Instructional Science **24:** 25–46, 1996.

An experiential system for learning probability: Stat Lady description and evaluation

VALERIE J. SHUTE[1], LISA A. GAWLICK-GRENDELL[2],
ROBERT K. YOUNG[3] & CLARKE A. BURNHAM[3]
[1]*Armstrong Laboratory* [2]*Galaxy Scientific, Inc.* [3]*University of Texas at Austin*

Abstract. This paper describes a computer-based instructional system called Stat Lady, and reports the results of an evaluation study that tested the efficacy of learning probability from this program in relation to a traditional Lecture and a no-treatment Control group. Results showed that both treatment groups learned significantly more than the Control group, yet there was no difference between the two treatment groups in outcome performance after three hours of instruction. This was viewed as encouraging because: (a) due to sampling error, students assigned to the Stat Lady condition were at a disadvantage, scoring about 20 points less on the quantitative SAT measure compared to the Lecture group, and about 25 points less than the Control group, (b) the lecture constituted a more familiar learning environment for these subjects, and (c) the professor administering the Lecture had more than 20 years experience teaching this subject matter, while this was Stat Lady's first teaching assignment. We also found a significant aptitude-treatment interaction where high-aptitude subjects learned more from Stat Lady than from the Lecture environment, but for low-aptitude subjects, there was no difference in learning outcome by condition. Implications of these findings will be discussed in relation to future computer-based instructional research.

Stat Lady is the name of a series of computerized experiential learning environments teaching topics in introductory statistics (e.g., probability, descriptive statistics). 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.

According to constructivist research, learners actively construct new knowledge and skills from what's already known. They do not come to a learning situation with a *tabula rasa*, but rather, as active-pursuers of new knowledge that will interweave with old knowledge (e.g., Bartlett, 1932; Collins, Brown, & Newman, 1989; Piaget, 1954). This construction process is believed to be enhanced by environments supporting experiential learning. Research in this area suggests that knowledge derived experientially is more memorable than passively received knowledge because the experience ('doing' rather than 'receiving') provides cognitive structure, and is intrinsically motivating and involving (e.g., Friedman & Yarbrough, 1985; Shute & Glaser, 1991). Furthermore, situating instruction within interesting and real-world problem-solving

scenarios is believed to enhance a topic's meaningfulness (Brooks, 1991; Brown, Collins, & Duguid, 1989; Clancey, 1992; Lave & Wenger, 1991; Suchman, 1987).

The general goal of the system is to enhance the acquisition of statistical knowledge and skills by making the learning experience both memorable and meaningful. To accomplish this, we started by creating a hands-on, experiential learning environment where learners actively engage in various on-line activities (e.g., painting cars, arranging sunbathers in lounge chairs, rolling dice and betting on the outcome). Moreover, each problem set was designed to be both entertaining (to impact memorability) and real-world related (to render it more meaningful).

Stat Lady takes a mastery learning approach to teaching, and learning is self-paced. That is, a relevant concept or rule is presented, then a problem is posed for students to solve in order to demonstrate comprehension (mastery) of each curriculum element. Learners are required to solve at least two problems correctly before advancing to the next element, but they may elect to solve more problems from a pool of related problem sets. If a learner gets a problem wrong, feedback is provided specific to the concept, formula, or rule that was in error, and learners are instructed to try again. Despite these features, Stat Lady is clearly not 'intelligent' in comparison to intelligent tutoring systems (cf. Shute & Psotka, in press); merely clever. That is, the version of Stat Lady presented in this paper does not engage in microadaptive, real-time diagnosis and remediation, but the program is sensitive to (and thus addresses) errors that are recognized by the system's 'bug' library (discussed in a later section).

The following research questions were investigated in this study: (a) Overall, what was the degree of change in subject's test scores from pretest to posttest (i.e., test-retest improvements)? (b) Given some general increase, was there differential pretest-to-posttest change (i.e., learning) as a function of condition? Specifically, did the two treatment conditions differ between themselves in terms of relative improvement? (c) Was there a main effect (or interactions) involving gender on pretest-posttest scores? This was of interest as females are often at a quantitative disadvantage, especially during high school and college (e.g., Hyde, Fennema, & Lamon, 1990). If there were differences, what might explain the obtained gender effect on outcome performance? Finally, (d) Was there evidence of any aptitude-treatment interaction?

A Control group was needed to provide baseline data assessing the first research question on test-retest changes (in the absence of instructional intervention). The next set of questions addressed whether a simple (i.e., non-intelligent), experiential learning environment could help students acquire a comparable degree of knowledge and skill as learning the same

material from an experienced human lecturer. Our hypothesis was that, even in the absence of any programmed 'intelligence,' Stat Lady could support the acquisition of knowledge and skills, possibly on par with the lecturer. In regard to gender effects on learning, we hypothesized that there would be a main effect of gender (with a male superiority), but no interaction involving treatment condition. In other words, we had no *a priori* reason to believe that our two treatment conditions would yield any differential advantages/disadvantages for males vs. females. Finally, with fixed learning time (3 hr) and identical curriculum underlying the two treatment conditions, we examined the interaction between condition and aptitude using subjects' (self-reported) SAT scores. We posited that high aptitude subjects would perform better in the Stat Lady environment given a greater facility in adapting to novel circumstances, while low aptitude subjects would be more comfortable (and hence learn better) within the familiar, structured Lecture environment.

