User Modeling and User-Adapted Interaction 5: 1–44, 1995.

SMART: Student Modeling Approach for Responsive Tutoring

VALERIE J. SHUTE
Armstrong Laboratory, AL/HRTI, 1880 Carswell Avenue, Lackland AFB, TX 78236–5507, U.S.A.
email: vshute@colab.brooks.af.mil

(Received 20 January 1995; accepted in final form 30 April 1995)

Abstract. This paper describes a new student modeling paradigm called SMART. The premise is that a single, principled approach to student modeling, involving both theoretical and empirical methods, can render automated instruction more efficacious across a broad array of instructional domains. After defining key terms and discussing limitations to previous student modeling paradigms, I describe the SMART approach, as embedded within a statistics tutor called Stat Lady (Shute and Gluck, 1994). SMART works in conjunction with a tutor design where low-level knowledge and skills (i.e., curricular elements) are identified and separated into three main outcome types. Throughout the tutor, curricular elements with values below a pre-set mastery criterion are instructed, evaluated, and remediated, if necessary. The diagnostic part of the student model is driven by a series of regression equations based on the level of assistance the computer gives each person, per curriculum element. Remediation on a given element occurs when a subject fails to achieve mastery during assessment, which follows instruction. Remediation is precise because each element knows its location within the tutor where it is instructed and assessed. I end with a summary of results from two controlled evaluations of SMART examining the following research issues: (a) diagnostic validity, (b) individual differences in learning from Stat Lady, (c) affective perceptions of the tutorial experience, and (d) contributions of mastery and remediation to learning outcome and efficiency. Comments about related and future research with this paradigm are offered.

Key words: Aptitudes, cognitive diagnosis, learning outcomes, mastery learning, macroadaptation, microadaptation, remediation

Intelligent instructional systems teach more efficiently by individualizing instruction (e.g., Regian and Shute, 1992; Shute and Psotka, in press; Wenger, 1987). A central component within these systems, as well as an important research area in the field, focuses on modeling a learner to support this individualization (i.e., developing a valid student model). While many approaches to student modeling exist, few modeling approaches, to date, have attempted to span a broad range of instructional domains. This paper describes a powerful, general, and comprehensive student modeling paradigm (Student Modeling Approach for Responsive Tutoring – SMART) with the goal of enhancing the efficacy of automated instructional systems. The premise is that a single, principled approach to student modeling, involving both theoretical and empirical methods, can render automated instruction more efficacious across an array of instructional domains.

1. Introduction

1.1. Overview

SMART relates to the dynamic interplay between cognitive diagnosis and remediation, thus it is broader than student modeling approaches that focus solely on diagnosis (e.g., Martin and VanLehn, 1993). Furthermore, SMART is unique in that it not only models *evolving* knowledge and skills (domain specific) for purposes of microadaptation, it also assesses *incoming* abilities (general and specific cognitive aptitudes) as predictors of subsequent learning, and indicators of suitable instructional environments for macroadaptation. For instance, knowing that an individual has low working-memory capacity suggests smaller units of instruction (Shute, 1991), and knowing that someone possesses an exploratory learning style suggests a less pedantic, more open type of learning environment (Shute, 1993a). Ongoing aptitude-treatment interaction (ATI) research is already providing information about initial states of learners that can be mapped onto optimal instructional environments (e.g., Shute, 1992; Shute, 1993b; Shute and Gawlick-Grendell, 1994; Swanson, 1990). Further studies in this area will enhance our ability to tailor, or optimize, instruction. Finally, whereas most other approaches focus on single outcome types (e.g., model-tracing for procedural skill), SMART models a range of outcome types, including: symbolic knowledge (SK), procedural skill (PS), and conceptual knowledge (CK).

1.2. Organization

In subsequent sections of this paper, I define key terms, then describe the SMART approach to student modeling, followed by a description of, and results from, two studies assessing its efficacy in predicting outcome performance in descriptive statistics and the impact on learning. I conclude with comments about current and future research with this paradigm.

2. Definition of Key Terms

The ability to diagnose student errors and tailor remediation based on the diagnosis represents the critical difference between intelligent and merely clever automated instructional systems. The working definition of intelligence used in this paper is that a system must be able to accurately diagnose students' knowledge structures, skills, and aptitudes using principles, rather than pre-programmed responses, to decide what to do next, and then adapt instruction accordingly (e.g., Clancey, 1986; Shute, 1992; Shute and Psotka, in press; Sleeman and Brown, 1982). Thus, I define intelligence as both diagnosis and remediation, working in concert.

A student learns from an intelligent tutoring system (ITS) primarily by solving problems – ones that are appropriately selected or tailor-made and serve as good learning experiences for that student. The system starts by assessing what the student already knows, the *student model*. The system concurrently must consider what the student needs to know, the curriculum. Finally, the system must decide what curriculum element (unit of instruction) ought to be instructed next, and how it should be presented. From all of these considerations, the system selects, or generates, a problem, then either works out a solution to the problem (via the domain expert), or retrieves a prepared solution. The system then compares its solution, in real-time, to the one the student has prepared, and performs a diagnosis based on differences between the two. Feedback is offered by an instructional system based on considerations such as how long it's been since feedback was last provided, whether the student already received some particular advice, and so on. After the feedback loop, the program updates the student model, and the entire cycle is repeated, starting with selecting or generating a new problem. Not all intelligent tutoring systems include these components, and the problem-test-feedback cycle does not adequately characterize all systems. However, this generic depiction does describe many existing systems.

The current assortment of student modeling approaches (embedded within intelligent tutoring systems) represent different philosophical as well as practical influences in their design. At a global level, the standard approach to building a student model involves representing emerging knowledge and skills of the learner. When the computer responds to updated observations with a modified curriculum that is minutely adjusted, this is called *microadaptive* modeling. In this case, instruction is very much dependent on individual response histories during tutoring sessions. Microadaptive modeling focuses primarily on student knowledge that is specific to the task domain, although some more general problem-solving and learning strategies may be modeled. An alternative, *macroadaptive* approach (Shute, 1992; Snow, 1990) involves assessing students' knowledge and skill prior to their use of the tutor, and focuses mainly on general, long-term aptitudes, such as working-memory capacity, associative learning ability, and impulsivity. In fact, many have argued that incoming knowledge is the single most important determinant of subsequent learning (e.g., Alexander and Judy, 1988; Dochy, 1992; Glaser, 1984). Combining these two approaches enables the curriculum to adapt to both momentary and persistent performance information, and to both domain-specific knowledge and general aptitudes (see Shute, 1993a, b).

3. Limitations of Prior Student Modeling Research

Historically, a common feature of student modeling research has been on *implementing*, not evaluating, the respective models. Therefore, there is limited data supporting a particular paradigm's validity. This paucity of empirical data may be remedied by more systematic research in the area. One suggestion is to begin a coordinated stream of methodical research and development, altering specific features of existing systems and evaluating the results of those changes in accordance with a principled approach. According to Self (1989), "Once a sounder foundation for ITS has been specified, it becomes possible to identify the elements of a theory of ITS. These elements lie within (formal) AI, in areas such as belief logics, reason maintenance, meta-level architectures, and discourse models – areas from which ITS research has been divorced" (p. 244).

In addition, there is a large cost associated with incorporating a student model into a tutoring system. This raises two important questions: (1) How much, and what kind of, information about a learner is required to adequately diagnose knowledge and skill acquisition and subsequently tailor instruction to the learner's needs? (2) What is the *payoff* of increasing a system's adaptability? Sleeman (1987) has argued that "...if one takes seriously the findings of the ATI work of Cronbach and Snow (1977), it would appear that there is little likelihood of producing instruction that is uniquely individualized" (p. 242). The key word in this statement is "uniquely." An exhaustive characterization of a learner would probably not warrant the effort and expense in terms of increases in final outcome. But the empirical question remains: How much is enough? And more generally, does inclusion of a student model in a tutor enhance overall learning outcome or efficiency? There is equivocal evidence in the literature concerning these issues. In some cases, researchers have reported no advantage of error remediation in relation to learning outcome (e.g., Bunderson and Olsen, 1983; Sleeman, Kelly, Martinak, Ward and Moore, 1989), whereas in others, an advantage has been reported for more personalized remediation (e.g., Anderson, 1993; Shute, 1993a, b; Shute and Regian, 1993; Swan, 1983).

Another question asks whether the student model is even the right framework around which to build good learning systems. Derry and Lajoie (1993) presented several reasons why the student modeling paradigm may be problematic. Among the more compelling reasons cited were that: (a) In complex domains, the student model cannot specify all possible solution paths, nor determine all possible "buggy" behaviors; (b) Reflection and diagnosis should be performed by the student, not the tutor; and (c) Model-tracing is only applicable to procedural learning, but the focus should be on critical thinking and problem solving.

The approach to student modeling taken in this paper (SMART) does not attempt to delineate all possible solution paths, buggy behaviors, or misconceptions. Rather, the information about curriculum elements (CEs), derived from cognitive task analysis and arranged in inheritance hierarchies, provides a basis for inferences about what knowledge and skills have been acquired, and to what degree. Furthermore, the intention of SMART is to not only model procedural skill acquisition, but also symbolic and conceptual knowledge acquisition. Finally, SMART is currently undergoing rigorous and controlled evaluations to ascertain the efficacy of its various features as well as its ability to promote learning across a variety of domains. Before presenting the actual framework underlying the new approach, I will present its theoretical basis.

4. Foundation of Student Modeling Framework

Student modeling paradigms emphasize distinct kinds of knowledge and skill outcomes; some specialized for modeling procedural skill acquisition (e.g., model tracing) and some more suitable for modeling conceptual knowledge (e.g., progression of mental models). The knowledge and skills represented within SMART were intended to be sufficiently comprehensive to give the model a wide range of applicability compared to existing models.

In general, the outcome of learning refers to any change within an individual's knowledge structure that results from a learning situation. Outcomes of learning can be quite diverse, differing in magnitude (e.g., learning a simple fact versus a complex technical skill) as well as content area (e.g., affective and social skills, motor skills, procedural knowledge). The specific outcomes included in SMART were derived from Anderson's ACT* theory of learning, as well as research conducted at the Armstrong Laboratory (see Anderson, 1983, 1987; Kyllonen and Christal, 1989; Kyllonen and Shute, 1989). These theories of cognitive skill acquisition posit two primary kinds of knowledge and skill outcomes, declarative and procedural, each arrayed along a continuum, from simple to complex. For a more detailed description of learning outcomes, see Shute (1994).

4.1. Declarative Knowledge

A *proposition* is the basic unit of information underlying declarative knowledge outcomes. It is represented by a single, isolated postulate. A collection of related postulates, constructed from experiences in the world, comprises a *schema,* defined as an interconnected set of propositions representing a situation. Schemas form the basis for comparing and interpreting incoming data. They also shape individuals' expectations, and hence, what is perceived. But schemas, based on prior knowledge and beliefs, can lead to erroneous inferences if the foundation is deficient or contains misconceptions.

4.2. Procedural Skill

While declarative learning outcomes relate to knowledge *about* something, procedural learning outcomes relate to knowledge of *how to do* something, as well as the ability to do it. A *rule* is the basic unit of action underlying procedural skill outcomes. Rules are typically represented by condition-action pairs. The condition may be defined as the "if" part of a rule, while the action may be defined as the "then" part, consisting of the associated steps of some procedure. The next level of procedural outcome is a *skill,* defined as a collection of related rules. A skill may be cognitive, motor, or even social or creative. Finally, a skill may become *automatic* after considerable practice applying that skill in many and varied situations. Eventually, an automatic skill requires little or no conscious effort. For instance, after years of practice driving a car (involving a complex coordination of skills), one can drive the car in traffic while listening to the radio and planning the day's activities. The execution of this procedure is almost unconscious, compared to the step-by-step manner of invoking procedures, outlined above.

4.3. Mental Model

Both declarative knowledge and procedural skill are believed to influence the formation of *mental models* – highly organized sets of concepts and rules relating them together. A mental model represents an integrated system, and is structured hierarchically, where different levels of analysis are possible. At each level of analysis, one can know: information about component parts, how they are connected, and how the system functions as a whole.

Similar to the above learning outcome distinctions, SMART makes a fundamental distinction between declarative knowledge and procedural skill. To simplify things, I've changed some of the terminology. For instance, the most rudimentary form of knowledge in SMART is called "*symbolic knowledge*" (SK) corresponding to the definition of propositional knowledge above (i.e., knowing about something), and to the more formal definition of something that represents something else by association or convention; the latter embodying symbols and formulas. Two examples of symbolic knowledge include: (a) the symbol "Σ" means "the sum of," and (b) the formula for computing a proportion from a frequency distribution = frequency/total sample size = f/N. The definition of "*procedural skill*" (PS) in SMART is the same as described above; namely, being able to apply some rule(s) in the solution of a problem. Some examples include: (a) arranging a set of numbers in ascending order, and (b) computing the variance from a set of data. Finally, "*conceptual knowledge*" (CK) in SMART maps onto the previously described mental model involving higher-level relationships among concepts, schemas, and rules relating them together. Examples of conceptual knowledge include: (a) knowing several ways to characterize the central tendency of a given distribution (e.g., mean, median, mode), and how they are related, and (b) understanding the notion of variability within a set of data, and how that relates to the specific distribution of data.

4.4. Knowledge and Skill Acquisition

In a well-designed curriculum, acquisition of the most basic form of knowledge (symbolic) typically, but not always, precedes either procedural skill or conceptual knowledge. In other words, SK comprises the building blocks for either the development of rules and skills, or more advanced concepts. In the first case (SK → PS); knowing the formula for computing variance is necessary in order to subsequently apply that formula in computing the actual variance. In the latter example (SK → CK), it is possible to bypass PS and achieve a functional understanding of a concept, but not be able to actually perform some procedure. For example, one can know that "variance" refers to the degree of dispersion in a distribution of data without being able to apply the formula: $[\Sigma(X - M)^2]/N - 1$. The more typical case of knowledge and skill acquisition, however, progresses from: SK → PS → CK. This hypothesized ordering of outcome types is based on current learning theory (e.g., Anderson, 1987; Kyllonen and Christal, 1989; Kyllonen and Shute, 1989). Finally, increased conceptual understanding can feed back to influence procedural skill (PS ↔ CK), in a top-down manner (e.g., VanLehn, 1990; White and Frederiksen, 1987).

