AL-TP-1991-0040

AD-A243 267

META-EVALUATION OF FOUR
INTELLIGENT TUTORING SYSTEMS:
PROMISES AND PRODUCTS

DTIC
ELECTE
DEC 12 1991
S D
D

Valerie J. Shute

HUMAN RESOURCES DIRECTORATE
MANPOWER AND PERSONNEL RESEARCH DIVISION
Brooks Air Force Base, TX 78235-5000

December 1991

Final Technical Paper for Period June 1990 – July 1991

91-17698

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5000

NOTICES

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

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

This paper has been reviewed and is approved for publication.

VALERIE J. SHUTE
Project Scientist

WILLIAM E. ALLEY, Technical Director
Manpower and Personnel Research Division

ROGER W. ALFORD, Lt Col, USAF
Chief, Manpower and Personnel Research Division

REPORT DOCUMENTATION PAGE

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE December 1991	3. REPORT TYPE AND DATES COVERED Final Technical Paper – June 1990 – July 1991

4. TITLE AND SUBTITLE
Meta-Evaluation of Four Intelligent Tutoring Systems: Promises and Products

6. AUTHOR(S)
Valerie J. Shute

5. FUNDING NUMBERS
PE – 61102F
PR – 2313
TA – T1
WU – 44

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Armstrong Laboratory
Human Resources Directorate
Manpower and Personnel Research Division
Brooks Air Force Base, TX 78235-5000

8. PERFORMING ORGANIZATION REPORT NUMBER
AL-TP-1991-0040

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

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

This paper examines two main promises of intelligent tutoring systems (ITS): (1) They will engender more effective and efficient learning in relation to traditional formats, and (2) they will reduce the range of learning outcome scores where a majority of individuals are elevated to high performance levels. Bloom (1984) has referred to these as the "two sigma problem"—to achieve two standard deviation improvements with tutoring over traditional instruction methods. Four ITS are discussed in relation to the two promises. These tutors have undergone systematic, controlled evaluations: (a) The LISP tutor (Anderson, Farrell, & Sauers, 1984); (b) Smithtown (Shute & Glaser, 1991); (c) Sherlock (Lesgold, Lajoie, Bunzo, & Eggan, 1990); and (d) The PASCAL ITS (Bonar, Cunningham, Beatty, & Weil, 1988). Results show that these four tutors *do* accelerate learning with no degradation in final outcome. Suggestions for improvements to the design and evaluation of ITS are discussed.

14. SUBJECT TERMS
aptitude
intelligent tutoring system
learning efficiency
learning environment
learning outcome
learning style

15. NUMBER OF PAGES
18

16. PRICE CODE

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

PREFACE

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 within are those of the author and do not necessarily reflect those of the Air Force. I would like to thank Pat Kyllonen, Dan Woltz, Wes Regian, and Bill Alley for contributing lucid comments on this paper. Correspondence concerning this paper should be addressed to Valerie J. Shute, Armstrong Laboratory, Human Resources Directorate, Brooks Air Force Base, Texas 78235-5000.

iii

META-EVALUATION OF FOUR INTELLIGENT TUTORING SYSTEMS: PROMISES AND PRODUCTS

SUMMARY

Intelligent tutoring systems (ITS) have been in existence for over a decade now. However, few controlled evaluation studies have been conducted comparing the effectiveness of these systems to more traditional instruction methods. This paper examines two main promises of ITS: (1) They will engender more effective and efficient learning in relation to traditional formats, and (2) They will reduce the range of learning outcome scores where a majority of individuals are elevated to high performance levels. Bloom (1984) has referred to these as the "two sigma problem"—to achieve two standard deviation improvements with tutoring over traditional instruction methods. Four ITS are discussed in relation to the two promises. These tutors have undergone systematic, controlled evaluations: (a) The LISP tutor (Anderson, Farrell, & Sauers, 1984); (b) Smithtown (Shute & Glaser, 1991); (c) Sherlock (Lesgold, Lajoie, Bunzo, & Eggan, 1990); and (d) The PASCAL ITS (Bonar, Cunningham, Beatty, & Weil, 1988). Results show that these four tutors *do* accelerate learning with no degradation in final outcome. Suggestions for improvements to the design and evaluation of ITS are discussed.