This paper will outline the program in detail, and then describe a study that tested the efficacy of Stat Lady's probability module compared to classroom instruction on identical curricular elements.

Method

Subjects

The subjects used in this study were undergraduate psychology students attending the University of Texas at Austin (N = 168) and participating in the study to satisfy a course requirement. The number of subjects per group was: Stat Lady = 63, Lecture = 36, and Control = 69.

Subjects assigned themselves to the various groups by signing up for the 5-hour experiment from a large pool of available experiments. They did not know the content of the condition they were signing up for, only the scheduling of the three conditions. The Lecture group required subjects to participate at a *specific time* for one hour on each of the five weekdays (i.e., Monday through Friday). In contrast, the Stat Lady condition (run on 10 computers) allowed subjects to come in at any time, between the hours of 8:00 a.m. and 6:00 p.m., for one hour on each of the weekdays. With the greater flexibility in scheduling, the Stat Lady condition was able to attract a larger number of volunteers than was the Lecture condition, hence the difference in sample sizes, per condition. The Control group participated later in the semester than did either the Lecture or Stat Lady groups, and signed up for only two hours of research. That is, they were only required to take the pretest (on a Monday) then the posttest four days later (on Friday). There was approximately an equal number of males and females, per group, with an age range from 17 to 27 years.

Materials

Program description

The Stat Lady module used in this experiment instructed probability concepts and rules (Shute & Gawlick-Grendell, 1992). It was written in Visual Basic 2.0 and runs on 386/486 computers under Windows 3.1. As mentioned, Stat Lady provides a very experiential learning environment with real-world scenarios and enticing displays (e.g., color, animation, sound), thus empowering and encouraging learners rather than simply providing formulas to memorize or tables of numbers to manipulate. Concepts to be learned are embedded in familiar situations (to draw on prior knowledge), and examples vary sufficiently to show the range/limits of applicability.

Knowledge base

Stat Lady's knowledge base was initially developed through careful inspection of six introductory statistics textbooks. The curriculum embodied within these textbooks typically spanned an overlapping range of knowledge and skills. A course outline of Stat Lady curricular elements was derived from the most commonly instructed probability concepts. That is, similar to the instructional content and sequencing within the majority of the textbooks, Stat Lady's curriculum began with the explication of simple concepts (e.g., sample space, elementary events, event class, sure events, mutually exclusive), then progressed to the instruction of various probability rules (i.e., addition rule, complementary rule, multiplication rule, and conditional probability of related and unrelated events). Finally, Stat Lady introduced several counting rules (i.e., Number of possible sequences for N trials, For sequences, Permutations, Ordered combinations, and Combinations) which ultimately led to the most difficult section involving Binomial Expansion. Several statistics experts reviewed the outline, and concurred that it represented a comprehensive curriculum for an introductory probability course at the college level. An overview of curricular elements is shown in Appendix 1.

Instructional examples

For each curricular element, we developed creative examples to illustrate and instruct each concept, as well as to capture the learner's interest. Furthermore, besides being an expert Statistician, Stat Lady is also an avid story-teller. She[1] begins teaching each concept or skill with a formal definition, then

[1] We chose to use the pronoun 'she' (rather than 'it') to emphasize the personality of the tutor over the code of the program.

immediately shares a 'true story' that just happens to contain the concept or skill currently under instruction. To illustrate a typical learning episode, Stat Lady first formally introduces a concept, such as Counting Rule 4: The number of ways of selecting and arranging r objects from among N distinct objects $= \frac{N!}{(N-r!)}$. (Note: In this case, $r \leq N$ must be true). Stat Lady then presents the learner with a 'true story,' recalling how she used to play Musical Chairs at birthday parties as a child. Learners are reminded how the game is played (e.g., five people dance around four chairs until the music stops, then sit in the closest chair; the person who doesn't get to a chair is eliminated from the game). This prompts the following question: '*Just how many different ways can N = 5 people be arranged in the r = 4 chairs*?' Stat Lady explains how this counting rule works ('There are 5 ways that the first chair can be filled, but only 4 ways that the second chair may be filled, because 1 of the 5 players is already seated'), then shows learners, step-by-step, how to work the equation and answer the question: $\frac{N!}{(N-r)!} = \frac{5!}{(5-4)!} = \frac{5!}{1!} = \frac{5\times4\times3\times2\times1}{1} = 120$. She also points out that another way to answer the question is to make a list of all possible combinations and count them up. Then it's the learner's turn to apply the rule in the solution of a new problem. This transitions into a problem-solving scenario.

Problem-solving scenarios

Real-world scenarios were created for learners to practice their evolving knowledge and skills that were acquired from the instructional examples. Following from the example illustrated above, Stat Lady describes another true story, this one about her nephew, Clarence, a maintenance worker at a local hospital. One day, while Clarence was at lunch in the cafeteria, a woman rushed into the hospital on the verge of delivering her baby. She was directed to go to the first room on the right. However, that door was locked, so she went into the second room, and the door automatically locked behind her. The doctor went into the third room to get some equipment for the delivery, and that door too accidentally closed and locked behind him. None of the doors could be opened without keys, and Clarence was the man with the keys. Clarence knew that two of his keys would unlock the doors, but unfortunately, the keys were unlabeled. Now that the scenario is set up, the question is posed: '*How many different arrangements of 2 keys to 3 doors are there*?'

