SMART Evaluation: Cognitive Diagnosis, Mastery Learning & Remediation

Valerie J. Shute
Armstrong Laboratory (AL/HRTI)
1880 Carswell Avenue
Lackland Air Force Base, TX 78236-5507, USA
e-mail: vshute@colab.brooks.af.mil

Abstract

SMART is the name of a student modeling approach that applies regression equations to learners' actions to predict knowledge and skill level. Ensuing instruction is based on this diagnosis. This paper presents an overview of the paradigm, followed by the results from two controlled evaluation studies where the main components (i.e., diagnostic updating routines and mastery/remediation control structures) are systematically tested. Study 1 examined the predictive validity of the diagnostic component, with mastery and remediation features disabled. Overall, the predicted student model values, in conjunction with an aptitude measure, accounted for a large amount of outcome variance. Furthermore, learners showed a 2.2 SD pretest-to-posttest improvement following only 4 hr of instruction. Study 2 examined the effects of mastery criterion and remediation on learning outcome and efficiency. Results indicated that posttest scores were even higher than in Study 1, but at the expense of greater learning time.

The purpose of this paper is to overview a student modeling paradigm called SMART (Student Modeling Approach for Responsive Tutoring), then provide a summary of some preliminary results from two controlled evaluation studies examining its predictive validity and contributions to learning.

Overview of SMART

SMART (Shute, in press) relates to the dynamic interplay between cognitive diagnosis and remediation, thus it's broader than most student modeling approaches that focus solely on diagnosis. 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 (see Shute, 1993). Finally, whereas most other approaches focus on single outcome types (e.g., model-tracing for procedural skill acquisition), SMART models a range of outcome types, including: symbolic knowledge (SK), procedural skill (PS), and conceptual knowledge (CK).

This paradigm may be implemented within any computer-based instructional system where low-level knowledge and skills (i.e., curriculum elements, or CEs) have been identified from a cognitive task analysis, and arranged in an inheritance hierarchy. For instance, SMART has recently been embedded within Stat Lady, an automated experiential learning environment teaching descriptive statistics (Shute & Gluck, 1994). To illustrate the three outcome types within this domain, consider one's knowledge of the Mean (a measure of central tendency). SMART assesses: (a) SK--being able to construct the formula for the Mean ($\Sigma X/N$), (b) PS--being able to collect data from the on-line "Number Factory" then compute the Mean from the obtained set of data using the on-line "Calculator," and (c) CK--distinguishing the definition and characteristics of the Mean from the Median and Mode, and showing an understanding that the Mean refers to the average or "balance point" within a distribution of data using the on-line "Grab a Graph" tool within the system. It is also being implemented within other systems, as well (e.g., a program teaching concepts associated with predicting underwater sound transmission paths and ranges). But the important empirical question is: Does inclusion of SMART in a program enhance learning beyond that afforded by simple computer-based training? More specific questions address particular features of SMART (diagnosis, mastery and remediation) and their relationships with different learning outcome and efficiency measures. These questions will be addressed following an overview of the model.

The chain of events in SMART is as follows: (a) Initialization: Each CE is initially evaluated following completion of an on-line pretest which yields a specific *p(CE)*--probable mastery value for a given CE; (b) Updating: Each CE is subsequently evaluated during problem solution and question-answering within the tutor; (c) Mastery or Remediation: Each p(CE) value provides the basis for deciding whether ensuing activities will involve new instruction, remediation, or continued practice on the current material, each of which represents distinct activities that

differ in essence (new instruction) or subtly (remediation or continued practice) from the current activity, depending on the diagnosis. Remediation (if indicated) is intended to be precise because each CE knows its exact location in the tutor (i.e., where it is instructed and evaluated). Data management occurs via a straightforward array of records, where each array element maps onto an individual curriculum element. All CE information is maintained within the record (e.g., its outcome type, the location in the tutor where it's instructed and assessed, its parents, siblings, and children CEs, and so on). Specific mechanisms underlying the student model will now be discussed.

Initialization

Initial values for SMART are obtained from a student's performance on a comprehensive, on-line pretest designed to assess incoming knowledge of *all* curriculum elements resulting from the cognitive task analysis (separated into three major outcome types: SK, PS, and CK). 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. An isomorphic posttest was also developed which contains the same collection of CEs, but with different questions, per CE.

