Are there other ways in which interacting with the tutor is expected to impact the student? For example, *Smithtown* explicitly instructed scientific inquiry behaviors, but it provided an environment that promoted learning about microeconomics. Less intentional side-effects of tutoring might include near or far transfer of skill, changes in perceived self-efficacy, or modification of attitudes about computers. If you believe such effects are probable and important, then appropriate and objective measures should be obtained to demonstrate the effect. So-called anecdotal evidence is usually the clearest indication of a missed opportunity during an evaluation.

In What Context Is It Supposed to Operate? Is the system intended to supplement existing instruction or provide stand-alone instruction? Is the system targeted to individuals or small groups? What prior knowledge, training, or demographic characteristics are assumed of 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? It is important to clearly specify the environment in which the tutor is intended to operate in order to give it a fair chance of succeeding in evaluation, and in deciding on appropriate control conditions.

Step 2: Clearly Define the Goals of the Evaluation Study

Evaluation studies should not be fishing expeditions. A thoughtful consideration of what you want to know enables you to develop an experimental design that will unequivocally give you that information. You should also be realistic about the difficulties involved in implementing various designs, and adjust your goals at the outset to those that are realizable. Consider the following questions:

What Would You Like to Know After the Study is Completed? What is the primary question you want answered, or alternatively, what is the most important claim you want to be able to make? You may want to know if the tutor is more effective, more efficient, or both, than some instructional alternative at producing criterion performance on some task. You may want to see how much students learn beyond their incoming knowledge and skills, or as a function of their incoming knowledge and skills. You may want to see how tutor effectiveness is influenced by students' individual learning style. You should clearly specify your research questions and hypotheses before you design the study.

How Will You Measure Success, and By What Standard Will You Judge It? Think carefully about how to measure what is being taught, and how you will judge success. Suppose your instructional goal is to teach nine test-taking strategies, and you found that 1 week after students received 2 hrs of tutoring, they were able to state five of these strategies, on average. Was your tutor a success? What if you also learned that 1 week after students received 5 min of instruction using a simple mnemonic approach, they were able to state eight of the nine strategies. By comparison, your tutor would seem ineffective. But what if you found that students learning from your tutor could reliably apply five of the nine strategies, whereas students trained with the mnemonic approach can state, but not apply, eight strategies? Because any human performance is extremely sensitive to the methods used to measure it, your measures of learning should closely reflect the goals of instruction.

You may want to capture quantitative indices, protocols, and/or observational data. Because your subjects will be working at computers, it will be possible to

plan for the online capture of quantitative measures of performance, such as latencies, accuracies, and behavioral counts. With considerably more effort and expense, protocol analyses can yield important information about learning that cannot be captured directly by the computer. Alternatively, trained observers may be employed to record aspects of learning and performance that are impossible to obtain otherwise (Schofield & Evans-Rhodes, 1989).

Step 3: Select an Appropriate Design to Meet Defined Goals

Only after reviewing the goals and methods of the tutor and clarifying the goals of the evaluation study is it appropriate to select an evaluation design. Researchers involved in the evaluation of automated instruction have sometimes chosen to distinguish between formative and summative evaluations (e.g., Kearsley, 1983) of courseware. Generally speaking, formative evaluations have an *internal* control condition, and ask the question: How can the system be improved? Summative evaluations have an *external* control condition, and ask the question: How does the system compare to other systems or approaches?

Originally, the formative/summative distinction was used to distinguish between diagnostic (formative) evaluation during student learning versus outcome (summative) evaluation after student learning (Bloom, Hastings, & Madaus, 1971). This distinction was only later adapted to the purpose of categorizing evaluations of courseware during development versus after completion. However, we have found that distinction too restrictive for our purposes. Concerning types of evaluation studies, we prefer to think in terms of the continuum described earlier (see Fig. 4.2) and a set of general evaluation categories. In particular, we present five broad design categories that may be used for evaluation studies. These include (a) *within-system designs* asking how two or more alternative versions of a single tutor compare to one another; (b) *between-system designs* addressing the effectiveness of one tutor in relation to another in terms of teaching the same subject matter; (c) *benchmark designs* asking how a tutor fares in relation to some standard instructional approach; (d) *hybrid designs* constituting combinations of the above options; and (e) *quasi-experimental designs* representing any of the previous categories, but without random assignment of subjects to conditions. Such an approach is often necessary for true field studies, but should be undertaken with some care (see Campbell & Stanley, 1968). Although these five categories do not represent an exhaustive set, they are common and useful design types for evaluation studies.