Three numbered doors appear on the screen (as buttons), along with buttons representing a white and a black key. The learner fills in a table of possible door-key pairs by buttoning first on a door and then on a key. This results in a picture of that selected pair appearing in the table (see Figure 1). The learner may not answer the question until the table has been completed with

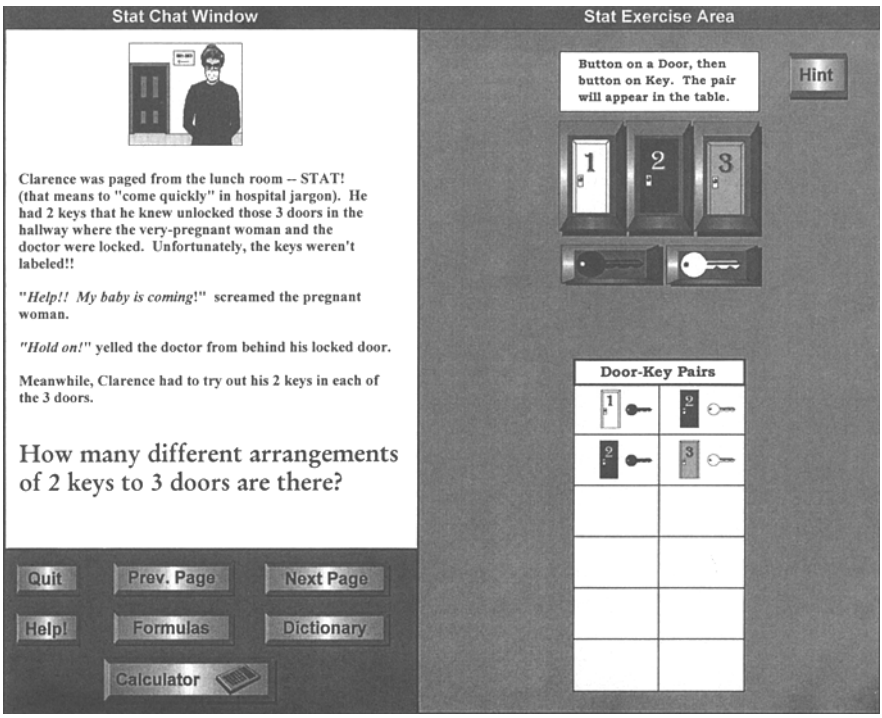

Figure 1. Example Stat Lady problem for Counting Rule 4: Doors and keys

all possible combinations (Door 1-Black Key & Door 2-White Key, Door 2-Black Key & Door 3-White Key, etc.). For learners having difficulty coming up with all of the unique pairings, a 'Hint' button is available (e.g., *'I see that you haven't selected the pair: Door 3-Black Key & Door 1-White Key'*). Once the table is completed, the learner is prompted to enter the correct answer to the initial question by using the on-line calculator: *'How many different arrangements of 2 keys to 3 doors are there*?' The answer may be figured out by applying the relevant counting rule ($\{3 \times 2 \times 1\} / 1 = 6$), simply counting the entries in the table (6 sequences of two doors-with-keys), or doing both.

A wrong response can occur if: (a) the wrong rule is computed, (b) the right rule is computed incorrectly, or (c) an incorrect count is made. In any of these cases, Stat Lady provides context-sensitive feedback. That is, the program knows about many possible mistakes that could be made for this particular problem (e.g., $2 + 3 = 5$, or $3^2 = 9$, etc.). If any of these known 'bugs' occur, they are specifically addressed. If the wrong answer is not in her 'buggy library' (described below) Stat Lady provides feedback by restating

Table 1. Example of buggy library of known mistakes for 'at least one ...' problem.

Response	Feedback
0.10	Hmmm ... your answer is incorrect! 0.10 refers to the chance of grabbing ONE Special Dark from the bag. The question asks about the probability of getting *at least one* out of two Special Darks. You need to sum the pairs which include at least one SD. Please try again.
0.01	Hmmm ... your answer is incorrect! 0.01 refers to the chance of grabbing TWO Special Darks from the bag. The question asks about the probability of getting *at least one* out of two Special Darks. You need to sum the pairs which include at least one SD. Please try again.
0.20	Hmmm ... your answer is incorrect! It looks like you may have added the {SD & SD} pair twice. After you computed the probabilities for each pair, you should have summed 7 (not 8) probabilities. Please try again.
0.18	Hmmm ... your answer is incorrect! It looks like you only added the pairs with just *one* Special Dark in the pair. The question asks about the probability of getting *at least one* out of two Special Darks, so you need to include the {SD & SD} pair. Please try again.
1.00	Hmmm ... your answer is incorrect! Remember, a probability of 1.00 refers to an event always happening. But there are events in your table that do not contain any Special Darks (for example, MC & MC or MG & K). This problem asks about the probability of getting at least one out of two Special Darks. You would need to sum the pairs which include at least one SD. Please try again.
0.19	Isn't that special? You are correct! 0.19 is the probability of grabbing at least one Special Dark from the bag. You summed the probabilities which included at least one SD: {MC & SD} or {MG & SD} or {K & SD} or {SD & SD} or {SD & MC} or {SD & MG} or {SD & K}.

the relevant counting rule and requiring the learner to re-do the problem. If the answer is correct, Stat Lady congratulates the learner, reiterates the counting rule, and may even perform a cartwheel if the problem was particularly difficult.