After the computer scores all test items, it communicates those values directly to the student model for initialization. Once the student model is initialized, a learner is placed in the curriculum at the beginning of instruction for a CE that they do not, or only partly, know. In other words, learners are able to skip CEs that they demonstrated an understanding of via their pretest performance. Further, the mastery criterion can be set at the outset of instruction to be greater than some value (e.g., .70). Any CEs with values falling below this threshold are candidates for instruction.

In the tutor, learning about a particular CE involves instruction and then relevant problem solving, during which time the learner can solve a problem correctly, 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), which is provided in response to erroneous inputs; it can not be explicitly requested. Each level of assistance is associated with progressively more explicit feedback to the learner, from the tutor's simply stating that the answer is incorrect, to identifying a particular bug in the solution, to finally, giving the learner the correct answer and telling him/her to input this solution (see Shute, in press, for more on the topic of error analysis and feedback). The simple presumption is: The more help required by the learner, the less understood the current CE. After a problem is completed, the relevant curriculum elements within the student model are updated. The updating scheme is outlined below.

Updating SMART values

The updating heuristics are based on a series of regression equations, which in turn are based on the level of assistance the computer gives each person, per curriculum element, during problem solution. The student model began with the identification of six main states: Remedial, intermediate, and mastery, with low vs. high divisions within each. Subsequently, a subject-matter expert (with over 30 yr of Statistics teaching experience) and I generated a set of rules for updating the model, per CE. These rules related to promotion and demotion through the curriculum, as a function of the levels of assistance needed. That is, if a person answers a problem without any tutorial assistance (level-0), that person should get a larger boost in his/her student model value compared to someone who required considerably more help from the system.

The simple promotion/demotion rules ultimately spawned four regression equations, each a function of the required assistance and used to estimate a learner's p(CE). The transformation from discrete rule to continuous function involved the assignment of a range of specific values to each rule, where the values ranged from 0 to 1 in sixths, corresponding to each of the six state divisions (e.g., If current state = high remedial = .17 to .33, and the student solved a problem with level-0 assistance, then new state = high intermediate = .50 to .67). Preliminary functions evolved from plotting the four levels of assistance by six states of mastery. The b-weights used in the regression equations were obtained from computing best-fitting curves on these preliminary functions (linear, quadratic, and cubic trends). These equations now provide a quick and easy way to update student model values. For example, to compute a new CE value, the system just needs to know the learner's 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 during a problem-solving episode. This gets plugged into the appropriate equation, and the new value results, allowing for a continuous representation of probable mastery values. The final graph of the best-fitting curves, plotted along with the discrete points, is presented in Figure 1.