INTRODUCTION

Advances and innovations in the history of education have been scarce. Of the few instructional breakthroughs (e.g., Head Start program, "mastery learning"), none have conveyed more potential and excitement than the emergence of intelligent tutoring systems over a decade ago. For a long time, researchers have contended that individualized tutoring engenders the most effective and efficient learning for most people (e.g., Bloom, 1956, 1984; Burton & Brown, 1982; Carroll, 1963; Cohen, Kulik, & Kulik, 1982; Lewis, McArthur, Stasz & Zmuidzinas, 1990; Woolf, 1987). Intelligent tutoring systems (ITS) *epitomize* this principle of individualized instruction. Thus, by extension, the two main promises of ITS are they can: (1) Engender more effective and efficient learning in relation to traditional formats, and (2) Possibly reduce the range of learning outcome scores where a majority of individuals are elevated to high performance levels. These promises have been called the "two sigma problem" (Bloom, 1984). The goal is to achieve two standard deviation improvements with tutoring over traditional instruction methods.

For those of us concerned with teaching and learning, these promises of ITS are profound. Unfortunately, although such systems have been in existence for over 10 years now, their efficacy has been equivocal for several reasons: ITS are often designed by seat-of-the-pants engineering, lacking principled design standards, and abounding in "intuition" underlying the implementation of system components (e.g., Koedinger & Anderson, 1990; Norman, 1989). Furthermore, systematic, controlled evaluations of ITS are rare (Baker, 1990;

Littman & Soloway, 1988). The few ITS that actually have been evaluated in relation to other learning situations have shown evidence supporting the first promise (facilitating learning), but have shown little evidence supporting the second promise (reducing individual differences in outcome performance). I view this as encouraging, however, because new technologies usually do not fare well compared against proven methods (Baker, 1990).

Bloom (1984) identified problems associated with "proven," conventional teaching methods (e.g., a teacher presenting material in front of 30 people). He asserted that this format provides one of the least effective techniques for teaching and learning. As teaching becomes more focused and individualized, learning is enhanced. For example, when a teacher supplements a lecture with diagnostic tests to determine where students are having problems, then adjusts the lecture accordingly, this is called "mastery teaching." Students learning under this condition typically generate test results around the 84th percentile. Bloom further reported that students involved in "one-to-one tutoring," with human tutors, performed around the 98th percentile (2 standard deviation increase) as compared with traditionally-trained students (see Figure 1). These results were replicated four times with three different age groups for two different domains. Bloom thus provides evidence that tutoring is one of the most effective educational delivery methods available.
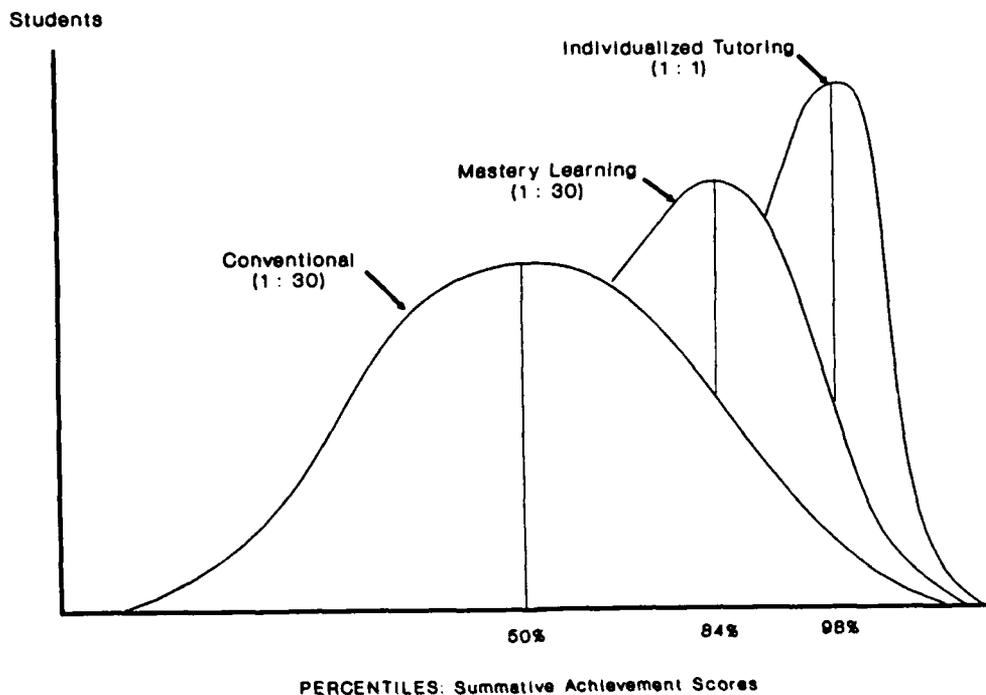