Buggy library

In addition to having knowledge of probability concepts and the ability to solve probability problems, Stat Lady's knowledge base also contains a library of known (and likely) incorrect responses. For each question that students are required to answer, we determined in advance (from pilot testing the system and consulting with statistics experts) a range of conceivable mistakes that

students commit in solving different problem types. Then, for each mistake, we generated specific feedback. To illustrate, consider the distribution of chocolate bars within a bag containing four different types of candy (e.g., 20 Milk Chocolates, 15 Mr. Goodbars, 10 Krackels, and 5 Special Darks). Students are posed the following question: '*Suppose you grabbed a chocolate bar, put it back, and grabbed another one. What's the probability that at least one of them will be a Special Dark?*' To answer this question, students must first determine all possible combinations of two chocolate bars (i.e., 16 candy pairs), then compute each probability, per pair. For the initial listing of paired chocolates, an empty table exists, along with pictures (buttons) of each of the four types of candy. Buttoning on any candy places it into the table as one of a pair. The table is completed when all 16 slots have been filled with different candy bar pairs. Following this specification of candy-pairs, they then need to fill in each of the associated probabilities. For instance, the probability of getting a Milk Chocolate and a Special Dark would be: $20/50 \times 5/50 = 0.04$ (application of the Multiplication Rule). Finally, they must add up all probabilities associated with the conditions containing *at least one* Special Dark (application of the Addition Rule), and input their response. Feedback related to this particular problem is shown in Table 1.

Learning environment

While the instructional examples and problem-solving scenarios were designed to be as creative as possible (and consequently, more memorable), Stat Lady's learning environment/interface was designed to be experiential in order to get subjects actively involved in the learning process. The tutor begins by engaging the learner in a little 'betting game' where learners are provided with $5.00 electronic cash for start-up funds, and bets are rendered on different combinations of numbers. A particular game is defined (e.g., Stat Lady wins if she gets a 9 or 10 on a roll of two dice, the learner wins if an 11 or 12 appears, otherwise, it's a draw), then the learner makes a bet (from 'no bet' up to $5.00) and left-buttons on the option 'roll 'em' which causes Stat Lady to shake and roll two dice. Over time (and usually, loss of cash), learners will realize that most of the games are unfair. To prove this (later on in the curriculum), they construct a table listing all two-dice events (2, 3, ... 12), all possible *outcomes* corresponding to each event (e.g., {1,1} {1,2 or 2,1} ... {6,6}), and all associated probabilities. Learners are then able to precisely assess the 'fairness' of the games (e.g., figuring out the probability of obtaining a 9 or 10 = 7/36 vs. an 11 or 12 = 3/36).

While Stat Lady was designed to be an experiential learning environment, it's still just a computer program, and thus inferior in many imporant ways to human instruction. Some general advantages of humans (over programs)

are that they are able to respond to questions about the material, adapt the presentation of material to the interests and abilities of the students, and condense or expand the presentation of material to match the time allotted for its presentation. We'll now describe the second treatment condition – the classroom lecture.

Lecture condition

A professor of psychology was employed as the lecturer in this condition. He has taught the introductory statistics course for more than 20 years, and is well-liked by the students. Using the course outline of curriculum elements, the lecturer was able to cover the entire curriculum in the designated three-hour time frame, but did not provide equal coverage to all of the curriculum topics. This stood in contrast to the Stat Lady condition where, because the program was self-paced, about half of the subjects failed to complete the whole curriculum.

The lecture sessions met for one hour on three successive days, comparable to the schedule for the Stat Lady group. The first lecture focused on basic aspects of probability (e.g., the concepts of sample space, events, mutually exclusive events, independent events, and the meaning of the term probability). The second lecture was devoted to probability rules for the combination of events (e.g., the 'AND/multiply' and the 'OR/add' rules). The counting rules were covered in the third lecture. The material covered in the prior lecture was reviewed at the beginning of the second and third lectures (which was not the case in Stat Lady).

The lecturer attempted to build upon the intuitions of the subjects. For example, the multiplication rule for independent events was introduced by posing the question, 'What is the probability of getting two heads when an honest coin is flipped two times?' The answer was then explained by the multiplication rule and the rule then used to determine the probabilities for more difficult problems. The lecturer also attempted to maintain the interest and involvement of the subjects, although this was not easy because the subjects knew that they were participating in an experiment and that their retention of the material would not affect their grade in a course. Problems were posed for the subjects to answer, and questions were encouraged. The lecturer also attempted to determine whether all of the subjects were understanding the material. Repetition and additional problems were used to ensure that they were. Because the lectures were typically aimed at the middle (and low) aptitude subjects, the lecture format may not have been optimal for the higher-ability subjects who would have benefited from a more rapid pace of learning.