In addition to representing students' specific learning outcomes (i.e., emerging knowledge and skills), I've suggested that student models can benefit by considering information about learners' general aptitudes. The basic idea is that learning outcomes can reflect differences in incoming aptitudes, such as associative skill, working-memory capacity, or reflectivity. The processes responsible for declarative knowledge outcomes appear to be chiefly associative (e.g., Shute, 1992; Kyllonen and Tirre, 1988) with some inductive reasoning required for the acquisition of complex schemas and mental models (Shute, 1994). The learning processes responsible for procedural skill acquisition involve primarily working-memory capacity and, depending on the domain, information processing speed (Shute and Kyllonen, 1990).

As part of my research on student modeling, I am currently contrasting the effectiveness of using general aptitudes and domain-specific knowledge and skills to predict and improve training/learning performance. Previous research in this area has demonstrated that various tests of basic cognitive abilities (e.g., CAM 4.0 tests, Kyllonen, 1994; Kyllonen, Christal, Woltz, Tirre, Shute and Chaiken, 1990) can accurately predict training performance (e.g., Shute, 1991; Shute and Kyllonen, 1990) and help assign students to training environments that fit their aptitudes (e.g., Shute, 1992; Shute, 1993a, b). However, information on student aptitudes is rarely employed in current instructional systems, which primarily model domain-specific student knowledge and skill. By explicitly comparing the usefulness of information concerning general aptitudes and domain-specific knowledge, I hope to provide concrete evidence that will enable tests of general cognitive abilities to be more widely used as part of intelligent instructional software. I will now present a student modeling framework designed to aid in the development and comparison of different modeling techniques by providing a standard formalism.

5. Smart Framework

Dillenbourg and Self (1992) outlined a two-dimensional framework and notation for student modeling. Their vertical dimension distinguishes among learner behavior, behavioral knowledge, and conceptual knowledge. This is crossed with the second dimension reflecting the representation of that knowledge – by the learner, the system, and the system's representation of the learner. I modified their original framework slightly (see Figure 1) to represent specific knowledge and skill

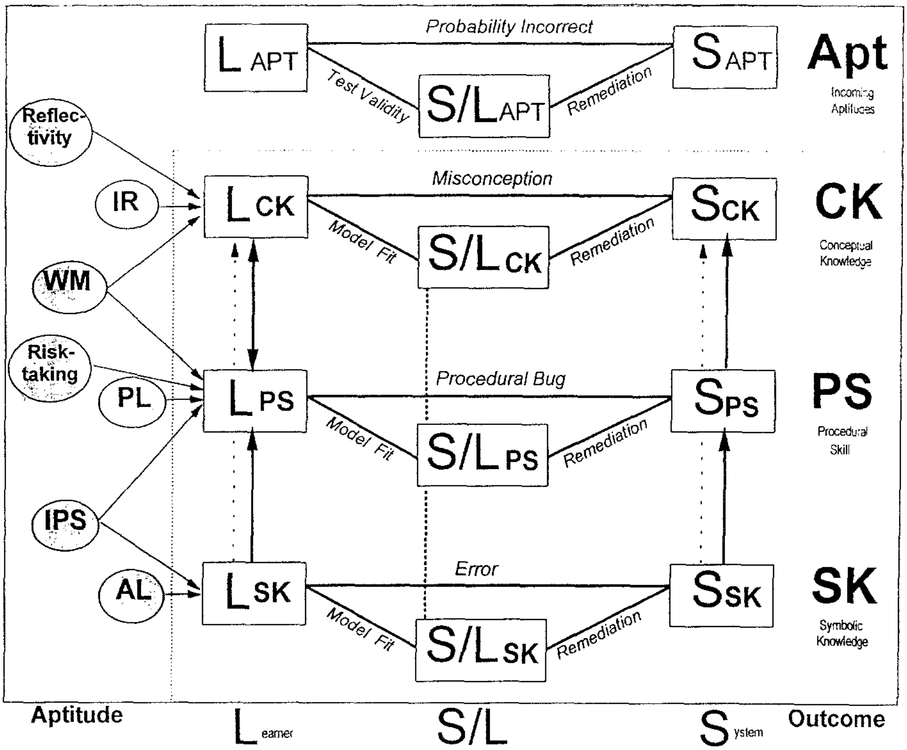


Fig. 1. SMART framework for student modeling (adapted from Dillenbourg and Self, 1992).

types required during the learning process (i.e., symbolic knowledge, **SK**; procedural skill, **PS**; and conceptual knowledge, **CK**), rather than overt behaviors. The horizontal axis remains basically the same (i.e., learner's representation of the knowledge or skill, **L**; the system's representation of the learner's knowledge, **S/L**; and the system's/expert's representation of the knowledge or skill, **S**). This modified framework reflects the standard microadaptive approach to student modeling.

5.1. Aptitudes

I also include in the framework (shown in Figure 1) a fourth element reflecting aptitudes (e.g., inductive reasoning skill – IR, working memory capacity – WM, procedural learning skill – PL, information processing speed – IPS, associative learning skill – AL, as well as various personality indices, such as reflectivity and risk-taking). This dimension is added to accommodate individual differences among learners (for macroadaptation), as well as provide a basis for representing the differential cognitive requirements of problems or tasks. This figure contains only a few aptitudes, for illustrative purposes. For example, SK acquisition is more a function of AL and IPS, while CK acquisition depends more on IR and WM. Connections between aptitude and outcome measures are based on findings reported in the literature; more aptitudes can be added later as research continues in this area.

Although not currently shown in the figure, I plan to incorporate a correlated time dimension because, as learning progresses, cognitive demands on the individual vary. For example, working-memory capacity and associative learning skills play an important role early in the learning process in determining the degree of new knowledge and skill acquisition. But over time, these become less important, and other cognitive factors gain importance (e.g., Ackerman, 1988, 1992; Woltz, 1988). Specifically, two findings from Ackerman's (1992) theory of skill development are relevant to the present framework: (1) overall task performance is predicted by general abilities and broad, task-specific abilities during acquisition, and (2) these correlations diminish, and correlations with perceptual speed and perceptual/motor abilities increase, as skill develops. Finally, various aptitudes are likely to be better/worse predictors of successful performance across different domains.

5.2. Interpreting the Framework

This framework is intended to function as follows. Suppose you were developing a computer program to teach statistical topics, and you were developing the part of the curriculum instructing the computation of a measure of central tendency – the mean (a procedural skill). Your system would begin by introducing relevant symbolic knowledge (e.g., Σ, X, N, ΣX/N), then would require learners to solve problems to demonstrate their acquisition of this new knowledge. Differences between the learner's (L_{SK}) and the system's representation (or understanding) of the knowledge (S_{SK}) would show up as *errors* on specific problems, such as failing to recognize the denominator (N) as being a part of the final formula (see bottom section of Figure 1, above). Next, the system directs students to apply their new knowledge by actually computing the mean from a set of data (procedural skill). Disparities between the learner's actual solution process (and product) (L_{PS}) and the system's representation of the correct procedure (S_{PS}) would be apparent in *procedural bugs*. Finally, learners are taught (and induce) conceptual information, like the mean represents the arithmetic average of a set of data, and relates to other measures of central tendency, such as the location of the mean in relation to the median and mode in a skewed distribution. Discrepancies between the learner's (L_{CK}) and system's (S_{CK}) conceptual representations can reflect fundamental *misconceptions*.

In general, the degree of similarity/dissimilarity between the learner's actual (L), and the model's inferred (S/L) representations, reflects the goodness/poorness of fit of the model. Next, discrepancies that exist between the system's (S) and model's (S/L) representations suggest appropriate remediation – what to instruct next, and how. For example, if the computer diagnoses a learner as having difficulty with specific PS acquisition (S/L_{PS}), then possible remediation could include specially-created problems for additional practice on certain elements.

The probability that a learner will get a specific problem correct or incorrect is also influenced by the degree to which a learner's aptitudes (L_{APT}) are matched to the cognitive requirements of a specific task (S_{APT}). That is, if the working-memory demands and requirements for task X were high (e.g., introduction of a new curriculum element requiring the integration of diverse knowledge), and the learner's actual working-memory capacity was low, the predicted probability of success would be low. Cognitive process measures (aptitudes) can easily be assessed. As mentioned in Section 4.4, the CAM 4.0 battery of computerized tests (Kyllonen et al., 1990) measures six different aptitudes (i.e., working-memory capacity, inductive reasoning skills, information processing speed, associative learning skills, procedural learning skills, and general knowledge) in each of three domains (quantitative, verbal, and spatial). The battery has been widely tested and validated, and individual tests can be extracted for different purposes. For instance, if you were teaching a three-dimensional navigational task and wanted to know an individual's spatial aptitude, you could select certain tests for administration (e.g., tests measuring spatial working memory). Differences between the learner's actual aptitude (L_{APT}) and the system's representation of the learner's aptitude (S/L_{APT}), denote the validity of the aptitude measure (e.g., CAM test validity). Moreover, discrepancies between the system's representation of the learner's aptitude (S/L_{APT}) and the system's aptitude requirements (S_{APT}) indicate what kind of remediation is required (e.g., decomposing the current task into more manageable units for a learner diagnosed as having low working-memory capacity).

5.3. Macroadaptation

To illustrate one way that the macroadaptive approach is intended to work, consider the results from the following study investigating learning from a flight engineering tutor (Shute, 1993b). Although there was no main effect on learning outcome due to treatment (i.e., abbreviated vs. extended practice conditions), I did find a significant aptitude-treatment interaction between an aptitude profile and treatment condition. Subjects with a lot of general knowledge (GK) but low working-memory (WM) capacity learned significantly more if assigned to the abbreviated, rather than the extended condition. Their broad GK provided a foundation that allowed for the extraction and interweaving of information from relatively few exemplars, and their low WM capacities weren't taxed within the abbreviated environment. However, these same subjects, assigned to the extended practice condition, performed poorly, probably due to boredom, fatigue, and/or working memory overload. The other type of subject was characterized by high WM and low GK. They learned more from the extended than the abbreviated practice condition. That is, they had the capacity, as well as the knowledge need, to profit from the extra practice afforded by the extended condition. The abbreviated condition was insufficient for these learners to induce the relevant principles.* This finding suggests that individuals characterized by specific aptitude profiles are differentially suited to instructional environments. In practical terms, one use of the macroadaptive approach involves *a priori* assessment of learners' WM and GK, then subsequent placement into the appropriate environment. This type of modeling complements the more common, microadaptive approach, which bases instructional choices only on student knowledge and performance within the training domain, and ignores information about students' more general abilities.

A second application of the macroadaptive approach will be discussed later in this paper following the presentation of the details of the student modeling updating scheme. Basically, it involves using modified updating heuristics for student model values, dependent on particular aptitude profiles.

I will now describe how SMART was integrated into a non-intelligent instructional system. The example of a statistics tutor is interesting because the domain encompasses (and attempts to instruct) a variety of knowledge and skill outcome types.

6. Employing the Smart Framework – Stat Lady

Stat Lady is the name of a series of computerized experiential learning environments teaching topics in introductory statistics, such as probability and descriptive statistics (Shute, Gawlick-Grendell, Young and Burnham, in press; Shute and Gluck, 1994). The design of the program reflects the theoretical postulates that learning is a constructive process, enhanced by experiential involvement with the subject matter that is situated in real-world examples and problems. The module focused on in this paper teaches descriptive statistics, and currently exists in two forms: non-intelligent and intelligent. The two versions of the tutor are the same, except for the presence or absence of the student model. Thus, they share a lot of features, including: (a) a humorous and experiential interface; (b) an identical set of curriculum elements, or CEs; (c) a pool of whimsical problem sets, per CE; and (d) a three-level feedback design. Nevertheless, the versions differ in two important respects: (a) subjects must complete *all* CEs in the non-intelligent version, but only those CEs identified as being not (or partly) known in the intelligent version; and (b) in the non-intelligent version, learners have to solve at least two (and may elect to solve up to five) problems before mastery is presumed and they go on to the next CE. In the intelligent version, however, learners may solve as few as zero**, but as many as five or more problems to demonstrate mastery of a given CE. Following performance assessment, learners may either be advanced to a new CE, receive remedial instruction, or continue to receive practice on the current CE.

* Note: the subjects who were characterized as being (a) high WM, high GK, and (b) low WM, low GK performed well and poorly on the outcome measures, respectively.

** Solving "zero" problems for a specific CE means that a learner solved the corresponding pretest items at a mastery level of performance.

6.1. Non-Intelligent Version of Stat Lady

Stat Lady's curriculum is presented in a fixed order to all learners, although learners are free to move forward and backward through the curriculum, at will. Moreover, students must solve at least two problems within each problem set,* which they select from a pool of potential problems, sufficiently varied in topic to be of interest to all. Further, they are required to answer each question within a given problem correctly before moving on to subsequent problems. If a learner is unable to answer a particular problem, Stat Lady intervenes with assistance.

The curriculum focused on in this paper is descriptive statistics, assembled from the results of a cognitive task analysis performed by two subject matter experts (for validity and reliability), and consisting of hundreds of curriculum elements (CEs). These CEs were arranged in a linear order, from simple to more complex concepts and skills, spread across three 4-hour sub-modules: (a) Data organization and plotting, (b) Measures of central tendency, and (c) Measures of variability. Students receive instruction and solve problems within each section, and Stat Lady provides specific feedback about each topic under investigation. The system also allows learners to engage in elective "extracurricular" activities, such as viewing items in the on-line Dictionary and Formula Bank, playing around in the Number Factory, or using the Grab-a-Graph tool.