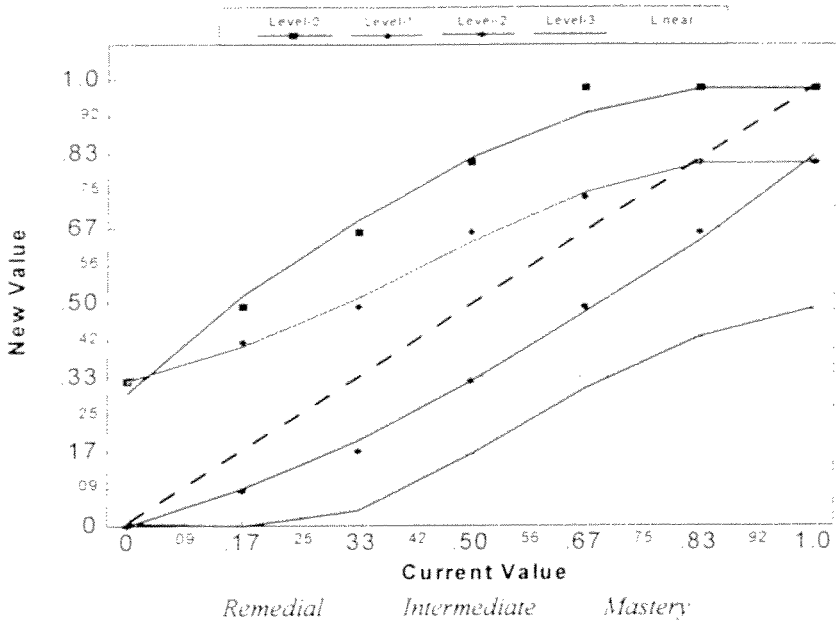

Figure 1. SMART's Updating Functions

There are several things to notice about Figure 1. First, the three main regions (i.e., dashed lines separating remedial, intermediate, and mastery boundaries) are arbitrary; they may be set at any value (e.g., mastery ≥ .67 or ≥ .83). Second, the curves themselves show some interesting properties, such as ceiling effects (level-0 curve) and floor effects (level-3 curve). Third, the best-fitting curves match the trends defined by the rule-based points quite well. While that is not entirely surprising (because they *are* best-fitting curves), it is encouraging that the fits are so close. Finally, note that these p(CE) values do not represent grades (although they are determined from individuals' response histories in the solution of tutor-specific problems). Rather, 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 1.00 would denote 100% confidence that a learner has achieved mastery, .67 indicating a more moderate degree of assurance that mastery has been attained, and so on.

Mastery or Remediation

When a learner attains mastery on all of the CEs in a given problem set, he/she is advanced to the next instructional section of the tutor containing CEs with below-mastery values (determined by pretest, or tutorial problem-solving data). But in the case where remediation is indicated, what happens is as follows. Each problem set evaluates several CEs, concurrently. 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 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 < .60 and has decreased from the previous value, then invoke remediation, or (b) If a CE has been remediated and the CE is < .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 .83 (too high), and remediation did not get invoked until meeting a "< .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. These represent the first of a series of controlled evaluations of SMART, specifically testing each of SMART's features. For instance, 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).

Study 1 Design and Results

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. 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 ≤ 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 represented 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, between 4-5 hours learning curriculum that focused on data organization and plotting. There were 77 CEs instructed and assessed in this segment of the tutor. Following instruction, all subjects completed the on-line posttest which was similarly decomposed into all CEs, tested in duplicate.

Learning

The first research question examined by this study explores the degree of learning that resulted from Stat Lady instruction (i.e., *without* SMART's mastery criterion and remediation features invoked). A t-test was computed on the pretest and posttest data, and the results showed this difference to be significant: Pretest $\underline{M} = 44.1\%$ (SD = 14), Posttest $\underline{M} = 75.3\%$ (SD = 15); $\underline{t}\,(103) = 25.75$, $\underline{p} < .001$. Thus, Stat Lady alone managed to elevate subjects, overall, from an incoming score of 44% to a final score of 75%, representing an increase of 2.2 standard deviations. This now becomes the "score to beat" in subsequent studies of Stat Lady/SMART where mastery criterion and remediation features are turned on. While these learning gains appear to be quite impressive, it should be noted that a posttest score of 75% correct is still not perfect, so there is room for improvement via SMART's mastery learning approach and provision of appropriate remediation.

Predictive Validity of the Diagnostic Component

The second question examined the goodness of fit between SMART's diagnostic heuristic (i.e., the p(CE) values) and the outcome data, analyzed in conjunction with individual differences variables in the equation. Some of these variables were extracted from the on-line demographic questionnaire, designed to assess educational background, gender, age, and so on. The cognitive ability measurement (CAM-4) battery provided even more individual differences measures, namely: working memory capacity, associative learning skill, inductive reasoning, and information processing speed (Kyllonen, Woltz, Christal, Tirre, Shute, and Chaiken, 1990). To reduce these data, I computed a principal components analysis on just the aptitude (cognitive) data. Four variables were used in the analysis (i.e., test scores from the quantitative subtests, above), and a single factor was extracted. The four variables accounted for 66% of the *aptitude* factor variance, and factor loadings were all high.

The primary research question concerned the accuracy of the computed student model values, p(CE)s, in predicting outcome CE scores. To test the goodness of fit of SMART's computed knowledge and skill mastery indices (i.e., diagnostic validity), 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 variable to enter into the equation was p(CE), with a multiple R = .73 (i.e., 54% of the unique outcome variance was explained by this variable alone). Next to enter the equation was aptitude, increasing the multiple R to .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 .82. None of the other variables reached criterion for inclusion in the equation. The final data are: p(CE) ($t = 5.01$, $p < .001$), aptitude ($t = 4.25$, $p < .001$), and pretest, $t = 2.90$, $p < .005$). A scatterplot of just the p(CE) and aptitude data predicting posttest score are shown in Figure 2.