Tests

Two parallel tests were constructed to assess statistical knowledge and skills across all three conditions – Forms A and B. Each item in Form A had a parallel item in Form B, and both forms assessed all curricular elements. For example, subjects had to solve a problem on both tests related to Counting Rule 2 ($K_1 \times K_2 \times \ldots \times K_N$). On Form A, this item was: *You have 3 pairs of pants, 4 shirts, and 2 pairs of shoes. How many different outfits are possible (note: an 'outfit' consists of pants, shirt, and shoes)*? On Form B, the item was: *At a salad bar, you can choose from 3 different macaroni salads, 2 soups, and 3 types of bread. How many different combinations of macaroni salad, soup, and bread are possible (note: a combination consists of one macaroni salad, one soup, and one bread)*?

Each test consisted of 47 items covering the entire curriculum, with some items having sub-parts. While informal pre-testing was done to assure that the items were parallel, no measure of reliability was taken until the experiment itself. The tests were counter-balanced in the actual experiment, with half the subjects in each group receiving the tests in A-B order and the other half receiving the tests in reverse B-A order.

Design and procedure

Two experimental groups and one no-treatment control group participated in a pretest-posttest control group design with intervening instruction for the two experimental conditions. The design was a $3\times2\times2$ factorial with two between-subjects factors (condition = Stat Lady, Lecture, and Control; and gender = male and female) and one within-subjects variable (test score = pretest and posttest). Thus, we compared two types of learning environments: Stat Lady and a typical classroom lecture, assessing the degree to which students learning from Stat Lady succeeded in learning probability compared to a group receiving standard lectures covering the identical curriculum, and a control group receiving no training (to measure baseline test-retest changes).

To ensure that both treatment groups were using the same curriculum (concepts, rules, symbols, and terminologies), we developed a comprehensive outline of curricular elements and provided it to the professor teaching the Lecture group who agreed to base his lectures on this outline. In addition, each of these elements formed the basis for item development in the two tests, as mentioned above.

Subjects signed up for one of the three conditions, and all subjects were administered a one-hour pretest (paper and pencil format) on a Monday. Subjects in the two treatment conditions then received three hours of instruction

Table 2. Pretest and posttest (percent correct) scores by condition.

Condition	Pretest	Posttest
Stat Lady(N = 63)	44.2	57.9
	(14.8)	(14.5)
Lecture(N = 36)	49.1	62.0
	(14.3)	(13.2)
Control (N = 69)	47.9	52.4
	(12.8)	(12.0)

Means are presented in columns with standard deviations in parentheses.

(on Tuesday, Wednesday, and Thursday, in one-hour units of instruction) from either a computer or a lecture. In all three conditions, subjects were tested in groups of about 7 persons, either at desks in a classroom, or in front of computers. For both treatment conditions, subjects were allowed to take notes and review these notes at home, if desired. Following instruction, a one-hour posttest (again, paper and pencil format) was administered on the ensuing Friday to all subjects (note: individuals in the no-treatment group simply returned after four days for their posttest).

To ensure objective scoring of the tests, we developed a scoring scheme providing all correct answers, and delineating situations where subjects could receive partial credit for an answer. For example, if a subject wrote the correct equation but had an incorrect final answer, half credit was given for that item because he or she showed knowledge of how to solve the item, but made a mathematical error. Two individuals were hired from a temporary employment agency to score all of the tests. These persons were unaware of different treatment conditions, they thoroughly double-checked each other's scoring, and there was very high agreement between the two scorers[2] Finally, aptitude data (e.g., Verbal and Quantitative SAT scores) as well as demographic data were also collected from the subjects.

In addition to not receiving instruction, the Control group was expected to differ in other ways from the two experimental groups due to differential testing requirements (i.e., needing them for only two, rather than five, hours). Instead, apparently through sampling error, the two *experimental* groups differed on the pretest more from each other than they did from the control

[2] There was perfect agreement on 160/168 > 95% of the tests scored. On the other 8 tests, item-score discrepancies were resolved on a case-by-case basis by the first author (blind as to the subjects' experimental condition).

Table 3. Pretest and posttest (percent correct) scores across parallel forms.

Form	Pretest	Posttest
Form A (N = 86)	48.6	56.7
	(13.2)	(13.5)
Form B (N = 82)	45.4	57.1
	(14.7)	(14.1)

Means are presented in columns with standard deviations in parentheses.

group (see Table 2 results), where subjects in the Stat Lady group scored lower on the pretest compared to the other conditions (albeit, this difference was not significant).

Results

Before reporting the results examining differential learning by condition, we first needed to determine if the tests we used were reliable measures. A split-half reliability coefficient was computed from the overall posttest scores (using the Spearman-Brown prophecy formula), equal to 0.71.

Next, recall that we had two matched forms for each test (A and B). To control for possible test bias, half of the subjects in each group were given the tests in A-B order while the other half were given the tests in reverse B-A order. For all 168 subjects, the overall percent correct on Form A = 52.8 (SD = 14.2) and the percent correct on Form B = 51.2 (SD = 15.2). A t-test comparing these means was not significant. In addition, when examined at the item level, there were no differences between any of the 47 matched pairs of A-B items within the tests. Finally, when the test-form data were compared as pretest or posttest, there were no differences between any of these values, shown in Table 3. Therefore, it appears that Forms A and B were indeed parallel. Because there was no main effect due to test form (A and B), all ensuing data analyses collapse across form.