Figure 1. Distributions for Different Learning Conditions.

2

This paper evaluates the two promises of one-to-one tutoring as embodied in four ITS: (a) The LISP tutor (Anderson, Farrell, & Sauers, 1984); (b) Smithtown, an intelligent discovery world that teaches scientific inquiry skills in *the context* of microeconomics (Shute & Glaser, 1991); (c) Sherlock, a tutor for avionics troubleshooting (Lesgold, Lajoie, Bunzo & Eggan, 1990); and (d) The PASCAL ITS, teaching PASCAL programming skills (Bonar, Cunningham, Beatty, & Weil, 1988; Shute, 1991). Results from these evaluations will be discussed in relation to the success criteria ("promises") as well as to ITS design issues.

FOUR EVALUATIONS

The LISP tutor. Anderson and his colleagues at Carnegie-Mellon University (Anderson, Farrell, & Sauers, 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). 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 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 apparently successful in promoting faster learning with no degradation in outcome performance compared to traditional instruction.

In a third study using the LISP tutor to investigate individual differences in learning, Anderson (1990) found that when prior, related experience was held constant, two "meta-factors" emerged (i.e., factor analysis on factor scores). These two meta-factors, or basic learning abilities, included an *acquisition* factor and a *retention* factor. Not only did these two factors explain variance underlying tutor performance, they also significantly predicted performance on a paper-and-pencil midterm and final examination.

Smithtown. Shute & 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. In one study (Shute, Glaser & Raghavan, 1989), three groups of subjects were compared: 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 while 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 hours, 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, *per se*, but rather on general scientific inquiry skills, such as hypothesis testing.

Another study conducted with Smithtown (Shute & Glaser, 1990) explored individual differences in learning and showed that scientific inquiry behaviors relating to a hypothesis generation and testing factor were significantly more predictive of successful learning in Smithtown than a standard measure of general intelligence. The five relevant indicators comprising this factor accounted for 42% of the criterion variance while a measure of general intelligence (composite of four tests) accounted for only 1% of the variance. These findings suggest that, in this tutor, individual differences in learning outcome are not simply a function of general intelligence. Rather, specific behaviors, presumably trainable, are predictive of outcome performance.