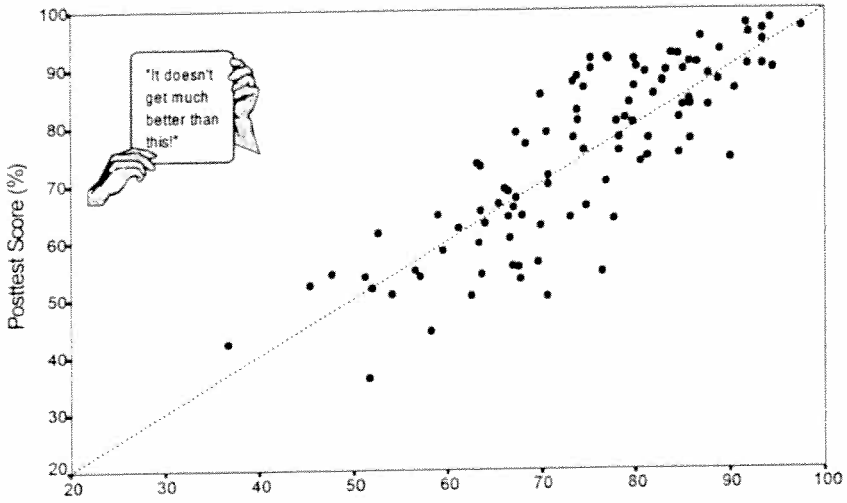

Figure 2. Scatterplot of p(CE) and Aptitude Data Predicting Posttest Score

Macroadaptation

The fact that aptitude significantly contributed to the equation, beyond that accounted for by p(CE), justifies the macroadaptive approach, and can prompt curve adjustments. A main effect of aptitude was reported above, but a full test of main effects and parameter interactions is needed to advise curve modifications such as those shown in Figure 3 (representing hypothetical adjustments). The graph on the left shows how a given curve may be systematically elevated or depressed given general aptitude information, and the graph on the right illustrates a possible interaction between aptitude and the *slope* of the curve. Ensuing p(CE) values, resulting from the modified regression equations, can subsequently reflect knowledge and skill mastery levels more accurately.

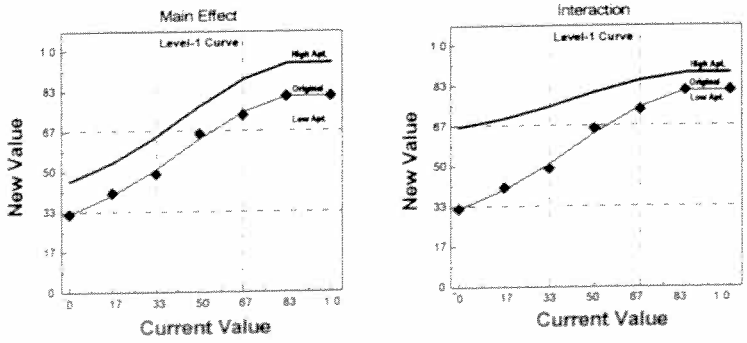

Figure 3. Hypothetical Macroadaptive Curve Adjustments

So, subjects in Study 1 learned a lot from the tutor, and the computed student model values turned out to be good predictors of outcome performance. But how did the students feel about their Stat Lady learning experience?

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 4, below). 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 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.