The next question addressed the equivalence of incoming abilities among the three groups of subjects. An ANOVA was computed on the pretest data, by condition, and found to be not significant: F (2, 165) = 2.19; $p > 0.10$. Thus, the groups all started out with approximately the same degree of incoming, domain-specific knowledge and skills.

We additionally wanted to examine subjects' quantitative SAT scores by condition because the curriculum was quantitative in nature. As with the

pretest data, above, the results of this analysis showed no significant differences among conditions on quantitative aptitude ($F < 1$). However, the mean scores per condition were of interest, especially as there was such a discrepancy between the Stat Lady and the other two conditions: Stat Lady M = 574 (SD = 81), Lecture M = 592 (SD = 94), and Control M = 598 (SD = 123). That is, students in the Stat Lady condition scored almost 20 points less on the quantitative SAT measure compared to the Lecture group, and about 25 points less than the Control group.

A repeated measures MANOVA was computed on the data where the dependent measures were pretest and posttest scores, and the two between-subjects factors were condition and gender. The first question addressed whether there was any evidence of overall change in subjects' test scores from pretest to posttest. There was a significant main effect due to the within-subjects variable (pretest vs. posttest): F (1, 164) = 88.46; $p < 0.001$. In other words, all three conditions showed, overall, higher posttest mean scores (percent correct) compared to pretest means. Pretest M = 47 (SD = 14, N = 168); Posttest M = 57 (SD = 14; N = 168).

For the next set of analyses, orthogonal comparisons were computed to determine two specific contrasts in relation to learning outcome: (1) Two treatment groups (Stat Lady and Lecture) vs. Control, and (2) Stat Lady vs. Lecture. The specific question asked: Was there differential change (i.e., learning) from pretest to posttest score by condition? The two experimental groups, combined, did learn a lot more of the material (i.e., pretest to posttest gains) relative to the control (baseline) condition. That is, the experimental vs. control group $\times$ pretest/posttest interaction was signficant with: $F(1, 164) = 21.23; p < 0.001$. The experimental group's average percent correct gain from pretest to posttest was 14% (SD = 13; N = 99) compared to the control group's increase of only 4% (SD = 10, N = 69). Furthermore, the contrasted groups began the study at about the same level of knowledge and skill: Experimental Pretest M = 46.2 (SD = 14.8), Control Pretest M = 48.2 (SD = 12.8).

The next orthogonal contrast asked: Did the two treatment conditions differ between themselves in terms of relative improvement? There was no significant experimental groups $\times$ pretest/posttest interaction. That is, there was no difference between Stat Lady and the Lecture conditions on pretest to posttest changes: F (1, 95) = 0.07. Thus, both groups learned equally well from the respective instructional interventions, and both groups improved about one standard deviation above their pretest scores, with only three hours of instruction.

We also investigated the data in terms of any differences in performance as a function of gender. Overall, there was a main effect due to gender: $F(1, 164) = 9.20; p < 0.01$, where males showed greater test scores compared to females.

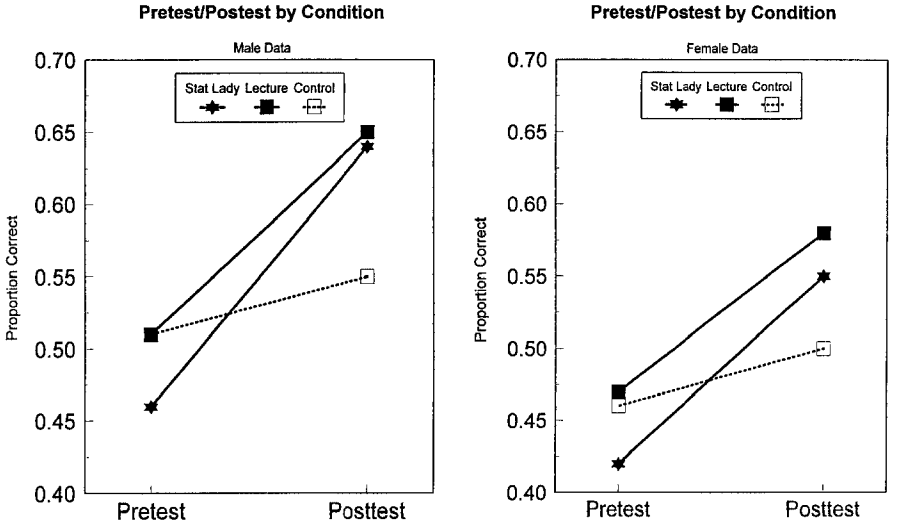

Figure 2. Changes in pretest to posttest scores as a function of learning condition and gender.

Pretest Scores: Male M = 49.1 (SD = 13.7, N = 94), Female M = 44.4 (SD = 14.1, N = 74). Posttest Scores: Male M = 60.5 (SD = 13.1, N = 94), Female M = 52.4 (SD = 13.4, N = 74). These data, plotted separately for males and females, are shown in Figure 2. It is interesting to note that, although subjects in the Stat Lady group scored about 5% less on the pretest than either the Lecture or Control conditions, the males learning from Stat Lady managed to increase to the same level as the Lecture group by the posttest. Next, none of the interactions involving gender approached significance, with the exception of gender × pretest/posttest. Overall, males performed marginally better than females in terms of pretest to posttest gains: F (1, 164) = 3.21; $p = 0.075$.