Step 4: Instantiate Design With Proper Measures, Subjects, and Controls

The next step is to carefully plan the details of the design. Carefully consider the selection of dependent and independent measures, the number and type of subjects, and the appropriate control conditions.

Dependent Measures. We have found that a common problem in failed evaluation is poor selection, design, or implementation of dependent measures to assess knowledge and skill acquisition. The dependent measures should directly reflect both the goals of the ITS and the goals of the evaluation study.

We prefer to obtain a variety of dependent measures. Because ITS instruction is done on computers, it is cheap and easy to capture data on virtually everything that happens during instruction. Given the expense and trouble involved in building an ITS and implementing a large-scale evaluation, we choose to err on the side of gathering too much data. Besides, it is the nature of learning and instructional research that the apparent effectiveness of an intervention will depend, in large part, on how you measure performance. If you measure performance in a variety of ways, you are more likely to pick up treatment effects if they exist. Possible dependent measures 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 or decay.

Independent Measures. It is likely that the effectiveness of an instructional intervention will vary with individual characteristics of students. Individuals come to any new learning situation with varying knowledge, skills, and abilities. Common individual difference measures include general intelligence, grade point average, standardized aptitude test scores, 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). Consider collecting these kinds of measures in order to control for potential confounds in your experimental design (e.g., two schools with different mean IQs for enrolled students).

Control Conditions. One of the most common arguments in interpreting the results of evaluation studies is over the suitability of the control conditions. The choice of treatment condition(s), as well as the proper control condition(s), must be principled, based on a reasonable consideration of the claims you hope to make. For example, if you choose only to include a no-treatment control, you may only be able to claim "My intervention is better than nothing." This assumes, of course, that data support the claim. Certain rules of thumb may be applied to help eliminate control-condition problems in ITS evaluation research (see Shute & Regian, 1993, for more on this topic). One problem that may arise in ITS research is the creation of Hawthorne effects (i.e., treatment differences due only to the fact that one group, usually the tutor-instructed group, receives special attention). Hawthorne effects, like placebo effects, are easily obtained, and thus must be carefully avoided. In the ideal case, the only difference between the control and treatment condition should be the treatment itself. Confounding

difference you may want to avoid including differences in motivation, time-on-task, exposure to certain information, background characteristics, and so on.

Subjects. In addition to specifying rigorous control and experimental conditions, you will need to identify the right type and number of subjects that are needed in the study. In this regard, the most important considerations are the target population to which you would like to generalize, and the effect size that you expect to obtain.

For whom is the tutor intended? If the purpose of your ITS is to teach university graduate students a certain curriculum, and your test subjects come from an undergraduate population, you won't be able to accurately assess the effectiveness of your tutor on the target population.

As a rule of thumb, we believe evaluation studies looking for main effects of instructional treatments should use at least 30 subjects per condition. For aptitude-treatment variables (ATI) studies, using individual difference measures as independent variables, studies should use at least 100 subjects per treatment (Cronbach & Snow, 1977). This estimate 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. Most basic experimental design textbooks describe how to estimate the required sample size for picking up a treatment effect of some hypothesized magnitude. When performing these calculations, keep in mind the difference between statistical significance and real-world importance. With enough power, you can pick up very small treatment effects, even though the effect size may be too small to be of practical importance.