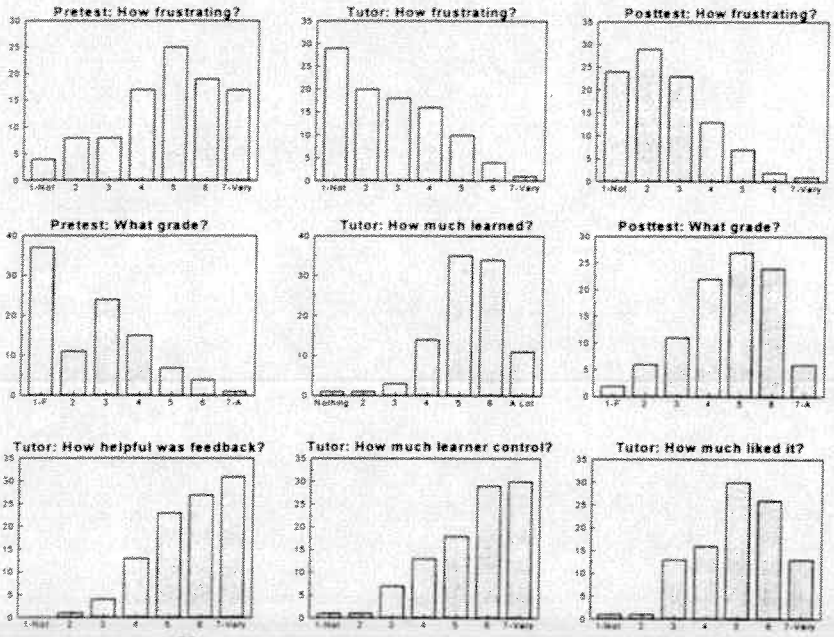

Figure 4. Affective Data Related to Testing and Tutorial Experiences

Study 2 Design and Preliminary Results

In addition to cognitive diagnosis, the second important component of SMART involves remediation as a function of the mastery requirement. These conjoined features have just been 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?

There were a total of 107 subjects in this study which spanned 2-3 days (dependent on the remediation requirements, per subject). Subjects were obtained from 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. Again, while all subjects were required to have a high school education, previous statistics coursework was restricted to ≤ 1 completed class.

Learning Improvements

The first question concerns how much (if anything) is added to learning outcome beyond that shown in Study 1 where the pretest to posttest gain = (posttest-pretest)/pretest score = 73% or 2.2 SD. First, it should be pointed out that, due to chance alone, Study 2 subjects started the tutor with slightly higher pretest scores ($\underline{M}$ = 50.3%) compared to subjects in Study 1 ($\underline{M}$ = 43.1%). To analyze differences in outcome performance, I thus needed to compute an ANCOVA on the posttest data with pretest as a covariate, and condition as a between-subjects variable (Note: the respective posttest scores were $\underline{M}$ = 75% and $\underline{M}$ = 83% for Studies 1 and 2, respectively). Results showed that the outcome difference between conditions was significant: $\underline{F}$ (1, 207) = 4.55; $\underline{p}$ < .04. Thus, we can conclude that remediation is helping learners acquire even greater degrees of domain-specific knowledge and skill. But we still have to ask "at what cost" with regard to training/learning time.

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 understood or knew how to do (like sorting data into ascending and descending order). Furthermore, Study 1 subjects were required to solve a fixed number (two) of 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 have required *less* time to complete the tutor, given that fewer CEs were instructed. On the other hand, remediation and mastery were operating in Study 2, which could have served to *add* time to the learning process. The actual tutor completion times were 4.4 hr and 7.5 hr for Studies 1 and 2, respectively.

Combining the learning outcome and efficiency data into a simple benefit-cost ratio reveals the following: As mentioned, Study 1 subjects increased 73% from pretest-to-posttest and required only 4.4 hr to attain that level of proficiency. On the other hand, subjects in Study 2 showed a 64% improvement from pretest-to-posttest, and required 7.5 hr to achieve that state. The simple ratio comparing percentage of improvement to learning time reveals that subjects in Study 1 demonstrated an average improvement rate of 16.6%/hr vs. subjects in Study 2 who showed an improvement rate of only 8.5%/hr.

While it may be tempting to conclude that the version of Stat Lady used in Study 1 (technically, computer-based training) represents a more efficient instructional tool compared to the intelligent version of the program, it should be pointed out that: (a) these data represent preliminary, global-level analyses, and (b) the results suggest specific ways to improve the remediation feature in SMART. That is, remediation was intended to be surgical and precise, but in reality, subjects often had to wade through extra instructional pages (to establish context), before receiving remedial instruction for the problematic CE. Furthermore, while there is a one-to-one link between questions in the problem set and specific, individual CEs, the system does not currently test single CEs independent of others in an integrated problem set. Although it is *possible* to remediate CEs individually, once a learner is returned to a problem set, they are currently re-tested on *every* CE in that section of the curriculum. This results in a substantial amount of time being added to the learning/assessing process, which may be easily reduced in the future. Overall, subjects in both studies showed impressive gains in a relatively short amount of instructional time. Finally, before dismissing the contribution(s) of intelligence, additional studies are needed, a few of which are listed below.

Conclusions and Future Directions

In summary, the unique contributions of SMART include: (a) modeling *emerging* as well as *incoming* knowledge and skills for broader, more accurate diagnoses, and (b) distinguishing among a range of knowledge and skill outcome types for differential representation and remediation. The results from the current round of controlled evaluations have been quite encouraging and have also suggested specific ways to enhance the system's performance. Study 1 showed that SMART's diagnostic accuracy is already quite high, overall, and may be improved even further. That is, results from the multiple regression analysis showed the computed student model values are singularly the most robust predictor of posttest performance, and with aptitude in the equation (Figure 2) the predicted values lie strikingly close to the actual posttest scores. Although the predictive validity is not perfect, I've suggested ways to utilize the empirical data (e.g., aptitude information) to increase p(CE) precision by adjusting regression equations.