The non-intelligent version of Stat Lady (SL) is often clever, but not intelligent in the classical sense (see working definition of computer-tutor intelligence, Section 2). The type of pedagogy embedded in the system may be called "near mastery learning." That is, relevant concepts and rules are presented, then SL poses various problems for students to solve in order to demonstrate comprehension of the curriculum elements. If a learner gets a problem wrong, SL provides progressively more explicit feedback. If the student fails to solve it correctly after three attempts, he or she is given the correct answer, and told to enter that answer. For example, suppose the current problem was assessing the SK element: "Formula for Computing the Mean." If the student (using the "Equation-Builder" tool) entered "ΣX" as the correct solution, Stat Lady would respond with level-1 feedback: "Your answer is almost correct. Please try again." If the student then created the solution "$\Sigma X/f$," level-2 feedback would inform the student, "Your numerator is correct, but your denominator is wrong (Do you really want to divide by the frequency?). Please try it again." Finally, if the learner then erroneously entered, "$\Sigma X/Xf$," Stat Lady would respond with level-3 feedback, "The correct answer for this problem is $\Sigma X/N$. Please create this solution now." The system thus *presumes* that the learner has actually acquired that concept or skill after explicitly being told. But this could be an erroneous presumption, with problems arising later on when a student tries to learn higher-level skills that have only partially-learned subskills as components.

* A "problem set" contains problems that relate to several previously-instructed CEs. A problem set consists of multiple questions, each one relating to (assessing) a particular CE.

The program begins by identifying a particular topic for instruction from an ordered list of curriculum elements. After presenting the topic in broad strokes, Stat Lady quickly illustrates the concept with a real-life, humorous example, to enhance memorability. Following instruction, learners have to apply the concept or skill in the solution of related problems. For problems that require data manipulation, they obtain their own unique data set from the Number Factory (an activity that's analogous to data collection), where they set their own parameters, such as the sample size, type of distribution, and minimum and maximum values. Then, if the learner successfully solves the problem on the first attempt, he or she is congratulated with positive auditory and textual feedback, and moves directly on to the next part of the exercise. If a student is having difficulty (e.g., enters one or more incorrect responses), each response is followed by the increasingly explicit feedback (and encouragement), mentioned above.

Even a quick illustration such as this makes the inherent problem in this approach apparent. While some students complete a problem set with a firm understanding of the knowledge or skill being taught, others may emerge with incomplete understanding. Ideally, all students would walk away from the tutor having demonstrated complete understanding. Hence, the primary motivator underlying SMART's development was: How can Stat Lady be modified to yield more guarantees of successful knowledge and skill acquisition?

6.2. Intelligent Version of Stat Lady

The intelligent version of Stat Lady (ISL) includes a student model based on the framework discussed earlier. Actually, three student models are implemented, one each for SK, PS, and CK (but in this paper, I'll refer to them collectively as "student model" because they apply similar updating heuristics).

6.2.1. *Classification and Organization of CEs*

The first step in rendering Stat Lady intelligent was to classify each CE from the cognitive task analysis into one of three categories: symbolic knowledge, procedural skill, or conceptual knowledge. This classification was initially done on the first descriptive statistics module (i.e., data organization and plotting), independently performed by three persons, and inter-rater reliability was very high (> 0.90). Simple operational definitions were employed as the basis for this sorting activity: SK = any symbol or formula, PS = application of a formula or rule, or performing a specific action within the tutor, and CK = definitions of, and relations among, the statistical concepts. A listing of all relevant CEs and associated descriptions may be found in Appendix A. Of the 77 CEs comprising data organization and plotting, there were only a handful of cases that required some debate as to category, but these were easily resolved with brief discussion.

The elements resulting from the classification were then arranged in three hierarchies (SK, PS, and CK) relating higher-level knowledge and skills to one another, and to successively lower-level knowledge and skills. These inheritance hierarchies have direct implications for SMART's updating scheme. That is, the acquisition of a closely-related, previously-learned CE (e.g., one sharing the same parent as the current CE) should predict the acquisition of its sibling. More distal parts of the inheritance hierarchy can also be considered when predicting student model values of particular CEs.* Appendix B shows a hierarchy of the tutor's PS elements.

Collectively, these CEs (in three learning outcome categories) represent the knowledge and skills I wanted learners to walk away with at the conclusion of their learning experience. But how does CE-acquisition information become translated into numbers or probabilities denoting a learner's attainment of mastery?

6.2.2. *General Operation of the Student Model*

The student model operates as follows: (a) Each CE is initially evaluated following completion of the on-line pretest, and these initial knowledge/skill estimates are passed to the student model for initialization; (b) Each CE is subsequently evaluated during problem solution and question-answering within the tutor; (c) CEs are linked to other CEs in inheritance hierarchies (i.e., parent, siblings, children) to provide student model updating information based on related CE values; and (d) Because each CE knows its exact location in the tutor (i.e., where it is instructed and evaluated), remediation is precise and efficient. These activities map onto the four main routines that drive the student model, whereby the data are managed by a straightforward array of records, and each array element maps onto an individual curricular element. All information for that CE is maintained within the record (see Table 1) such as its outcome type and location in the tutor where it's instructed and assessed.

6.2.3. *Student Model: Initialization*

As indicated, initial values for SMART are obtained from students' performance on a comprehensive, on-line pretest designed to assess incoming knowledge of *all* curriculum elements resulting from the cognitive task analysis (separated into SK, PS, and CK items). The pretest contains at least two test items per CE (for reliability), the scores of which are combined to yield an average initial CE value. In addition, many test items contain multiple parts, so these items can be scored with partial credit.** Thus, pretest scores (the initial "best guess" of incoming knowledge/skill) can range from 0 (completely incorrect) to 1 (completely correct), with intermediate values reflecting *degree* of incoming mastery. After the computer

* Currently, SMART does not employ hierarchical information in its updating scheme, but I plan to add this capability and then test its contribution to predictive validity. For more on this topic, see Section 9.2.

** I also created an isomorphic posttest containing the same CE's, but with different questions.

TABLE I. Data contained within each CE record

Curricular element number (e.g., CE-1)
Curricular element description (e.g., *Sort Data by Ascending Order)*
Current (X) probable mastery level (e.g., 0.85)
Previous (X – 1) probable mastery level (e.g., 0.60)
Probable mastery level from X – 2 (e.g., 0.33)
Remediation flag set to True when remediation is needed (default = F)
Knowledge type (SYMBOLIC, PROCEDURAL, or CONCEPTUAL)
First Pre/Post question to evaluate CE (item #1)
Second Pre/Post question to evaluate CE (item #2)
Range (inclusive) of page numbers involved in the *instruction* of this CE
Range (inclusive) of page numbers involved in the *remediation* of this CE
CE description [variable] used when remediation required ("Sorting by Ascending Order")
An *instruction done* flag set to True when instruction is completed (default = F)
An *mastery achieved* flag set to True when mastery has been reached (default = F)
An *remediation done* flag set to True following remedial instruction (default = F)
A count of the number of times remediation has been required
Children CEs (from hierarchy)
Sibling CEs (from hierarchy)
Parent CEs (from hierarchy)

scores all test items, it communicates those values directly to the student model for initialization.

Once the student model is initialized with the pretest values, a learner is placed in Stat Lady's curriculum at a CE that they do not, or only partly, know. The mastery criterion can be set at the outset of instruction to be greater than some value (e.g., > 0.83). Then, any CEs with values falling below this threshold would become candidates for instruction. In the tutor, learning about a particular CE involves instruction, and then relevant problem solving. Following Stat Lady's introduction of the CE material, a learner can solve a problem without any assistance from the tutor (i.e., on the first try, with no negative feedback). This is called level-0 assistance. Alternatively, the learner may require various degrees of assistance (i.e., level-1, level-2, or level-3, see Section 6.1, above), which is provided to the learner in response to erroneous inputs, not explicitly requested. The simple presumption is – the more help required by the learner, the less understood the current CE; and hence, the lower the associated *probable* degree of mastery (or p(CE) value). After a problem is completed, the appropriate curriculum elements within the student model are updated. The derivation of the updating scheme is outlined below.

6.2.4. *Student Model: Discrete Representation*

The student model began with the identification of three main states: Remedial, intermediate, and mastery, with low vs. high divisions within each. This yielded six

			New State			
		Levels of Help:	Level-0	Level-1	Level-2	Level-3
Current State	REMEDIAL	Low	Low Intermediate	Low Intermediate	Low Remedial	Low Remedial
		High	High Intermediate	Mid-Low Intermediate	Mid-Low Remedial	Low Remedial
	INTERMEDIATE	Low	Low Mastery	High Intermediate	High Remedial	Low Remedial
		High	High Mastery	Low Mastery	Low Intermediate	High Remedial
	MASTERY	Low	High Mastery	Low Mastery	High Intermediate	Low Intermediate
		High	High Mastery	High Mastery	Low Mastery	Mid-Low Intermediate

Fig. 2. SMART updating rules for promotion and demotion.

states, total (e.g., low-remedial, high-remedial). I then generated a set of *rules* for updating the model, per CE. These rules relate to promotion and demotion through the curriculum, as a function of the levels of assistance that are needed. That is, if a person answers a problem without any tutorial assistance (level-0), that person should get a larger boost in the student model value compared to someone who required considerably more help from the system (level-3). In fact, the learner who shows obvious difficulty with some specific part of the curriculum should receive a substantial decrease in the student model value (and, if necessary, remediation) rather than being allowed to flail unsuccessfully. Figure 2 shows the current and new state mappings, as a function of the levels of help required during problem solution.

While these decision rules may appear somewhat arbitrary, the soundness for the promotion/demotion rules was agreed upon by two expert instructors in the domain, both of whom were familiar with mastery learning approaches to teaching. Further, the rules may be thought of as qualitatively applying Bayes' theorem – the conditional probability of success given certain performance within the tutor (i.e., p(mastery) | performance). To illustrate how this works, consider Joe, classified as "high remedial" based on his pretest performance for a particular CE (i.e., he knew a little about the topic). Following tutorial instruction on this CE, suppose he required no tutor assistance (level-0 feedback) in the solution of a problem. He'd

	Current Value	New Value			
	X-axis	Level-0	Level-1	Level-2	Level-3
Remedial	$0.0 \le X \le .17$	.33	.33	.00	.00
	$.17 \le X \le .33$	.50	.42	.08	.00
Intermed.	$.33 \le X \le .50$	.67	.50	.17	.00
	$.50 \le X \le .67$	.83	.67	.33	.17
Mastery	$.67 \le X \le .83$	1.00	.75	.50	.33
	$.83 \le X \le 1.0$	1.00	.83	.67	.42

Fig. 3. Transformation from discrete to continuous representation: states and values.

thus be elevated to "high intermediate" (see white cell in Figure 2, above) which becomes his new "current" status. Furthermore, if he solved the next problem again on his own, his new state would be probable "high mastery" (shown by the black cell in Figure 2), *given* two prior instances of successful, independent performance. Alternatively, suppose Jane began the tutor at a more ambiguous status (e.g., low intermediate), and required level-3 help from the tutor to solve her first problem. She would be demoted to a "low remedial" state. If, on the next problem, she again needed level-3 assistance, she would be returned to the curriculum for remedial instruction *given* obvious difficulty with the current topic. This action results from the inclusion of various higher-level rules operating in conjunction with the student model to prevent subjects from becoming trapped in an endless cycle of solving problem-after-problem and getting nowhere (i.e., not advancing to criterion performance). Section 6.2.6 (below) discusses the specific decision rules concerning whether to provide continuing, new, or remedial instruction.

While this discrete representation provides a simple (and empirically testable) solution to the problem of deciding who should move where and by how much, it retains a categorical flavor, and thus, is not very refined. To transition from a discrete to a more continuous representation, I needed to associate values with each of the six states, and then derive functions to use as a basis for making decisions about probable mastery levels: p(CE). Figure 3 shows the six states with their associated ranges of student model values, along with a method (i.e., table lookup) to compute a new value from the current value as a function of how much assistance a person required.

Each of these discrete points can now be plotted (see Figure 4), separated into the four types of assistance required (Levels 0, 1, 2, or 3). The x-axis shows the current value, and the y-axis becomes the new value. To illustrate, suppose a learner had a current CE value of 0.50 and required a single level of help during problem

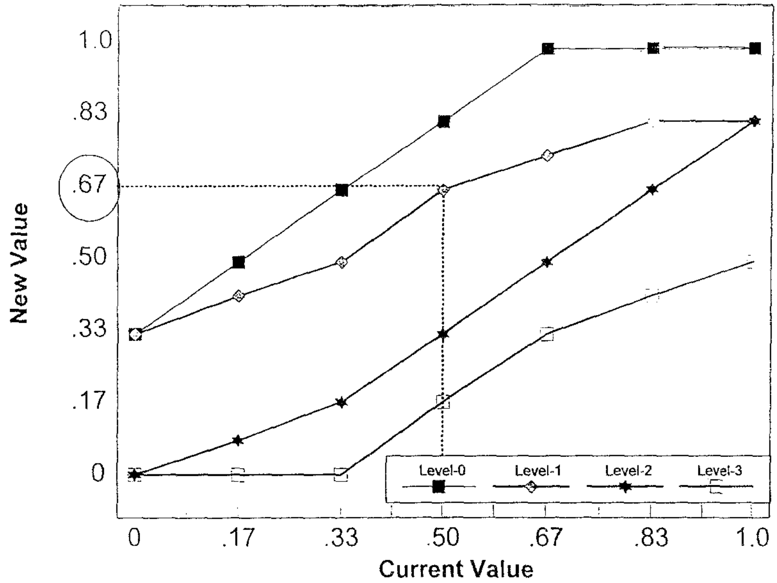


Fig. 4. Preliminary student model update functions.

solution. Draw a line up from 0.50, and intersect with the level-1 "curve." Reading over to the y-axis, the new value = 0.67.