Given the obtained main effect of gender on outcome performance, we sought a *post hoc* explanation for these findings. First, gender-related performance differences were not attributable to differential number of previous math courses. Data concerning each subject's number of algebra, geometry, and statistics courses were collected. None of these courses differed significantly by gender, and when the number of math courses was included as a covariate in the equation, the main effect of gender remained: F (1, 161) = 9.12; $p < 0.01$. On the other hand, there was a significant difference between males and females, overall, on their Quantitative SAT scores:[3] Male M = 608

[3] Submitting SAT scores was voluntary, and in this study, 146 out of 168 persons elected to report them. Analyses involving SAT score (either cumulative or separated into Quantitative and Verbal scores), thus show a slightly lower df than other analyses.

(SD = 108, N = 88), Female M = 561 (SD = 94, N = 60), F (1, 145) = 7.37, $p < 0.01$ (but no difference between males and females on Verbal SAT scores: $F < 1$). Thus, performance differences between the sexes may be attributed to a differential quantitative aptitude. This was supported by a reanalysis of the MANOVA (reported above) including quantitative SAT as a covariate. In this reanalysis (controlling for incoming quantitative skill), the main effect of gender disappeared: F (1, 145) = 2.26; $p = 0.14$.

The final research question we investigated concerned evidence for aptitude-treatment interactions. Originally, we hypothesized that high-aptitude subjects would learn better from Stat Lady, and low-aptitude subjects would perform better within the structured and familiar Lecture condition. First, there were no significant differences among the three conditions on their cumulative SAT scores: F (2, 145) = 0.10, Stat Lady M = 1100 (SD = 129), Lecture M = 1114 (SD = 139), and Control M = 1108 (SD = 154). Next, we computed a stepwise regression analysis predicting posttest score from the following independent variables: pretest score, cumulative SAT score (our measure of general aptitude), treatment condition (i.e. Stat Lady , Lecture, or Control), and an explicit interaction variable: treatment $\times$ SAT score. Results showed a Multiple $R = 0.72$; F (4, 143) = 37.49; $p < 0.001$. Hence, 50% of the outcome variance may be accounted for by the following variables: Pretest score ($t = 7.30$, $p < 0.001$), SAT score ($t = 4.14$, $p < 0.001$), and treatment condition ($t = 2.23$; $p < 0.05$). Furthermore, the interaction between treatment $\times$ aptitude was significant ($t = -2.80$; $p < 0.01$).

Using the regression equation to plot expected values for our outcome variable, we plotted subjects' data who were one standard deviation above and below the mean to best illustrate the obtained interaction. This is shown in Figure 3. Here, we see that for low aptitude individuals, there was no difference in predicted outcome score by condition, but for high aptitude subjects, there was a significant difference in outcome performance by condition – with a Stat Lady advantage over the other two conditions. Clearly, the high aptitude subjects benefited more from the experiential environment compared to subjects learning from the more didactic lecture condition.

Discussion

Stat Lady was designed to be an experiential learning environment (albeit, non-intelligent) to render the learning of statistics more memorable and meaningful. We sought to achieve this goal by embedding the curriculum elements in real-world (and colorful) examples, drawing on prior knowledge on which to seat new knowledge, and making the learning experience enjoyable by presenting the material in a game-like environment. The hands-on nature of

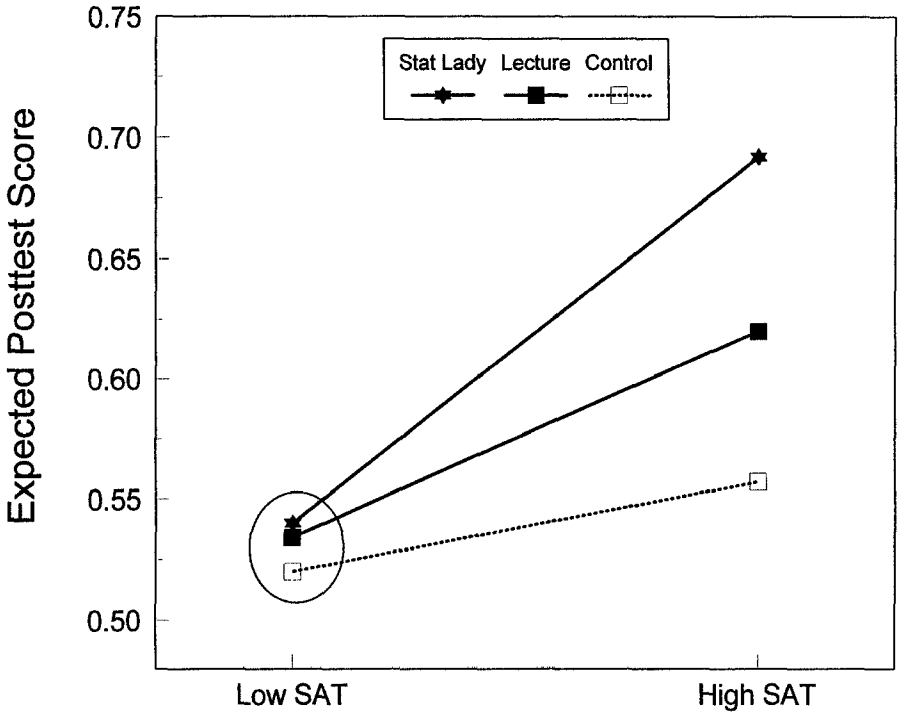

Figure 3. Aptitude-treatment interaction

Stat Lady was believed to make abstract probability concepts concrete and 'alive.'