CONCLUDING REMARKS

In this chapter, we have described our general approach to research and development of intelligent tutoring systems. The approach was based on the fundamental belief that ITS research should be driven by learning theory and constrained by evaluation data. We further described and illustrated a principled progression from laboratory studies using artificial tasks (with high internal validity), to field studies of fully implemented ITS teaching real tasks (with high external validity). We believe that early in this progression it is appropriate to identify effective instructional interventions in controlled laboratory settings using carefully designed laboratory tasks. Interventions that appear promising in this context should be applied to real-world tasks in controlled laboratory studies and eventually in field studies. Finally, we presented four steps we believe are useful in

organizing the design of an evaluation study: (a) Delineate the goals and methods of the tutor, (b) define the goals of the evaluation study, (c) select the appropriate design to meet the defined goals, and (d) instantiate the design with appropriate measures, subjects, and controls.

It is important to note that even the most carefully designed evaluation study can fail during implementation due to incomplete logistical planning and preparation. Any evaluation effort has a multitude of details to attend to, and it is important to try to anticipate all of these in advance. Problems can be avoided with careful planning, training of personnel, and general preparation. For example, you can avoid a lot of problems by providing testing personnel with clear "scripts" and procedural checklists. You should also consider, in advance, the possible "worst-case" scenarios, such as what you would do if your hardware or software fails. These kinds of questions are best considered before the study begins (i.e., an ounce of prevention is worth a pound of cure).

If you succeed in carrying out a large evaluation study, you may be surprised at the difficulties involved in dealing with very large and diverse data sets. We recommend that you automate the storage, moving, recoding, and formatting of data as much as possible, and carefully check your automated procedures with dummy data sets having known distributions. 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.

We have found the evaluation of instructional interventions to be every bit as exciting as the development of these interventions. We are sometimes supported, sometimes humbled by data about how our instruction influences learning. Always, however, we benefit from the process.

ACKNOWLEDGMENTS

This chapter represents a synthesis and extension of some of our other work, especially Shute and Regian (1993) and Regian and Shute (1994). The research reported in this chapter was conducted by personnel of the Armstrong Laboratory, Human Resources Directorate, Brooks Air Force Base, Texas. The opinions expressed in this chapter are those of the authors and do not necessarily reflect those of the Air Force.

REFERENCES

Anderson, J. R., Boyle, C., & Reiser, B. (1985). Intelligent tutoring systems. *Science, 228*, 456–462.

Anderson, J. R., Farrell, R., & 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, & C. Luckhardt (Eds.), *Intelligent tutoring systems: Evolutions in design* (pp. 151–161). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.

Bonar, J., Cunningham, R., Beatty, P., & Weil, W. (1988). *Bridge: Intelligent tutoring system with intermediate representations* (Tech. Rep. No. 88-7). Pittsburgh, PA: Learning Research & Development Center, University of Pittsburgh.

Burton, R. R., & Brown, J. S. (1982). An investigation of computer coaching for informal learning activities. In D. Sleeman & J. S. Brown (Eds.), *Intelligent tutoring systems* (pp. 77–123). London: Academic Press.

Campbell, D. T., & Stanley, J. C. (1968). *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., & 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 Psychologia, 71*, 1–15.

Farquhar, J. (1992). *Loader*. Unpublished computer program, Brooks Air Force Base, Texas.

Farquhar, J., Shebilske, W. L., & Regian, J. W. (1992). *Dynamic graphical models during training*. Unpublished manuscript.

Kearsley, G. (1983). Computer-based training: A guide to selection and implementation. Reading, MA: Addison-Wesley.

Kimura, D. (1992, September). Sex differences in the Brain. *Scientific American*, 119–125.

Lesgold, A. M., Bonar, J., Ivill, J., & Bowen, A. (1989). An intelligent tutoring system for electronics trouble-shooting: DC-circuit understanding. In L. Resnick (Ed.), *Knowing and learning: Issues for the cognitive psychology of instruction* (pp. 66–93). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lesgold, A., Lajoie, S. P., Bunzo, M., & Eggan, G. (1992). A coached practice environment for an electronics troubleshooting job. In J. Larkin, R. Chabay, & C. Sheftic (Eds.), *Computer-assisted instruction and intelligent tutoring systems: Establishing communication and collaboration* (pp. 49–80). Hillsdale, NJ: Lawrence Erlbaum Associates.