Notice that the example given above (current value = 0.50, new value = 0.67) could be directly obtained from the table presented in Figure 3. The difference between the table and these preliminary functions is that, with these curves, we can interpolate values much more easily. That is, we're not limited to using only those values explicitly presented in the table. Rather, *any* value, between 0 and 1, can be used as a starting point. For instance, a current value of 0.59 with level-0 assistance can yield a new value of 0.90. Again, that involves drawing a line up from 0.59, intersecting the relevant curve, then reading over to the y-axis for the new value. While these four curves can serve as preliminary functions to estimate a learner's p(CE) (probable mastery value for a given CE), they really are little more than a "jury-rigged" state representation. The lines between the points are simply drawn in; they are not actual functions. One final step is required to transform these values into real functions that can be used in the computation of continuous student model values.

6.2.5. *Student Model: Continuous Representation*

For each of these four preliminary functions, I computed best-fitting curves: linear, quadratic, and cubic fits. In all cases, the linear trends fit the worst, so, on the basis of the R^2 values and visual inspection of the fits, I selected optimal curves, per function. For level-0 and level-2 assistance, the quadratic function was optimal, and for levels 1 and 3, the cubic trend best fit the values. The b-weights from

$$\hat{Y}_{(\text{Level-0})} = .3026 + 1.4377X - 0.7207X^2$$

$$\hat{Y}_{(\text{Level-1})} = .3316 + 0.2946X + 1.1543X^2 - 0.9507X^3$$

$$\hat{Y}_{(\text{Level-2})} = -.0117 + 0.5066X + 0.3518X^2$$

$$\hat{Y}_{(\text{Level-3})} = .0071 - 0.6001X + 2.5574X^2 - 1.4676X^3$$

Fig. 5. Regression equations for student model updating.

the best-fitting curves provided the basis for regression equations that can now be employed in the computation of new student model values. See Figure 5 for these predictive equations.

For instance, to compute a new CE value, the system just needs to know the current value (X), which is either the pretest score or the currently computed p(CE) value, as well as the level of help the person required from the tutor. This gets plugged into the appropriate equation, and the new value results, allowing for a true continuous representation of probable mastery values. The final graph of the best-fitting curves, plotted along with the discrete points, is presented in Figure 6.

There are several things to notice about this graph. First, the three main regions (i.e., dashed lines separating remedial, intermediate, and mastery boundaries) are arbitrary; they may be set at any value the instructor wants (e.g., setting mastery as ≥ 0.67 or ≥ 0.83). Second, the curves show some interesting properties, such as ceiling effects (level-0 curve) and floor effects (level-3 curve). Finally, notice how well the best-fitting curves match the trends defined by the rule-based points. While that is not entirely surprising (because, after all, they are best-fitting curves), it is encouraging that the fits are so close. It should be re-emphasized that these values do not represent grades, although they are determined from individuals' response histories in the solution of tutor-specific problems. Instead, these computed p(CE)s reflect probable mastery values, a *best guess* about the level to which a learner has probably achieved mastery. As such, they may also be considered confidence scores, where a 1.00 would denote 100% confidence that a learner has achieved mastery, a score of 0.67 would indicate a moderate degree of confidence that mastery has been attained, and so on.

6.2.6. *Student Model: Mastery/Remediation Decisions*

When a learner attains mastery on all of the CEs in a problem set, he or she is advanced to the next section of the tutor containing a CE that shows a below-mastery threshold value (determined by performance on the pretest, or tutorial

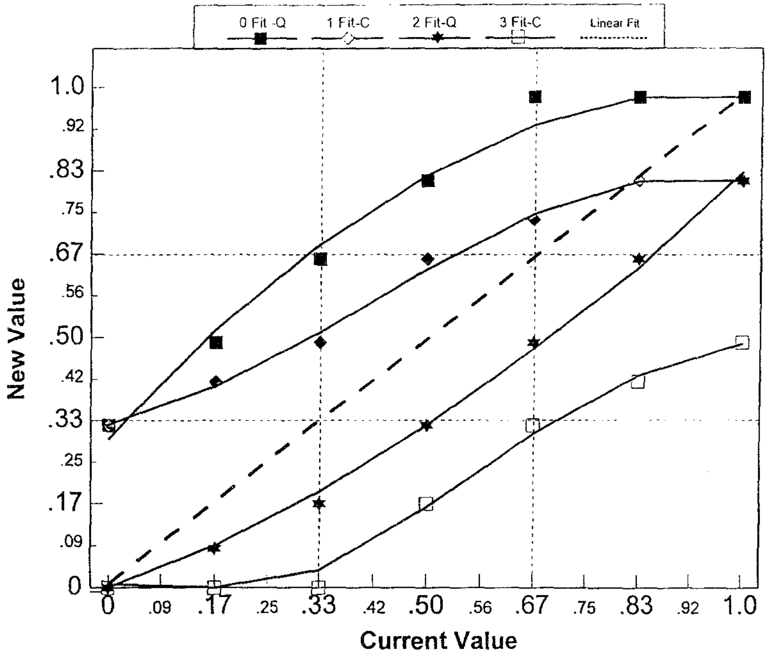


Fig. 6. Final student model update functions.

problem solving). But in the case where remediation is indicated, what happens is as follows. Each problem set evaluates several CEs, concurrently, and each CE is linked to a set of specific questions, for blame assignment. Any CE whose computed value falls below some value on one or two consecutive trials (depending on its remediation history) becomes a candidate for remedial instruction. Suppose there were three CEs that required remediation following completion of a particular problem set. Each CE can be addressed in turn (or sometimes, in parallel) as each one knows exactly the place in the tutor's curriculum where it is first instructed, as well as the specific problems which assess it. After remediation and assessment, new updating occurs.

Additional rules operate in conjunction with the student model to prevent subjects from either not advancing expeditiously when they actually have mastered the topic, or becoming trapped in an endless cycle of solving problem-after-problem and getting nowhere. Specifically, there is a two-part rule concerning whether to let the student continue solving problems or be placed back into the curriculum for remedial instruction: (a) If a CE has not yet been remediated, and the CE is < 0.60 and has decreased from the previous value, then invoke remediation, or (b) If a CE *has* been remediated and the CE is < 0.50 and has decreased across *two* consecutive trials, then remediate. The rationale underlying this decision is that I needed to "narrow" the window of floundering. That is, results from a pilot study of SMART's mastery and remediation features showed that when the mastery criterion was set at 0.83 (too high), and remediation did not get invoked until meeting a " $<$ 0.50 on two trials" criterion (too low), many learners got entangled in a frustrating cycle of problem solving. And, as frustration levels rose, motivation, and consequently learning, plummeted.

Obviously, these initial SMART functions are not going to be perfect indicators of knowledge and skill acquisition. But how well do these computed values predict outcome performance (*diagnostic validity*)? Furthermore, does adding remediation and a mastery learning criterion to the program increase outcome performance, and what effect does it have on learning time (*microadaptive validity*)? In a related sense, how much (if anything) do aptitudes and other individual differences measures contribute to the prediction of learning outcome (*macroadaptive validity*), beyond that provided by domain-specific knowledge and skill? Finally, how do subjects feel about their testing and tutorial experiences (*affective* data)?

Results will now be presented from two studies – one completed, and one very close to completion (i.e., 90% of the data have been collected). These represent the first of a series of controlled evaluations of SMART, specifically testing various features underlying the model. For example, when assessing the predictive validity of the diagnostic component of SMART (Study 1), the mastery and remediation features were disabled. Evaluating the remediation component involves comparing the relative degree of learning with this feature "turned on" (Study 2) versus when it's not employed (Study 1). I will also present data regarding the overall degree of learning that occurred from Stat Lady instruction, as well as more refined analyses that examine the data separately by outcome type and aptitude level. Finally, I will present data examining subjects' affective perceptions of their Stat Lady experience.

7. Testing SMART and Stat Lady – an Empirical Evaluation (Study 1)

7.1. Diagnostic Validity

In this study, SMART's mastery criterion and remediation features were disabled because I needed to increase outcome variability. That is, with no mastery criterion or remediation in operation, subjects would differ more in knowledge and skill acquisition than they would if everyone were elevated to the same high level of mastery. This decision allowed me to analyze the correlation between computed student model values [p(CE)s], and corresponding outcome CE scores. While this decision rendered Stat Lady "unintelligent," it serves as a necessary baseline for subsequent studies where different features of SMART are selectively turned back on.

7.1.1. *Design and Procedure*

A total of 104 subjects (61% male, 39% female) participated in this study which spanned two 8-hour days. Subjects were obtained from several local temporary employment agencies, paid for their participation, ranged in age from 17–30 years old (mean age = 22), and tested in groups of approximately 25 persons. While all subjects were required to have a high school education, previous statistics coursework was restricted to $\leq$ 1 completed class. The first day of the study involved completing an on-line demographic questionnaire, as well as an on-line pretest which assessed domain-specific knowledge of pertinent symbols, rules, and concepts (note: the learning criterion task used in Studies 1 and 2 represents about 1/3 of the curriculum from Stat Lady's descriptive statistics module). Subjects were also administered a computerized battery of cognitive and personality test measures which they completed prior to learning from the tutor. Subjects completed the tutor on the second day, spending, on average, about 5 hours learning curriculum and solving problems that focused on data organization and plotting. There were 77 CEs instructed and assessed in this segment of instruction. Following the tutor, all subjects were administered the on-line posttest.

7.1.2. *Individual Differences*

Before explicitly testing the goodness of fit between SMART's diagnostic values and the outcome data, I needed to identify individual differences variables to include in the predictive equation. Some of these variables were extracted from the demographic questionnaire, designed to assess educational background, interest in statistics, computer experience, and gender. The cognitive ability measurement (CAM 4.0) battery provided even more individual differences measures, namely: working memory capacity, associative learning skill, inductive reasoning, and information processing speed (Kyllonen et al., 1990).

To reduce these data, I computed a factor analysis on just the aptitude (cognitive) data. Four variables were used in the principal components analysis (i.e., the test scores from the quantitative measures, above), and a single factor was extracted. The four variables accounted for 66% of the *aptitude* factor variance, and factor loadings were all high: working memory (0.88), inductive reasoning (0.87), processing speed (0.79), and associative learning skill (0.69). Demographic data were used, as is.

7.1.3. *Predictive Validity of SMART*

The first research question underlying this study concerns the diagnostic validity of the computed student model values, p(CE)s, in predicting outcome CE scores. To test this relationship, I computed a stepwise multiple regression analysis. Posttest score was the dependent variable, and the following were used as independent variables: Pretest score, p(CE) data, aptitude factor score, education (years of school), and gender (male or female). Results showed that the first (i.e., strongest) variable to enter into the equation was p(CE), with a multiple $R = 0.73$ (i.e., 54% of the unique outcome variance was explained by this variable alone). Next to enter

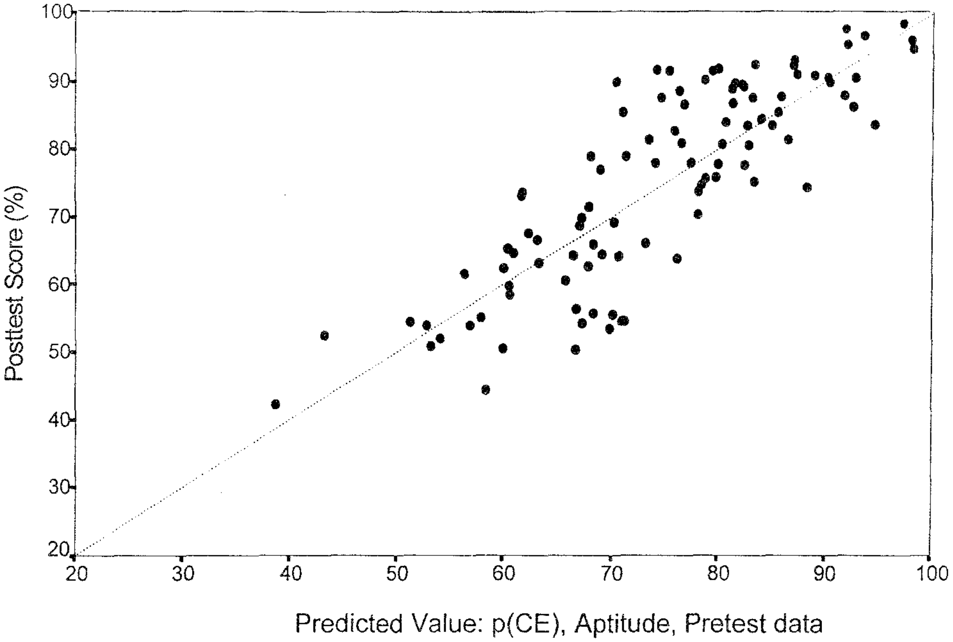


Fig. 7. Scatterplot of p(CE), aptitude, and pretest data predicting posttest score.

TABLE II. Correlations among pretest, p(CE), aptitude, school, gender, and posttest data

Variable	PRETEST	P(CE)	APTITUDE	SCHOOL	GENDER	POSTTEST
PRETEST	–	0.66**	0.63**	0.01	0.16	0.71**
P(CE)		–	0.57**	0.06	0.02	0.73**
APTITUDE			–	0.01	−0.02	0.69**
SCHOOL				–	0.07	0.06
GENDER					–	0.06

** $p < 0.001$

the equation was aptitude, increasing the multiple R to 0.81 (accounting for an additional 11% of unique outcome variance). On the third and final step, pretest data entered into the equation, accounting for an additional 4% variance, and increasing the multiple R to 0.82. None of the other variables reached criterion for inclusion in the equation.* The final data are: p(CE) ($t = 5.01, p < 0.001$), aptitude ($t = 4.25$, $p < 0.001$), and pretest, $t = 2.90$, $p < 0.005$). A scatterplot of these three variables predicting posttest score is shown in Figure 7, and intercorrelations among the relevant variables are shown in Table 2.