Sherlock. "Sherlock" is the name given to a tutor which provides a coached practice environment for an electronics troubleshooting task (Lesgold, Lajoie, Bunzo, and Eggan, 1990). The tutor teaches troubleshooting procedures for dealing with problems associated with an F-15 manual avionics test station. The curriculum consists of 34 troubleshooting scenarios with associated hints. A study was conducted evaluating Sherlock's effectiveness using 32 trainees from two separate Air Force bases (Nichols, Pokorny, Jones, Gott, & Alley, in preparation). Pre- and post-tutor assessment was done using verbal troubleshooting techniques as well as a paper-and-pencil test. Two groups of subjects per Air Force base were tested: (1) subjects receiving 20 hours of instruction on Sherlock, and (2) a control group receiving on-the-job training over the same period of time. Statistical analyses indicated that there were no differences between the treatment and the control groups on the pretest (means = 56.9 and 53.4, respectively). However, on the verbal posttest as well as the paper-and-pencil test, the treatment group (mean = 79.0) performed significantly better than the control group (mean = 58.9) and equivalent to experienced technicians having several *years* of on-the-job experience (mean = 82.2). The average gain score for the group using Sherlock was equivalent to almost four years of experience.

PASCAL ITS. An intelligent programming tutor was developed to assist novice programmers in designing, testing, and implementing PASCAL code (Bonar, Cunningham, Beatty, & Weil, 1988). The goal of this tutor is to promote conceptualization of programming constructs or "plans" using intermediate solutions. A study was conducted with 260 subjects who spent up to 30 hours learning from the PASCAL ITS (see Shute, 1991). Learning efficiency rates were estimated from the time it took subjects to complete the curriculum. This measure involved both speed and accuracy since subjects could not proceed to a subsequent problem until they were completely successful in the current one. To estimate learning outcome (i.e., the breadth and depth of knowledge

and skills acquired), three criterion posttests were administered measuring retention, application and generalization of programming skills.

The PASCAL curriculum embodied by the tutor was equivalent to about 1/2 semester of introductory PASCAL (J.G. Bonar, personal communication, March 1990). That is, the curriculum equaled about 7 weeks or 21 hours of instruction time. Adding two hours per week for computer laboratory time (conservative estimate), the total time spent learning a half-semester of PASCAL the traditional way would be at least 35 hours. In the study discussed above, subjects completed the tutor in considerably less time (i.e., mean = 12 hours, SD = 5 hours, normal distribution). So, on average, it would take about three times as long to learn the same PASCAL material in a traditional classroom and laboratory environment as with this tutor (i.e., 35 vs. 12 hours).

While all subjects finished the PASCAL ITS curriculum in less time compared to time needed to complete the curriculum under traditional instructional methods, there were large differences in learning rates found at the end of the tutor. For these subjects (having no prior PASCAL experience), the maximum and minimum completion times were 29.2 and 2.8 hours, a range of more than 10:1. In addition, while all 260 subjects successfully solved the various programming problems in the tutor's curriculum, their learning outcome scores reflected differing degrees of achievement. The mean of the three criterion scores was 55.8% (SD = 19, normal distribution). The range from highest to lowest score was 96.7% to 17.3%, representing large between-subject variation at the conclusion of the tutor. In an attempt to account for these individual differences in outcome performance, Shute (1991) found that a measure of working-memory capacity, specific problem-solving abilities (i.e., problem identification and sequencing of elements) and some learning style measures (i.e., asking for hints and running programs) accounted for 68% of the outcome variance.

SUMMARY AND CONCLUSIONS

Intelligent tutoring systems have been around for over a decade now, so it is not unfair to ask: What is the verdict? Four ITS have been discussed in this paper which have undergone systematic evaluations. The results of the evaluations, as a whole, were very encouraging. The common finding is that learning efficiency with ITS was enhanced in relation to traditional instruction (e.g., LISP tutor, Smithtown, Sherlock, PASCAL tutor). That is, learning rates were accelerated whereby students acquired the subject matter faster from various ITS than from more traditional environments: (a) Subjects working with the LISP tutor learned the knowledge and skills in 1/3 to 2/3 the time it took a control group to learn the same material; (b) Subjects working with Smithtown learned the same material in 1/2 the time it took a classroom-instructed group; (c) Subjects working with Sherlock learned in 20 hours skills which were comparable to those possessed by technicians having almost 4 years of additional experience; and (d) Subjects learning from the PASCAL ITS acquired,

in 1/3 the time, equivalent knowledge and skills as learned through traditional instruction.