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., & Soloway, E. (1988). Evaluating ITSs: The cognitive science perspective. In M. C. Polson & J. J. Richardson (Eds.), *Foundations of intelligent tutoring systems* (p. 209–242). Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Regian, J. W., & Shute, V. J. (1994). Basic research on the pedagogy of automated instruction. In T. de Jong, H. Spada, & D. M. Towne (Eds.), *The use of computer models for explication, analysis, and experiential learning* (pp. 121–132). New York: Springer-Verlag.

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* (pp. 238–243). Amsterdam, the Netherlands: IOS.

Shebilske, W. L., & Regian, J. W. (1992, October). *Video games, training, and investigating complex skills*. Paper presented at the annual meeting of the Human Factors Society, Atlanta, GA.

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 & V. J. Shute (Eds.), *Cognitive approaches to automated instruction* (pp. 15–47). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shute, V. J. (1993a). A comparison of learning environments: All that glitters . . . In S. P. Lajoie & S. J. Derry (Eds.), *Computers as cognitive tools* (pp. 47–74). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shute, V. J. (1993b). A macroadaptive approach to tutoring. *Journal of Artificial Intelligence and Education*.

Shute, V. J., & Gawlick-Grendell, L. A. (1993, August). *An alternative approach to learning probability: Stat Lady.* Proceedings of AI & ED 93, Edinburgh, Scotland.

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

Shute, V. J., & 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* (pp. 333–366). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shute, V. J., Glaser, R., & Raghavan, K. (1989). Inference and discovery in an exploratory laboratory. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences* (pp. 275–326). San Francisco: Freeman.

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

Shute, V. J., & Regian, J. W. (1993). Principles for evaluating intelligent tutoring systems. *Journal of Artificial Intelligence & Education*, 245–271.

Shute, V. J., Pellegrino, J. W., Hubert, L., & Reynolds, R. W. (1983). The relationship between androgen levels and human spatial abilities. *Bulletin of the Psychonomic Society, 26*(6), 465–468.

Skinner, B. F. (1957). *Verbal behavior.* Englewood Cliffs, NJ: 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.

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

Woolf, B. P. (1988). Intelligent tutoring systems: A survey. In H. Schrobe (Ed.). Exploring artificial intelligence (pp. 1–44). Palo Alto, CA: Morgan Kaufman.

5 Assessment of Intelligent Training Technology

Alan Lesgold
University of Pittsburgh

Over the past decade, there has been considerable research and development in applications of artificial intelligence to education and training (e.g., studies in Larkin & Chabay, 1992; Polson & Richardson, 1988; Psotka, Massey, & Mutter, 1988). In several cases, training systems have been produced that are receiving practical use (e.g., Anderson, 1990; Corbett & Anderson, 1992; Govindaraj, 1988). More commonly, so far, managers are starting to face decisions about whether a prototype research system has potential utility. In this chapter, I view the assessment of intelligent training systems from a long-term perspective, discussing the different kinds of decisions that require assessment of intelligent training technology and a number of specific assessment issues, considered in light of current theory and experience. In particular, I draw on experiences with the Sherlock coached practice environment for electronics troubleshooting (Lajoie & Lesgold, 1990; Lesgold, Lajoie, Bunzo, & Eggan, 1992; Lesgold, Lajoie, Logan, & Eggan, 1990).

IMMEDIATE EFFECTIVENESS VERSUS POTENTIAL

Technology assessment in the world of intelligent training systems must consider not only the effectiveness of a training system but also the likelihood that it can be assimilated by the organizations that could use it. This can be seen either superficially as a marketing problem or more deeply as a problem in changing schooling or training. In either case, though, it is not enough for a product to be effective; it also must either fit the existing organizational structure and available technology or be so attractive as to bring about adaptive changes that make it usable.