* The finding that education and gender failed to predict outcome performance suggests that the system is doing its job quite well – reducing the impact of potentially influential sources of individual differences in learning.

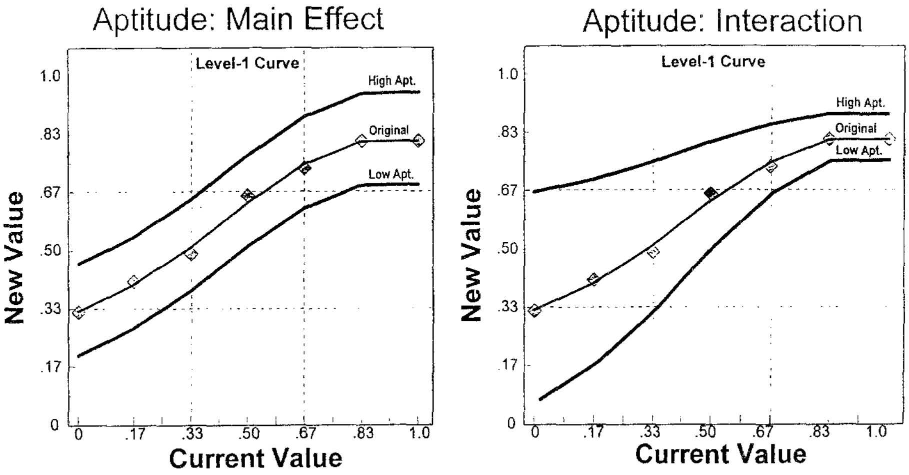


Fig. 8. Hypothetical macroadaptive curve adjustments.

7.1.4. *Macroadaptation*

The fact that aptitude significantly contributed to the equation (beyond the variance that was accounted for by the p(CE) data) justifies the macroadaptive approach, and theoretically could prompt curve adjustments. A main effect of aptitude was reported above, but a full test of main effects and interactions is needed to advise curve modifications such as the one shown on the left of Figure 8 (the main effect curve). The graph on the right illustrates a possible interaction between aptitude and the *slope* of the curve. Ensuing p(CE) values, resulting from modified regression equations, can progressively reflect knowledge and skill acquisition levels more accurately.

7.2. Learning from Stat Lady

Within the same study, I wanted to examine the overall degree of learning as a function of interacting with the Stat Lady program. Moreover, I wanted to test a series of hypotheses concerning individual differences in learning by outcome type.

7.2.1. *Hypotheses*

I hypothesized that subjects would show, on average, a treatment-effect size = 1 (1 sigma) with regard to pretest-to-posttest improvement as a result of Stat Lady instruction and problem solving. The rationale for this hypothesis was based on the following. Kulik, Kulik, and Bangert-Drowns (1985) reported that computer-based instruction (CBI) typically yields a 0.5 effect size in relation to a classroom lecture control group. As mentioned, the version of Stat Lady used in this study was technically CBI, and although there was no classroom control group for comparison, I did have data available from a no-treatment control group (N = 79) who came from the same population as the Study 1 subjects. Thus, the predicted 1-sigma increase refers to what we would expect in comparison to a no-treatment control group (i.e., no-treatment to lecture = 0.5 sigma, lecture to CBI = 0.5 sigma, thus no-treatment to CBI = 1.0 sigma increase). Computing treatment-effect size (on posttest data), does presume, however, that pretest data between the groups are comparable.

I also hypothesized that the ordering of outcome types (i.e., posttest scores in each of the three categories) by acquisition difficulty would be arrayed from SK to PS to CK. This hypothesis was based on current learning theory (e.g., Anderson, 1987; Kyllonen and Shute, 1989). Further, aptitude has historically been shown to be a powerful predictor of learning, thus providing the basis for my next hypothesis that there would be a main effect of aptitude level on knowledge and skill acquisition. And in regard to affective predictions, I expected subjects to report an overall positive feeling about their Stat Lady experience. This was based on data collected from an earlier study involving the Probability Tutor (see Shute and Gawlick-Grendell, 1994) which reported that, overall, students not only learned a lot, but also greatly enjoyed their learning experience from that Stat Lady module.

The only interaction that I predicted involved the relationship between learning and outcome types. In particular, given the experiential nature of the tutor, in conjunction with scores of studies showing how skill acquisition improves in relation to the amount of practice opportunities (e.g., Bryan and Harter, 1899; Fisk and Rogers, 1992; Schneider and Shiffrin, 1977; Woltz, 1988), I expected that subjects would show more dramatic improvements in procedural skill acquisition compared to the more declarative knowledge acquisition. That is, a salient feature of Stat Lady is the provision of hands-on exercises, especially useful for practicing PS elements.

7.2.2. *Results*

I computed a MANOVA using Gain (pretest to posttest changes in scores) and Type (SK, PS, CK) as within-subject variables, and Aptitude (high or low)* as the between-subjects variable. The first question examined the degree of learning that resulted from Stat Lady instruction. First, the main effect of Gain was significant: $F(1, 102) = 794.14, p < 0.001$; Pretest $M = 44.1\%$ (SD = 14), Posttest $M = 75.3\%$ (SD = 15). Thus, Stat Lady managed to elevate subjects, overall, from an incoming score of 44% to a final score of 75%, representing an increase of more than *two* standard deviations from their incoming status.

* Aptitude level was computed from CAM-4 data submitted to a principal components factor analysis. A single factor (Aptfac) was extracted, 66% explained variance, and factor loadings on each of the four variables were all high. Factor scores for Aptfac were saved for each subject, and a median split was computed to arrive at high and low aptitude categories.

TABLE III. Descriptive statistics on pretest and posttest by outcome type and aptitude

Low Aptitude (N = 52)	Mean	SD	Min	Max
SK-Pretest	38.82	16.39	00.00	75.00
SK-Posttest	69.47	17.73	25.00	100.00
PS-Pretest	32.35	14.66	03.74	75.00
PS-Posttest	71.18	14.49	33.42	96.65
CK-Pretest	36.29	11.00	03.33	60.00
CK-Posttest	59.49	13.90	30.83	86.53
High Aptitude (N = 52)				
SK-Pretest	50.36	18.21	12.50	100.00
SK-Posttest	83.65	14.31	50.00	100.00
PS-Pretest	51.08	19.69	18.50	88.61
PS-Posttest	85.65	13.87	39.02	100.00
CK-Pretest	51.22	12.27	27.20	80.00
CK-Posttest	79.66	12.69	42.78	99.17

To interpret the Stat Lady learning gains in relation to simple pretest-to-posttest change, I examined data from a demographically-matched group of 79 subjects who received no instructional intervention and whose pretest M = 48% (SD = 16) and posttest M = 55.5% (SD= 18). Although this group started out with significantly higher pretest scores than the Stat Lady group: $F(1, 181) = 4.34$; $p < 0.05$, the Stat Lady subjects ended up with significantly higher posttest scores: $F(1, 181) = 61.04$; $p < 0.001$. Furthermore, the Stat Lady condition showed a 74% pretest-to-posttest improvement rate compared to the baseline group's 15%. The effect size from these data = 1.1, an underestimate of the true effect size given the control group's higher pretest scores.

The analysis also showed a main effect of aptitude: $F(1, 102) = 45.12$, $p < 0.001$ where, predictably, high-aptitude subjects performed better across all tests compared to low-aptitude subjects. Results also showed a main effect of Type: $F(2, 204) = 3.20$, $p < 0.05$, a significant Gain × Type interaction: $F(2, 204) = 5.44$, $p < 0.01$, as well as a significant 3-way interaction involving Gain × Type × Aptitude: $F(2, 204) = 4.03$, $p < 0.03$. No other interactions were significant. Table 3 shows the means and standard deviations of these data, while Figure 9 depicts the obtained 3-way interaction.

7.2.3. *Discussion*

Although I had expected to see some learning gain resulting from the Stat Lady experience, I was surprised by the magnitude of improvement. That is, the two standard deviation increase from pretest to posttest performance is quite impressive, especially in the absence of any active intelligent component (e.g., mastery and

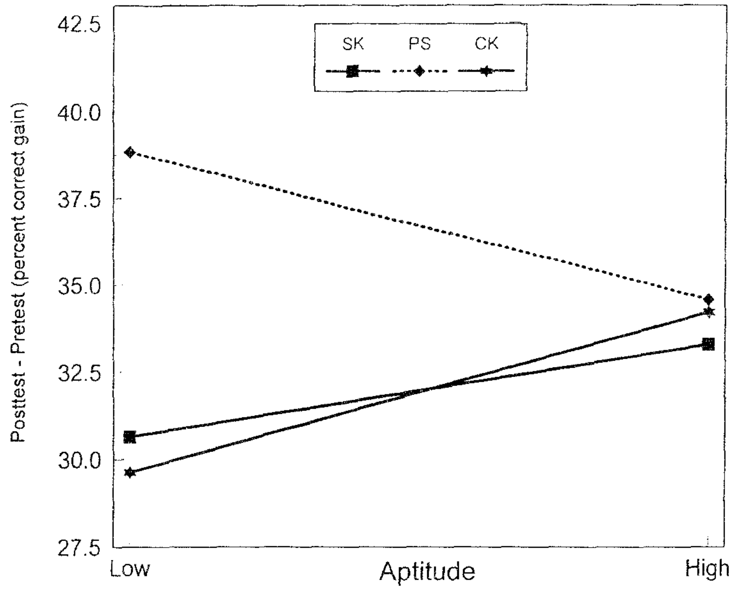


Fig. 9. Learning gain by artifact type and aptitude interaction.

remediation) and given the relatively short amount of instructional time. This now becomes the "score to beat" in subsequent studies of Stat Lady/SMART where mastery criterion and remediation features are turned on. The finding that high aptitude subjects performed better on the outcome measures compared to the "cognitively challenged" was no surprise. In general, subjects with higher incoming abilities are able to more quickly acquire new subject matter, regardless of learning environment. What was encouraging from these findings was that the difference between aptitude levels was not that large (see Figure 9, above). The hypothesis concerning a main effect of outcome type was similarly upheld. Subjects did have a more difficult time solving CK test items than either SK or PS ones. The mean percent correct data across the three outcome categories (collapsing across pretest, posttest, and aptitude level) are: SK = 61%, PS = 60%, and CK = 57%. This is in keeping with the previously-hypothesized difficulty ranking of outcome types, in general.

The Gain × Type interaction shows differential performance improvements by outcome type. Learning gains for procedural skill elements are much higher than either SK or CK elements (as predicted). This result, most likely, is a function of the effects of greater practice on outcome performance. That is, the probability of having a procedural outcome established is related to the number of times an item is exercised (or reentered into working memory), and strength increases as a power function of practice (Anderson, 1987; Carlson and Yaure, 1990). The average number of questions answered during the tutor separately by outcome type reveals that subjects did, in fact, receive more practice on PS elements than either SK or

CK ones. As mentioned, the experiential nature of the program was designed to provide subjects with extra practice opportunities. For instance, when constructing a frequency distribution, learners had to perform the same low-level procedure several times within a given problem set (e.g., filling in the cells of frequencies required learners to (a) choose a value in the range of the sample, (b) count the number of times that value occurred in the sample, and (c) enter that frequency in the correct cell). However, there is a fine line between enough and not enough practice. Thus, I attempted to strike a balance between giving learners a sufficient opportunity to automate the process(es), while not requiring so much repetition that learners would become bored. I settled on a requirement of completing five cells per column, which allows sufficient practice for that skill, but not undue tedium. Stat Lady filled in the remainder of the cells. This amounts to five times the opportunities that learners have acquiring most SK and CK elements within the tutor.

The 3-way interaction (Gain × Type × Aptitude) shows that differences exist between high- and low-ability subjects in terms of their relative gains across SK, PS, and CK elements. As seen in Figure 9, high-ability subjects achieve equivalent learning gains on all three outcome types. Low-ability subjects' gain is slightly less, except in the case of procedural skill. In that case, they demonstrate dramatic gains, far beyond even those of the high-ability learners. This result generates at least two interpretation issues. First, why would low-ability subjects' procedural skills increase so much more than their symbolic and conceptual knowledge? This again is readily accounted for via the practice opportunities explanation above. The added number of opportunities subjects have to work on their procedural skills allows the low-ability subjects a chance to improve to a substantial degree. Second, why don't the high ability subjects improve to the same degree? This result is undoubtedly a function of ceiling effects. As seen in Table 2, high-aptitude subjects perform at a fairly elevated level (relative to their low-aptitude peers) on the pretest. By the time they improve 35%, they have increased to roughly 85% correct on the posttest. That represents about as high an average as any educator could reasonably expect given such abbreviated instruction.