For learning outcome measures, the LISP tutor yielded the same (or in one study, 43% better) criterion scores than a control group not using the tutor. Results from the Smithtown analysis showed that subjects learned the same material as a classroom group, despite the fact that the tutor focused on the instruction of scientific inquiry skills, *not* the subject matter. And the outcome data from subjects using Sherlock showed increases in scores comparable to an advanced group of subjects and significantly better than a control group. In all cases, individuals learned faster, and performed at least as well, with the ITS as subjects learning from traditional environments.

The second promise, concerning a reduction in the range of outcome scores, was less straightforward to assess. The outcome variance of the Smithtown data was fairly restricted (mean = 72.7, standard deviation = 10). However, the posttest data from the Sherlock analysis showed a less restricted range in outcome scores (mean = 79, standard deviation = 17). And the results from the PASCAL ITS study similarly showed a relatively large variability on the final performance measure (mean = 55.8, standard deviation = 19).

As stated earlier, Bloom (1984) reported that individualized tutoring resulted in a two standard deviation increase in outcome performance for the majority of learners (see Figure 1). He suggested that treatment-effect size be computed as follows: (Mean of the experimental group—Mean of the control group)/SD of the control group. Ideally, we could compute effect size for each tutor evaluation and then make comparisons among tutors to see relative effectiveness in relation to Bloom's "2 sigma" goal. From the four tutors discussed in this paper, however, relevant data is only available from two of the four tutor studies: Sherlock and Smithtown.[1]

To illustrate how effect size may be computed and evaluated, data from the Sherlock study yields an effect size = (79.0 - 58.9)/19.7 = *1.02*. This implies a 1 standard unit increase in performance above the control group of subjects (84th percentile). Similarly, data from the Smithtown study (see Shute and Glaser, 1990) yields an effect size = (72.7 - 57.3)/15.9 = *0.97*, or a 1 standard unit increase in performance above the control group of subjects. Although both represent significant improvements of ITS over traditional on-the-job training or classroom instruction, these data still fall short of attaining "2 sigma" status.

The problem with finding evidence from the various tutor studies for a "reduction in range" may be due, in part, to the unreasonableness of the

---

[1] For the LISP tutor, no standard deviation data was cited for the control groups, and for the PASCAL study, no control group was included in the design.

second promise. In a footnote to his article, Bloom reported, "The control class distributions were approximately normal, although the mastery learning and tutoring groups were highly skewed" (1984, p. 16). Skewness and kurtosis data were, unfortunately, not presented. It may be more reasonable to evaluate ITS success in terms of another criterion: the reduction in the *correlation* between incoming knowledge and skills and learning outcome. That is, for a tutor to be really effective, it should be able to compensate for (or remediate) incoming cognitive weaknesses, and reinforce strengths to maximize learning outcome. In terms of this criterion, Anderson (1990) reported two basic learning abilities (acquisition and retention factors) that were highly predictive of LISP outcome performance. A possible enhancement to the design of this system would include adapting to differences in learning abilities. For instance, on-line measures could be monitored for rates of acquisition and retention of the subject matter. Then subjects demonstrating deficits in either of these areas could receive compensatory instruction, as needed. In another study, Shute and Glaser (1990) identified certain inquiry skills that significantly predicted outcome performance for microeconomics. While this system did monitor inquiry skills, not enough adaptability was built into the design (i.e., it was created to be more exploratory so the "coach" intervened infrequently). A suggested system modification would include increasing intervention as needed, rather than only after a fixed number of "buggy" behaviors. Finally, findings from the PASCAL tutor (Shute, 1991) showed that learning outcome was strongly predicted by a working memory factor, two problem solving abilities, and some learning behaviors. Information about an individual's working memory capacity could be used to vary instruction, such as teaching smaller chunks of relevant knowledge for those with less working memory capacity. Moreover, this tutor could benefit from the inclusion of supplemental instruction on relevant problem solving skills (e.g., part-task training of sequencing skills). In summary, by restructuring curricular materials (i.e., adapting to individuals' needs in real-time), learning from tutors could become less dependent on aptitudes, thereby providing everyone with a "fair shake" at learning. Obviously this is an hypothesis that can be empirically verified with more research.