Study 2 indicated that when the mastery criterion and remediation features are enabled, 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 example, Shute (in press) discusses a significant 3-way interaction involving learning gain × outcome type × aptitude level.

As Study 1 reported, and Figure 2 depicted, average p(CE) and posttest data (collapsed across all CEs) 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. For instance, it will be informative to see if the curves consistently (or differentially) result in estimates that either over- or underestimate the actual posttest values. If so, then systematic curve adjustments are called for, achieved by altering b-weights, accordingly.

Further refinements and extensions to the program that are currently being implemented include: (a) employing a Bayesian network of hierarchically-arranged CEs for additional inferencing capabilities, (b) programming an adaptive pretest based on IRT (item-response theory) to assess intermediate-level CEs and make deductions about lateral and lower-level ones, as well as derive inferences about related, more complex CEs, (c)

utilizing different "flavors" of remediation, per outcome type, and (d) testing the generalizability of the paradigm across a variety of domains.

Bayesian Networks. Student model values may be propagated through an inheritance hierarchy to provide estimates of mastery for complex CEs that contain lower-level ones as components. Software exists (e.g., Noetic Systems, ERGO, 1993) that enables one to easily establish a network of related nodes (CEs) in terms of conditional probability distributions. The network is updated based on probabilistic relations. I've recently acquired this software and have begun generating hierarchical Bayesian nets, separately by outcome type. An empirical test is underway to examine predictive validity of network data beyond that already shown by SMART's regression-driven p(CE) values.

Flavors of Remediation. Knowledge and skill acquisition may be enhanced by using different kinds of remediation; capitalizing on the best aspects of theoretically-grounded modeling approaches by pairing them with appropriate types of knowledge/skill remediation. For instance, SK elements are remediated in SMART by drill and practice, while PS elements are remediated in the context of problem solving. However, CK instruction/remediation currently (and simply) involves re-instructing problematic CE(s), followed by re-assessment within the context of new problem sets. Alternatively, explicitly instructing CK via analogies (for either initial or remedial instruction) may enhance the memorability of CK elements and also help learners solve related PS elements. While the system already employs some analogies for CK instruction (e.g., animated slices of pie to illustrate proportions), I am now systematically creating analogies for *all* CK elements (e.g., using a see-saw analogy with a manipulable fulcrum for instructing the Mean).

Adaptive Pretest. While the current pretest assesses knowledge/skill on every CE in the tutor, a more efficient procedure would be to utilize the already-established CE hierarchies as the basis for a more adaptive pretest. Then inferences could be made about a learner's actual understanding of one (or more) CE(s) based on performance data from related CEs. The result would be a much shorter pretest, which would have the added benefit of being less negatively perceived (see Figure 4). For instance, 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 (actual $r_{xy} = 1.0$, $N = 104$ between these two CEs). The strength of the inference would depend on the relatedness of the CEs in question.

Generalizability of the Paradigm. 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 example, 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) that teaches 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 curriculum. We (John Schuler and I) are currently incorporating SMART into the delivery system, and the process has been pleasantly straightforward.

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. The current and planned research will result in an iteratively-enhanced system that can optimize learning outcome and efficiency across a broad range of learners and outcome types.

References

Kyllonen, P. C., Woltz, D. J., Christal, R. E., Shute, V. J., Tirre, W. C., & Chaiken, S. (1990). *CAM-4: Computerized battery of cognitive ability tests.* Unpublished computer program, Brooks Air Force Base, TX.

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

Shute, V. J. (in press). SMART: Student Modeling Approach for Responsive Tutoring. To appear in: *User Modeling and User-Adapted Interaction* (Special issue on student modeling).

Shute, V. J. (1993). A macroadaptive approach to tutoring. *Journal of Artificial Intelligence & Education, 4*(1), 61-93.

Shute, V. J., & Gluck, K. A. (1994). *Stat Lady: Descriptive Statistics Module.* [Unpublished computer program]. Brooks Air Force Base, TX: Armstrong Laboratory.