7.2.4. *Affective Data*

After completing the tutor, all subjects completed an on-line "affective survey" rating various aspects of the tests and tutor on a 7-point scale. Subjects in this study had no (or very minimal) prior statistics education or training. Consequently, most of them felt very frustrated by their pretest performance, despite the fact that they were told to "do their best" and not to worry if they could not solve some of the items (see upper-left corner of Figure 10). In contrast, subjects perceived the tutor itself, and subsequently the posttest, to be considerably less frustrating. In fact, as shown in the bottom row of the figure, subjects really enjoyed the instructional part

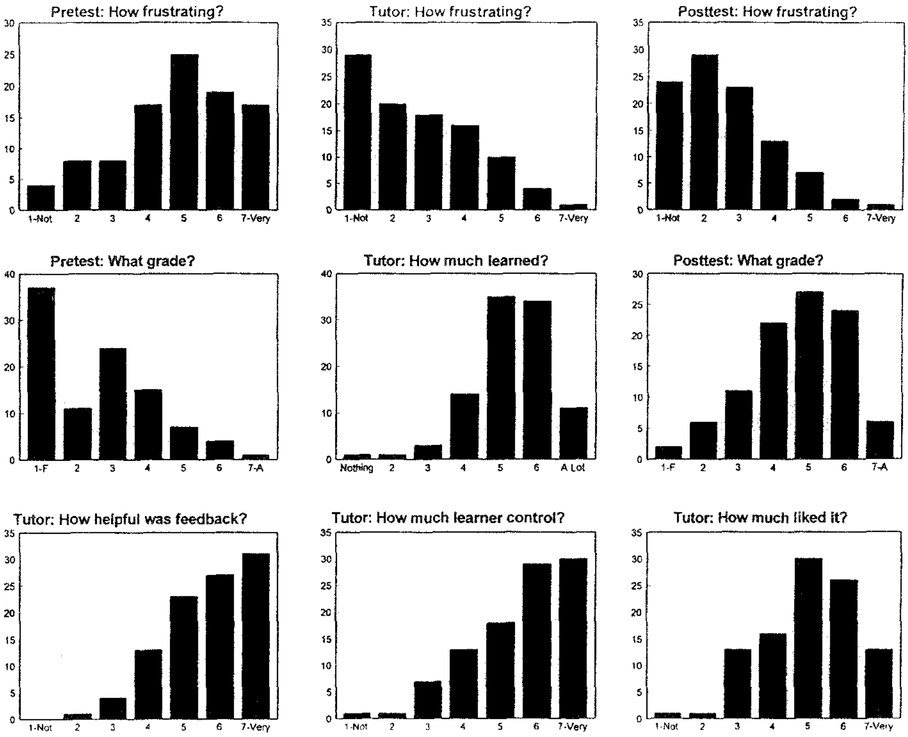


Fig. 10. Affective data related to testing and tutorial experiences.

of the study where they reported that (for the majority) the tutor's feedback was helpful, they felt in control of their learning, and they liked the program.

8. Testing Mastery and Remediation Features in Smart – Preliminary Findings (Study 2)

In addition to cognitive diagnosis, the second important component of SMART involves remediation as a function of the mastery requirement. These conjoined features are currently being assessed in a controlled evaluation to determine the degree to which learning is improved with the systematic inclusion of these components. Two basic questions include: (a) How does the inclusion of a mastery criterion and remediation influence learning outcome, in general? and (b) How does their inclusion impact learning efficiency, in a cost-benefit sense?

8.1. Design and Procedure

Currently, I have data from four groups of subjects (N = 100). When the study is completed, I will have collected data from over 120 subjects. Subjects in this study represent the same population, demographically, as those who participated in Study 1. That is, they are obtained through the same local temporary employment agencies, are roughly 60% male and 30% female, and approximately 22 years old.

8.2. Learning Improvement

The first question concerns how much is added to learning outcome beyond that shown in Study 1. Currently, subjects in Study 2 (N = 100) are showing higher posttest scores (M = 82.1%) compared to those from Study 1 (M = 74.9%). However, they also are arriving with higher pretest scores (M = 49.8%) compared to subjects in Study 1 (M = 43.5%); F (1, 199) = 9.61; $p < 0.01$. To control for these differences in incoming knowledge and skill, I computed an ANCOVA on the posttest data with pretest as a covariate and condition as a between-subjects variable. Results showed that there was a significant difference in learning outcome due to condition: F (1, 199) = 4.16; $p < 0.05$, with superior outcome performance evidenced by subjects in Study 2 (i.e., the Stat Lady condition that invoked a mastery criterion and remediation). Thus, we can tentatively conclude that remediation appears to help learners acquire even greater degrees of domain-specific knowledge and skill. But we still have to ask "at what cost" in regard to training/learning time.

8.3. Learning Efficiency

The time it takes to complete the tutor, with mastery and remediation in place, reflects learning efficiency (i.e., speed and accuracy of acquisition). Regarding a comparison between the two studies, I had no *a priori* hypotheses about relative efficiency. That is, in Study 1, subjects had to complete all 77 CEs in the curriculum, even those that most subjects already knew about (like sorting data into ascending and descending order). Furthermore, Study 1 subjects were required to solve two problems per problem set, regardless of pretest performance. In contrast, subjects in Study 2 were only exposed to those CEs that were not, or were only partially, known (i.e., CEs that were answered correctly on the pretest were not explicitly instructed by the tutor). Thus, in general, Study 2 subjects may require *less* time to complete the tutor given fewer CEs instructed. On the other hand, remediation and mastery were operating in this study, which could have *added* time to the learning process.

Because of the incoming differences in pretest scores, noted above, I computed another ANCOVA, using the total amount of time spent learning and problem solving within the tutor as the dependent measure, pretest as the covariate, and condition as the between-subjects factor. Preliminary results show that learners required significantly more time to complete Stat Lady in Study 2 with the mastery criterion and remediation features operational (M = 7.6 hr) compared to subjects in Study 1, where Stat Lady had those features disabled (M = 4.4 hr); F (1, 200) = 223.66; $p < 0.001$. More refined cost-benefit analyses will be forthcoming once all of the data have been collected and separated into respective times for: instruction, remediation, self-initiated review (e.g., accessing the on-line dictionary), and so on.

8.4. Discussion

In summary, the results from Study 2 (not entirely completed) has tentatively indicated that when mastery criterion and remediation are invoked by the system, learning increases even beyond that seen in Study 1. This is exciting because the large (2 SD) pretest to posttest improvement shown in Study 1 was viewed as difficult to surpass. Results from these studies (as well as future SMART experiments) will enable me to progressively optimize learning for a variety of individuals and outcome types. For instance, data from Study 1 can provide information that will enable me to "tweak" the initial regression curves/equations to more accurately predict outcome in future studies, as well as improve instruction, if necessary (for more on this, see Section 9.1, below).

Average p(CE) and posttest data were used in the initial, global assessment of SMART's diagnostic accuracy. However, more refined analyses are needed, such as examining these data separately by the four levels of assistance, and even at the individual (or aggregate) CE level. It will be informative to see if these four curves consistently over- or underestimate the true posttest values. If so, then curve adjustments are called for, achieved by altering b-weights, accordingly. The new regression equations should more accurately predict learners' actual knowledge and skill levels in future studies. This interplay among theoretically-based, rule-driven, and empirically-validated issues is just what is needed to advance the field of artificial intelligence and education. I'll now present some additional research questions that I'm in the process of examining with SMART.

9. Current Research with Smart

9.1. Empirical and Theoretical Modifications to the Curves

Findings related to the predictive validity of SMART's updating heuristics not only provide information about the overall goodness-of-fit between SMART and the outcome measures, but can also serve as an basis for modifying the regression equations. For instance, the degree to which SMART's functions over- or underestimate the actual posttest scores suggest downward or upward adjustments to the curves which could easily be achieved by altering the b-weights, accordingly. An alternative, more theoretical approach to modifying the curves has been suggested by Anthony Jameson (personal communication, March 1995). Specifically, he suggested using a formal Bayesian approach to generate equations. At any given time, there is a given p(mastery) for a particular CE. For each of the four possible outcomes (i.e., levels of help) per CE, the ratio of the likelihoods of that outcome given mastery vs. non-mastery may be defined. For example, we may postulate that level-0 behavior is 5 times as likely given mastery than given non-mastery,

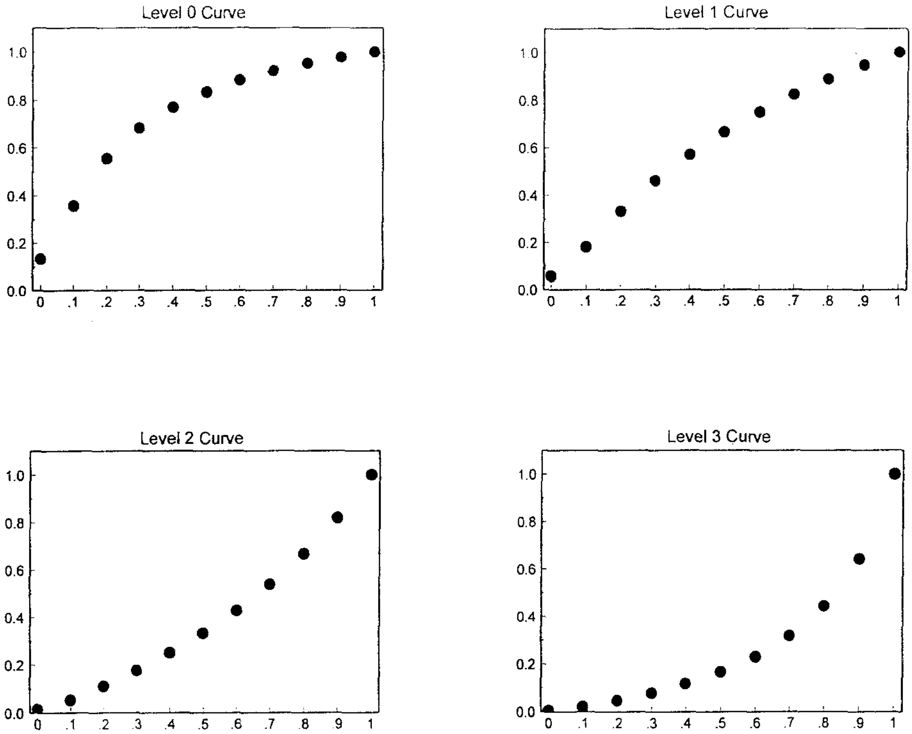


Fig. 11. Four new curves based on Bayes' rule.

and that the ratios are 2, 0.5, and 0.2 for level-1, level-2, and level-3 actions, respectively. We can then view each behavioral outcome for a problem during the tutoring phase as an observation, applying Bayes' Rule to update p(mastery). The updating formula that results from Bayes' Rule can be rewritten for this case as follows:

$$p(\text{mastery} \mid \text{outcome}) = \left\{ \frac{1}{\left(\frac{1/p(\text{mastery})-1}{\text{likelihood ratio (outcome)}}\right)+1} \right\} \qquad (1)$$

where *p(mastery)* is the probability before the observation (i.e., current value), *p(mastery |outcome)* is the updated value, and *likelihood ratio(outcome)* is the likelihood ratio associated with the outcome in question. For each of the four outcomes, we then get a continuous curve with the same basic meanings as shown in Figure 6. Figure 11 shows the four curves derived from the four likelihood ratios listed above (chosen to coincide exactly with Figure 6's values when current value = 0.50).

The above approach to deriving the curves has the advantages that the form of the curves follows from some theoretical assumptions, and each curve is characterized by a single, interpretable parameter (the likelihood ratio). In a preliminary analysis of the validity of these curves (relative to the original ones), I've recomputed students' data from Study 1 (N = 104) using the new values to compute student model values. Results show that the values arising from the new curves, alone, can account for even more of the unique outcome variance compared to the original curves (i.e., new curves $R^2 = 0.70$, original curves $R^2 = 0.54$). The next step is to systematically test a series of parameters against the outcome data to further optimize the fit. This has obvious implications for the next iteration of SMART.

9.2. Testing the Contribution of the Inheritance Hierarchies

The student model values are intended to be propagated through the inheritance hierarchy to provide preliminary estimates for complex curriculum elements that contain element X as a component knowledge or skill. Also, information concerning the acquisition of a sibling CE can potentially be used to enhance "lateral" predictive validity of a current CE. The preliminary data analyses described above do not include relational data from the inheritance hierarchies. Thus, another logical series of regression analyses may compare predictive validities of the p(CE) scores both with and without "lateral-sibling " CE data in the equation (see Appendix B for an example of one of the hierarchies). For example, CE-2 represents the procedural skill of sorting data in descending order, while CE-1 assesses sorting data in ascending order. These two "sibling" CEs are perfectly correlated for both pretest and posttest performance ($r_{xy} = 1.00$; N = 104). Thus, using posttest CE score as the dependent variable, I can test for any additional, unique variance accounted for by including p(CE-1) in the equation, along with the computed value for p(CE-2) in the same equation. This test will show the relative contribution of sibling information in predicting targeted CEs. For example, terminal p(CE-2) with level-1 assistance = $[0.332 + 0.295X + 1.154X^2 - 0.951X^3]$ + information from terminal p(CE-1). Terminal p(CE-1) data can then be used to either increase or decrease the estimate for p(CE-2). Extending this rationale to the entire hierarchy, software currently exists (e.g., ERGO, Noetic Systems, 1993) that enables one to easily establish a network of related nodes (CEs) in terms of conditional probability distributions for their relationships. The network (or hierarchy) then is updated based on probabilistic relations. Proximal nodes are impacted more strongly, while more distal nodes have weaker increments/decrements. I'm currently in the process of acquiring the software needed to set up the Bayesian network. Because the hierarchies have already been established, what remains is to assign values to the linkages between CEs. This will provide the basis for an empirical test of relative predictive validities between p(CE) values inferred from the Bayesian net with those derived from just the regression equations in SMART.

In addition to some of the research described above, there is also a plethora of additional questions that I plan to examine in the upcoming years. Some of these research areas will now be outlined.

10. Future Research with Smart

10.1. Remediation

The diagnosis of learner difficulty in regard to a particular CE suggests (and may invoke – dependent on the severity of the problem) remediation. This can involve either presenting a review of previous, related instruction, or the creation of specially-designed problems that focus on the particular bug or misconception. Each of the three student models (SK, PS, CK) may ultimately require different *flavors* of remediation. For instance, symbolic knowledge may best be remediated by drill and practice, procedural skill may be optimally remediated by presenting problems that specifically relate to the explicit (or inferred) bug, and conceptual knowledge may be remediated with carefully-designed analogies. Further, explicitly instructing CK via analogies (for either initial or remedial instruction) may serve a dual purpose: (a) enhance the memorability of CK elements (over time, a *retention* issue), and (b) help learners solve related PS elements (in real-time, a *transfer* issue). This attempts to capitalize on the best aspects of a variety of theoretically-grounded student modeling approaches by pairing them with their most appropriate type of knowledge/skill remediation. Currently, remediation within SMART involves specific re-instruction of the problematic CE(s), followed by re-assessment within the context of new problem sets.