What else could bring ITS closer to achieving these promises? A principled approach to the design and evaluation of ITS would be very helpful. One such approach is exemplified by a taxonomy of learning skills, developed and currently in use for both basic and applied research at the Armstrong Laboratory (see Kyllonen & Shute, 1989). This taxonomy defines four interactive dimensions: subject matter, learning environment, desired knowledge outcome, and learner styles. It is believed that interactions among these dimensions influence outcome performance. For example, it is misleading to generalize that one type of learning environment (e.g., exploratory) is best for all persons. Rather, aptitude-treatment interactions (Cronbach & Snow, 1977) are believed to occur where certain learner characteristics (aptitudes and styles) are better suited to certain learning environments for optimal outcome performance. Controlled studies using the taxonomy are needed in order to test various combinations of interactive dimensions in ITS designs. Then controlled studies comparing ITS versus traditional instruction are needed to calculate effect size

measures and be related back to Bloom's "2 sigma problem." The taxonomy provides a useful structure for comparing and evaluating tutors.

In conclusion, the evaluation results are, overall, encouraging. This is rather surprising given the enormous differences among the four tutors in design structure as well as evaluation methods. The findings indicate these four tutors do accelerate learning with no degradation in final outcome. In addition to measuring the reduction in range of learning outcome (as indicated by the second promise), it was suggested that a supplemental criterion would be the attenuation of correlation between outcome score with incoming aptitude measures.

Obviously, further basic research is needed to add more "psychology" and control into ITS designs. Rather than continuing to build tutors randomly, a more efficient route to the goal of optimizing ITS is to systematically alter the design of existing ones and evaluate the results of those changes in accordance with a principled approach (as is possible with the learning skills taxonomy). Many outstanding questions continue to beg for answers: What types of learners do better in what types of environments? Are certain domains better suited for specific instructional methods? When should feedback be provided, what should it say, and how is it best presented? How much learner control should be allowed? In conclusion, a principled approach to the design and evaluation of ITS is badly needed before we can begin to obtain answers to these questions. Only then can we reassess the "verdict" of ITS success.

REFERENCES

Anderson, J.R. (1990). Analysis of student performance with the LISP tutor. In N. Fredericksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

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*. 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.

Bonar, J., Cunningham, R., Beatty, P., & Weil, W. (1988). *Bridge: Intelligent tutoring system with intermediate representations* (Technical Report). Pittsburgh, PA: University of Pittsburgh, Learning Research & Development Center.

Burton, R.R., & 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.

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.

Koedinger, K.R., & Anderson, J.R. (1990). Theoretical and empirical motivations for the design of ANGLE: A new geometry learning environment. *AAAI Spring Symposium Series*. Stanford, CA: Stanford University.

Kyllonen, P.C., & Shute, V.J. (1989). A taxonomy of learning skills, In P.L. Ackerman, R.J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences*. New York: Freeman.

Lesgold, A., Lajoie, S.P., Bunzo, M., & Eggan, G. (1990). 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*. 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, CA: Stanford University.

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

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

Norman, D.A. (1989). *The psychology of everyday things*. New York: New York, Basic Books.

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

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*. 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*. New York: Freeman.

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

Woolf, B. P. (1987). A survey of intelligent tutoring systems. In Proceedings of the *Northeast Artificial Intelligence Consortium*, Blue Mountain Lake, NY.