10.2. Adaptive Pretest

While the current pretest assesses knowledge and skill on every CE in the tutor, a more efficient procedure, deriving from the inheritance hierarchies (and alluded to in Section 9.2, above), would be to utilize that information to make a more adaptive pretest. In that case, inferences could be generated about a learner's actual understanding of one (or more) CE(s) based on performance data from related CEs. The result would be a considerably shorter pretest, which would have the advantage of being less negatively perceived (see Figure 11, above, regarding affective data related to the pretest). To illustrate, if a learner obviously knew how to sort numbers by ascending order, we could confidently infer that he or she could also sort them by descending order. The strength of the inference (Bayesian rule) would depend on the relatedness of the CEs in question. Again, elements that were more distal would have lesser associative strengths.

10.3. Generalizability of Smart

Initially, I have implemented and tested SMART within a single domain – Stat Lady. I am currently testing the generality of the paradigm using other criterion tasks, selected as being dissimilar to Stat Lady's domain. For instance, the Navy Personnel Research and Development Center is supporting the development of an instructional system called PC-IMAT (PC version of the classroom Interactive Multisensor Analysis Trainer [IMAT]). The domain includes teaching concepts required for the successful planning and execution of antisubmarine warfare (ASW); specifically, concepts associated with determining/predicting underwater sound transmission paths and ranges. Navy-standard models of the elements of the oceanographic environment which affect acoustic propagation are necessary to illustrate these concepts, and are being incorporated into the lessonware, written in Visual Basic. We (John Schuler and I) are currently incorporating SMART into the delivery system, and the process is quite straightforward. The general issue concerns how well (and easily) SMART can characterize an assortment of tasks using the relatively simple framework described herein. To expedite this process, another (planned) parallel research stream involves incorporating SMART into an existing intelligent tutoring system authoring shell – RIDES (Munro, 1993; Towne, 1993). A top-down series of questions will appear for the subject-matter expert to answer. This will ease the burden associated with conducting a less systematic approach to cognitive task analysis, categorization, and hierarchical organization of ensuing CEs, the first steps in the SMART process.

10.4. Visible Student Model

Another research question concerns whether the explication of a learner's knowledge and skill acquisition actually helps or hinders the learning process. To test this question, I plan on making the student model visible to the learner in the form of an on-line Report Card, presented as a bar graph showing the level of mastery the system believes the student has achieved (i.e., $\mathrm{S/L}_{SK}$, $\mathrm{S/L}_{PS}$, $\mathrm{S/L}_{CK}$). The Report Card will represent concepts and skill acquisition at a global level (e.g., SK of the Mean, PS in computing the Mean), then will be further decomposable into individual elements (e.g., success on Σ, N, X, $\Sigma X/N$). This information may be helpful to the learner who is concerned about his/her progress. Additional records will be maintained on students' usage of the elective tools (e.g., number of times they accessed the on-line Dictionary). This information is intended to be inspectable as well, and perhaps will work its way into the formal student model. That is, a person who supplements his/her tutorial instruction by engaging in self-initiated reading of the hypertext Dictionary, for instance, may be more successful in knowledge and skill acquisition (compared to passive learners) given this evidence of active, motivated learning behavior.

10.5. Alternative Mastery Criteria

Does altering the mastery criteria within Stat Lady have any impact on learning? I plan to run subjects in each of three separate conditions where the mastery criterion will be fixed at > 0.50, 0.60, and 0.70. Results from some pilot test data that I have collected showed that the original criterion (> 0.83) was much too stringent (and consequently, frustrating) for most individuals. That is, to attain mastery at that

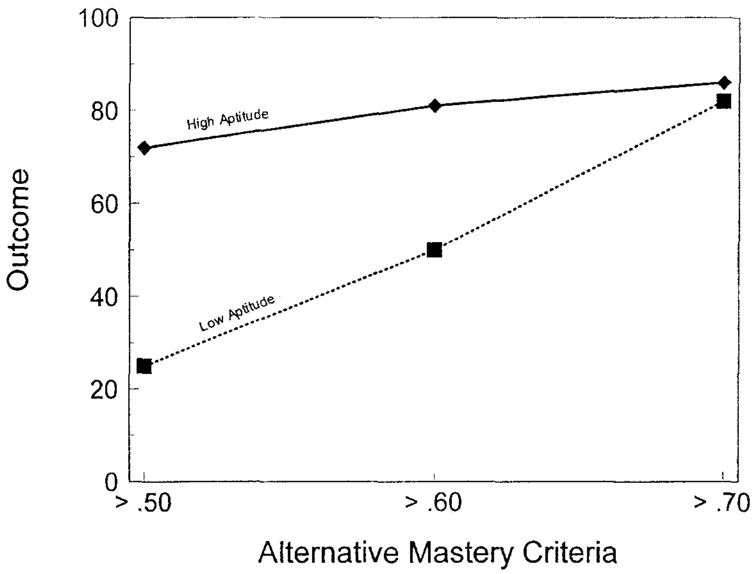


Fig. 12. Hypothesized interaction between aptitude and criterion level.

level, too many problems had to be solved. This not only added time to the learning process, it also frustrated learners, forcing them to continue solving problems once they had already understood the concept or rule. Consequently, in the studies reported in this paper, the mastery criterion value was reduced to 0.70. Perhaps even lower criterion values may suffice for a demonstration of acquisition for most (or certain) knowledge and skills. But this remains an empirical question. I also envision an interaction between criterion level and aptitude on learning outcome. Figure 12 depicts a hypothetical interaction, where low-aptitude subjects perform better when the mastery criterion is set higher, while high-aptitude subjects perform well, regardless of criterion level. The basis for this hypothesis is that if lower-aptitude learners are allowed to advance to a new CE before it is truly "mastered" (e.g., using a criterion > 0.50), then consequently their retention, assessed by posttest performance, may be adversely affected. Moreover, as was seen in Figure 9, low-aptitude subjects were substantially helped by the extra practice opportunities; and the higher the criterion, the more practice that is required to attain "mastery" status.

10.6. More Refined Aptitude Measures

I would like SMART to eventually represent and utilize information about learners' aptitudes across a *variety* of cognitive (and other) measures. The macroadaptive analysis, discussed earlier, only considered the contribution of a global aptitude measure. But given that I did find a significant influence of general aptitude, further tests are warranted, using specific aptitudes in the equation. These more refined analyses can readily be made via the CAM battery, in conjunction with other available data I have from the two studies (e.g., personality data), testing the contribution of a variety of aptitudes in predicting outcome performance. Another way that aptitude data can potentially enhance learning is to employ findings from ATI research. For instance, a person judged to be "high exploratory," could go through Stat Lady with the explicit feedback (levels 2, 3) turned off. Similarly, low exploratory learners could learn with the less explicit feedback turned off (levels 1, 2). Finally, as mentioned earlier, aptitude information can also result in curve modifications (entire curve, slope or intercept adjustments). This information can potentially make the model even stronger and hence, more predictive.

11. Summary and Conclusion

This paper describes a novel student modeling paradigm (SMART) that employs both microadaptive (domain-specific) and macroadaptive (domain-independent) information. That is, SMART not only assesses emerging knowledge and skills, but also incoming aptitudes, using this information to subsequently adapt instruction for a particular learner. Moreover, the system assesses a range of outcome types, unlike other approaches that focus on a single learning outcome.

SMART operates in conjunction with a tutor design where low-level curriculum elements (CEs) are identified and arranged in inheritance hierarchies, separated into symbolic knowledge, procedural skill, and conceptual knowledge. Throughout the course of the tutor, these CEs are instructed, evaluated, and remediated, if necessary. Three main routines drive the student model: (1) Initializing student model values, (2) Updating student model values via regression equations, and (3) Determining what to do next (i.e., remediate, continue with the same CE, or advance to the next relevant CE). Because SMART uses a mastery learning approach, it requires a learner to achieve a certain level of mastery before being advanced to the next unit of instruction. Several executive control rules are responsible for the decision to either advance or remediate the learner in a timely fashion, thus providing benefits to a range of aptitude levels.

The efficacy of the program's diagnostic capabilities was shown to be quite accurate, accounting for 54% of the unique outcome variance on the basis of just the computed student model values, and 67% of the outcome variance when aptitude and pretest data entered the equation. An alternative set of curves, discussed in Section 9.1, have demonstrated even greater predictive validity, accounting for 70% of the outcome variance. Results from a series of tests currently in progress, across a range of parameter values, will disclose optimal fits. Once determined, these will provide the computational (as well as theoretical) basis for SMART's updating heuristic.

Evaluating the mastery/remediation components of SMART compared the relative degree of learning with this feature "turned on" versus when it was not employed. Because the system knows precisely where each CE is instructed and assessed, remediation is surgical and efficient. Recall that in Study 1, this feature was not operational; subjects simply proceeded through the entire curriculum, solving two problems per CE before moving on. Despite the absence of this important feature, subjects still managed to show a two standard deviation increase from pretest to posttest performance. In Study 2 (with remediation operating, and mastery level set to > 0.70), outcome performance was shown to be even higher. However, this greater outcome was achieved at the expense of greater learning time – Study 2 required about 3 more hours to complete Stat Lady compared to Study 1. Cost-benefit analyses will reveal whether the obtained increase in outcome is justified by the additional cost in learning time. However, the results of this analysis may not be completely definitive because the content of the instructional material (i.e., descriptive statistics) is really not that difficult. Thus, the potential benefits of remediation may obscured due to ceiling effects on learning. Another use of the results from this kind of analysis is that it can provide valuable decision-making data to instructors who either: (a) have a fixed amount of instructional time and do not mind if outcomes vary, or (b) are concerned with fixing outcome (e.g., to some mastery criterion), but who can afford to have learning times vary. The cost-benefit data (relative outcomes and training times under various conditions) would impact the decision to employ either an intelligent or non-intelligent version of some software.

In addition to the specific research questions examined in this paper, other general research questions include: What types of learning environments are more appropriate for learners with certain characteristics? Do these same person-environment interactions manifest across domains? Are certain domains better suited for specific instructional methods? How much learner control should be allowed? Answers to these questions will inevitably boost the instructional capabilities of automated systems of the future.

In conclusion, the SMART approach to intelligent tutoring attempts to facilitate learning by bringing together assessment, diagnosis, remediation, and mastery-based learning in a dynamic synergy that will tailor instruction to students' particular profiles of abilities and needs.

Acknowledgements

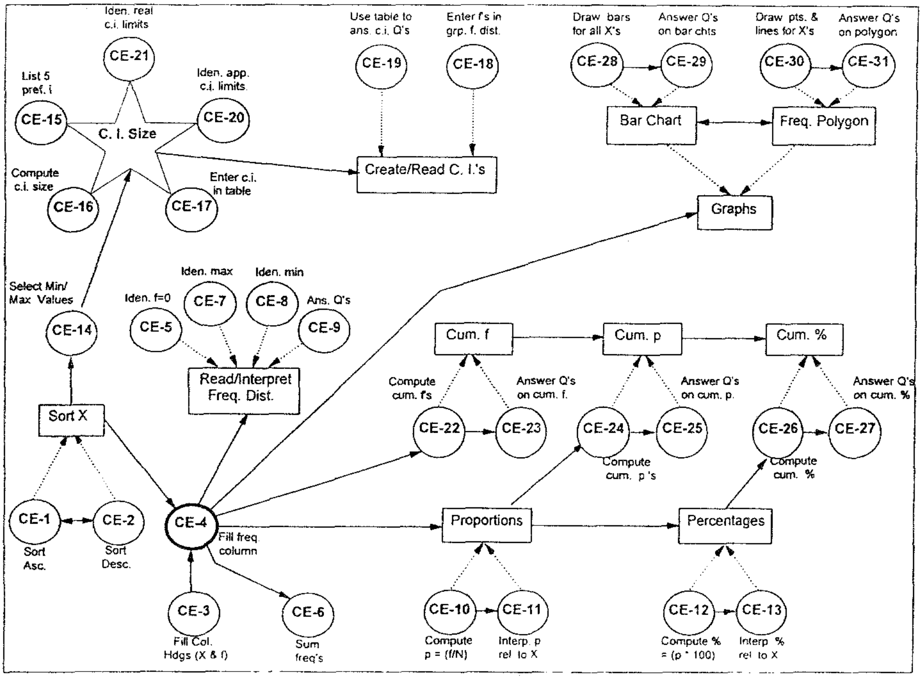
I gratefully acknowledge the contributions that Lee Gugerty and Wes Regian made to this research project. I'd additionally like to thank Kevin Gluck, Nancy Lefort, and Roger Dubbs for their very sagacious assistance in creating the student modeling paradigm (i.e., their SMART contributions), as well as Pat Kyllonen and Martin Ippel for their insightful discussions about experimental studies. I'd like to thank Lisa Gawlick-Grendell, Oscar Orta, Linda Robertson-Schüle, and Cathy Gomez for their help in data collection. Finally, I greatly appreciate AFOSR for their funding support (via TRAIN), as well as the helpful comments made on an earlier draft of this paper by anonymous reviewers.

Appendix A
Curriculum Elements from the Stat Lady: Descriptive Statistics Tutor

CE #	Type	Description
Section 1		
1	PS	Sort data in ascending order.
2	PS	Sort data in descending order.
3	CK	Definition of "frequency."
4	CK	Definition of "distribution."
5	CK	Definition of "frequency distribution."
6	CK	Definition of "variable."
7	CK	Definition of "value."
8	PS	Fill in the column headings for a simple freq. dist.
9	PS	Fill in the cells of the frequency column.
10	SK	Symbol for a variable (X).
11	SK	Symbol for frequency (*f*).
12	PS	Identify values with a frequency of 0.
13	PS	Identify least frequently occurring value.
14	PS	Identify most frequently occurring value.
15	PS	Answer interpretive questions about a sample.
16	SK	Symbol for summation (σ).
17	SK	Formula for the sum of all the frequencies in a sample.
18	PS	Sum all the frequencies in a sample.
19	SK	Symbol for the total sample size (N).
Section 2		
20	CK	Definition of "proportion."
21	SK	Formula for computing proportions.
22	SK	Symbol for proportion (p).
23	PS	Compute proportions.
24	PS	Answer interpretive questions about the proportions in a frequency distribution.
25	CK	Knowledge that the sum of the proportions in a frequency distribution equals 1.
26	CK	Knowledge of the computational relationship between proportions and percentages.
27	SK	Symbol for percent (%).
28	SK	Formula for computing percentages.
29	PS	Use proportions to compute percentages.
30	PS	Answer interpretive questions about percentages in a frequency distribution.
Section 3		
31	CK	Definition of "class interval."
32	PS	Identify the lowest and highest values in a sample.
33	CK	Knowledge that roughly 12 class intervals is an appropriate number for a grouped frequency distribution.
34	SK	Symbol for the class interval size (i).
35	CK	Knowledge that a "Preferred" class interval size should be used in creating grouped frequency distributions.

36	PS	Compute the best class interval size for a given sample.
37	SK	Formula for computing class interval size.
38	PS	Choose an appropriate Preferred *i*.
39	CK	Knowledge that the lower limit of the intervals in a grouped frequency distribution should be a multiple of the class interval size.
40	PS	Enter the class intervals for a grouped frequency distribution.
41	PS	Enter correct frequencies in the cells of a grouped frequency distribution.
42	PS	Answer interpretive questions about a grouped frequency distribution.
43	CK	Knowledge that any one value in a sample can apply to one and only one class interval.
44	CK	Knowledge that there can be no gaps between class intervals. They must encompass all values within the range of the sample.
45	CK	Knowledge that all intervals within the range are included in a grouped frequency distribution, even if there are no values that fall within those class intervals.
46	CK	Definition of “apparent limits.”
47	CK	Definition of “real limits.”
48	PS	Identify the real limits of a class interval.
49	PS	Identify the apparent limits of a class interval.
Section 4		
50	SK	Symbol for cumulative frequency (*Cum. f*).
51	CK	Definition of “cumulative frequency.”
52	CK	Knowledge of the purpose that cumulative frequency distributions serve.
53	PS	Compute cumulative frequencies.
54	CK	Knowledge that the final value in a cumulative frequency column should be equal to the sample size.
55	PS	Answer interpretive questions about cumulative frequencies.
56	SK	Symbol for cumulative proportion (*Cum. p*).
57	SK	Symbol for cumulative percentage (*Cum. %*).
58	CK	Definition of “cumulative proportion.”
59	CK	Definition of “cumulative percentage.”
60	SK	Formula for cumulative proportion.
61	PS	Compute cumulative proportions.
62	PS	Answer interpretive questions about cumulative proportions.
63	SK	Formula for cumulative percentage.
64	PS	Compute cumulative percentages.
65	PS	Answer interpretive questions about cumulative percentages.
Section 5		
66	CK	Definition of “graph.”
67	CK	Knowledge of the location and purpose of the x-axis.
68	CK	Knowledge of the location and purpose of the y-axis.
69	CK	Definition of “bar chart.”
70	PS	Create a bar chart.
71	PS	Answer interpretive questions about bar charts.
72	CK	Definition of “ratio” scale of measurement.
73	CK	Definition of “nominal” scale of measurement.
74	CK	Definition of “frequency polygon.”
75	PS	Create a frequency polygon.
76	PS	Answer interpretive questions about frequency polygons.
77	CK	Knowledge that frequency polygons can only be made for data on the interval or ratio scales of measurement.

Appendix B
Hierarchy of Procedural Skill Curriculum Elements



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.

Ackerman, P. L.: 1992, 'Predicting individual differences in complex skill acquisition: Dynamics of ability determinants'. *Journal of Applied Psychology* **77**, 598–614.

Alexander, P. A. and J. E. Judy: 1988, 'The Interaction of Domain-Specific and Strategic Knowledge in Academic Performance'. *Review of Educational Research* **58**(4), 375–404.

Anderson, J. R.: 1983, *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R.: 1987, 'Skill Acquisition: Compilation of Weak-Method Problem Solutions'. *Psychological Review* **94**, 192–210.

Anderson, J. R.: 1993, *Rules of the Mind*. Hillsdale, NJ: Erlbaum.

Bryan, W. L. and N. Harter: 1899, 'Studies on the telegraphic language: The acquisition of a hierarchy of habits'. *Psychological Review* **6**, 345–375.

Bunderson, C. V. and J. B. Olsen: 1983, 'Mental Errors in Arithmetic Skills: Their Diagnosis in Precollege Students'. Final Project Report No. NSF SED 80–125000: WICAT Education Institution, Provo, UT.

Carlson, R. A., and R. G. Yaure: 1990, 'Practice schedules and the use of component skills in problem solving'. *Journal of Experimental Psychology: Learning, Memory, & Cognition* **16**, 484–496.

Clancey, W. J.: 1986, 'Intelligent Tutoring Systems: A Tutorial Survey'. Report No. KSL-86–58: Stanford University, Stanford, CA.

Cronbach, L. J. and R. E. Snow: 1977, *Aptitudes and Instructional Methods: A Handbook for Research on Interactions*. New York: Irvington.

Derry, S. J. and S. P. Lajoie: 1993, 'A Middle Camp for (Un)Intelligent Instructional Computing: An Introduction'. In: S. P. Lajoie and S. J. Derry (eds.), *Computers as Cognitive Tools*. Hillsdale, NJ: Erlbaum, pp. 1–11.

Dillenbourg, P. and J. Self: 1992, 'A Framework for Learner Modelling'. *Interactive Learning Environments* **2**(2), 111–137.

Dochy, F. J. R. C.: 1992, *Assessment of Prior Knowledge as a Determinant for Future Learning*. Heerlen: Open University of the Netherlands.

Fisk, A. D. and W. A. Rogers: 1992, 'The application of consistency principles for the assessment of skill development'. In: J. W. Regian and V. J. Shute (eds.), *Cognitive approaches to automated instruction*. Hillsdale, NJ: Erlbaum, pp. 171–194.

Glaser, R.: 1984, 'Education and Thinking: The Role of Knowledge'. *American Psychologist* **39**(2), 93–104.

Kulik, J. A., C. C. Kulik and R. L. Bangert-Drowns: 1985, 'Effectiveness of computer-based education in elementary schools'. *Computers in Human Behavior* **1**, 59–74.

Kyllonen, P. C.: 1994, 'A Theoretical Framework for Cognitive Abilities Measurement'. In: D. Detterman (ed.), *Current Topics in Human Intelligence: Volume IV, Theories of Intelligence*. Norwood, NJ: Ablex.

Kyllonen, P. C. and R. E. Christal: 1989, 'Cognitive Modeling of Learning Abilities: A Status Report of LAMP'. In: R. Dillon and J. W. Pellegrino (eds.), *Testing: Theoretical and Applied Issues*. San Francisco: Freeman, pp. 112–137.

Kyllonen, P. C. and V. J. Shute: 1989, 'A Taxonomy of Learning Skills'. In: P. L. Ackerman, R. J. Sternberg and R. Glaser (eds.), *Learning and Individual Differences*. New York: Freeman, pp. 117–163.

Kyllonen, P. C. and W. C. Tirre: 1988, 'Individual Differences in Associative Learning and Forgetting'. *Intelligence* **12**, 393–421.

Kyllonen, P. C., D. J. Woltz, R. E. Christal, W. C. Tirre, V. J. Shute and S. Chaiken: 1990, 'CAM-4: Computerized Battery of Cognitive Ability Tests'. Unpublished computer program, Brooks Air Force Base, Texas.

Martin, J. D., and K. VanLehn: 1993, 'OLAE: Progress toward a multi-activity, Bayesian student modeler'. In: P. Brna, S. Ohlsson and H. Pain (eds.), *Artificial Intelligence in Education, 1993*. Charlottesville, VA: AACE, pp. 410–417.

Munro, A.: 1993, 'Authoring Interactive Graphical Models for Instruction'. In: D. M. Towne, T. de Jong and H. Spada (eds.), *Simulation-Based Experiential Learning*. Berlin: Springer-Verlag, Series F, Vol. 122, pp. 33–46.

Noetic Systems: 1993, '*ERGO* v. 1.2' (software), Baltimore, MD.

Regian, J. W. and V. J. Shute (eds.): 1992, *Cognitive Approaches to Automated Instruction*. Hillsdale, NJ: Erlbaum.

Schneider, W. and R. M. Shiffrin: 1977, 'Controlled and automatic human information processing: Detection, search, and attention'. *Psychological Review* **84**, 1–66.

Self, J. A.: 1989, 'The case for formalising student models (and intelligent tutoring systems generally)'. In: D. Bierman, J. Breuker and J. Sandberg (eds.), *Artificial intelligence and education: Synthesis and reflection*. Springfield, VA: IOS, p. 244.

Shute, V. J.: 1991, 'Who is Likely to Acquire Programming Skills?'. *Journal of Educational Computing Research* **7**, 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, pp. 15–47.

Shute, V. J.: 1993a, 'A Comparison of Learning Environments: All that Glitters... '. In: S. P. Lajoie and S. J. Derry (eds.), *Computers as Cognitive Tools*. Hillsdale, NJ: Erlbaum, pp. 47–74.

Shute, V. J.: 1993b, 'A Macroadaptive Approach to Tutoring'. *Journal of Artificial Intelligence and Education* **4**(1), 61–93.

Shute, V. J.: 1994, 'Learning Processes and Learning Outcomes'. In: T. Husen and T. N. Postlethwaite (eds.), *International Encyclopedia of Education* (2nd Edition). New York, NY: Pergamon Press, pp. 3315–3325.

Shute, V. J. and L. A. Gawlick-Grendell: 1994, 'What does the computer contribute to learning?'. *Computers and Education: An International Journal* **23**(3), 177–186.

Shute, V. J. and K. A. Gluck: 1994, 'Stat Lady: Descriptive Statistics Module'. Unpublished computer program, Armstrong Laboratory, Brooks Air Force Base, Texas.

Shute, V. J. and P. C. Kyllonen: 1990, 'Modeling Individual Differences in Programming Skill Acquisition'. Technical Paper AFHRL-TP-90-76. Armstrong Laboratory, Brooks Air Force Base, Texas.

Shute, V. J. and J. Psotka: in press, 'Intelligent Tutoring Systems: Past, Present, and Future'. To appear in D. Jonassen (ed.), *Handbook of Research on Educational Communications and Technology*. Scholastic Publications.

Shute, V. J. and J. W. Regian: 1993, 'Principles for Evaluating Intelligent Tutoring systems'. *Journal of Artificial Intelligence in Education* 4(3), 245–271.

Shute, V. J., L. A. Gawlick-Grendell, R. K. Young and C. A. Burnham: in press, 'An Experiential System for Learning Probability: Stat Lady Description and Evaluation'. To appear in *Instructional Science*.

Sleeman, D. H.: 1987, 'PIXIE: A Shell For Developing Intelligent Tutoring Systems'. In: R. Lawler and M. Yazdani (eds.), *AI and Education: Learning Environments and Intelligent Tutoring Systems (Vol. 1)*. Norwood, NJ: Ablex Publishing, pp. 239–265.

Sleeman, D. H. and J. S. Brown: 1982, *Intelligent Tutoring Systems*. London: Academic Press.

Sleeman, D. H., A. E. Kelly, R. Martinak, R. D. Ward and J. L. Moore: 1989, 'Studies of Diagnosis and Remediation With High School Algebra Students'. *Cognitive Science* 13(4), 551–568.

Snow, R. E.: 1990, 'Toward Assessment of Cognitive and Conative Structures in Learning'. *Educational Researcher* **18**(9), 8–14.

Swanson, J. H.: 1990, 'The Effectiveness of Tutorial Strategies: An Experimental Evaluation'. Paper presented at the annual conference of the American Educational Research Association, Boston, MA.

Swan, M. B.: 1983, 'Teaching Decimal Place Value: A Comparative Study of Conflict and Positively-Only Approache' (Research Rep. No. 31). Nottingham, England: University of Nottingham, Sheel Center for Mathematical Education.

Towne, D. M.: 1993, 'Teaching and Learning Diagnostic Skills in a Simulation Environment'. In: D. M. Towne, T. de Jong and H. Spada (eds.), *Simulation-Based Experiential Learning*. Berlin: Springer-Verlag, Series F, Vol. 122, pp. 149–164.

VanLehn, K.: 1990, *Mind Bugs: The Origins of Procedural Misconceptions*. Cambridge, MA: MIT Press.

Wenger, E.: 1987, *Artificial Intelligence and Tutoring Systems*. Los Altos: CA, Morgan Kaufmann Publishers.

White, B. Y. and J. R. Frederiksen: 1987, 'Qualitative Models and Intelligent Learning Environments'. In: R. Lawler and M. Yazdani (eds.), *AI and Education*. Norwood, NJ: Ablex Publishing, pp. 281–305.

Woltz, D. J.: 1988, 'An investigation of the role of working memory in procedural skill acquisition'. *Journal of Experimental Psychology: General* **117**, 319–331.

Author's Vita

Dr. V. Shute:
Armstrong Laboratory, 7909 Lindbergh Drive, Brooks A.F. Base, TX 78235-5352, U.S.A.

Valerie Shute received her M.A. (1981) and Ph.D. (1984) degrees from the University of California, Santa Barbara. Her multidisciplinary studies centered around: cognitive psychology, artificial intelligence, and statistics. After receiving her doctorate, Dr. Shute accepted a post-doctoral position at the University of Pittsburgh, developing intelligent microworld environments. From 1986 to present, Dr. Shute has been employed at the Armstrong Laboratory, Brooks Air Force Base, Texas, conducting basic research on cognitive process measures, as well as designing, developing, and evaluating intelligent tutorial systems. She has also been involved in conducting tests of aptitude-treatment interactions using the controlled environments offered by intelligent tutoring systems, and is currently involved in student modeling research. Dr. Shute has written over 65 technical papers and chapters, and co-edited a book on cognitive processes and automated instruction.