The present experiment compared the instruction of introductory probability from two diverse methods: Stat Lady and a traditional lecture method. The main research questions we addressed included: Was there differential change (i.e., learning) from pretest to posttest score by condition, and if so, did the two treatment conditions differ between themselves in terms of relative improvement? We were also interested in examining the data for a main effect or interactions involving gender on pretest-posttest scores. Finally, we investigated the data for the presence of an aptitude-treatment interaction.

In the current experiment, we included a no-treatment control group to test the degree to which subjects increased from pre- to posttest in the absence of any instructional intervention. This provided a baseline measure for our subsequent analyses on learning. We found that both the Stat Lady and Lecture methods resulted in greater pretest-to-posttest performance increases than did a control with no instruction (indicating the relative degree of learning), yet the two treatment groups did not differ from one another. We had originally postulated that the experiential nature of Stat Lady would improve learning,

perhaps comparable to that from a classroom lecture. There are several reasons why we view this 'null finding' as heartening. That is, the two treatment conditions differed along several important dimensions, often with a Stat Lady disadvantage.

Differential experience

Based on the obtained results, it was concluded that Stat Lady (non-intelligent, 5 months old, and on her first teaching assignment) was no less an effective instructional tool than was an intelligent, skilled, and popular lecturer with more than 20 years of experience teaching the same concepts. In general, a human instructor can easily outperform an unintelligent computer program in terms of being able to adapt to individual strengths and weaknesses, as well as gauge time. Specifically, the instructor in this experiment was able to conduct instructive dialogs at several levels, appealing to both slower and brighter students during the same lecture. In addition, the instructor was able to guide the timing of the lecture, spending relatively more time on important concepts and less time on those that are either less important or known from experience to be easier for the students to grasp. As a result, all subjects in the lecture condition were taught the entire curriculum within the allotted time frame. In contrast, learning was designed to be self-paced in Stat Lady. Therefore, given the time constraints imposed by this study (i.e., 3 hr), many subjects learning from the Stat Lady condition failed to complete the entire curriculum. But even with incomplete coverage of relevant material, these 'handicapped' subjects still scored comparable to the Lecture condition on their posttest scores.

Instructional familiarity

In addition to a large discrepancy in teaching experience, another important difference between these two methods of instruction was familiarity. In general, students are quite familiar with the traditional method of classroom instruction: Each student in the experiment had at least 12 years experience with learning by lecture, and hence, was very accustomed to, and comfortable with this method of instruction. In contrast, many (perhaps most) students were receiving computer-delivered instruction for the first time. Some individuals had never interacted with a computer before, and others had only minimal computer experience. For those individuals in the Lecture group, the important information provided by the instructor was easy to identify (i.e., his clothes, mannerisms, and habits were immediately identified as irrelevant to the learning task). On the other hand, the Stat Lady subjects did not know at the outset of the experimental session just what was important and what

was not important. As a consequence, much of the effort expended by the Stat Lady students (especially in their first computer session) was spent simply trying to distinguish important from unimportant information. In other words, although the interface to the system was quite user-friendly, some computer-related information and procedures had to first be acquired by the learners (e.g., how to insert a floppy, type on the keyboard, use a mouse and make menu-item selections) before they could move on to actually acquiring the domain-specific knowledge and skills. Despite this disadvantage, subjects in the Stat Lady group not only learned how to run a computer and survive in a novel learning environment, they also learned the probability concepts presented by the program to the same degree as the Lecture group. At the same time, the Lecture group enjoyed the familiarity of a mode of learning which they had been exposed to for at least 12 years.

Element of surprise

A third variable differentially impacting the two treatment conditions involved the administration of the outcome measure. The posttest was intended to be a 'surprise' for all subjects in the treatment conditions. However, it turned out that many subjects in the Lecture group were informed that there would be a posttest, so some of them actually studied for it. In contrast, no one in the Stat Lady group knew about the posttest. Thus, if there was any posttest advantage, it was clearly on the side of the Lecture group.

Motivation

The final factor which entered into the learning equation (perhaps equally per condition) involved subject motivation. Subjects in this study were introductory psychology students who, by serving in this experiment, were fulfilling a course requirement. As such, it is likely that these subjects were both relatively uninterested and/or unmotivated as subjects in the research. Furthermore, comparisons should not be made to some standard of performance for motivated students taking a statistics course.

We also examined the role of gender in learning this curriculum. We found a main effect of gender on outcome performance (with a male superiority), but that wasn't surprising as males entered the study with higher quantitative aptitude scores. When this measure was controlled (i.e., covaried out of the equation), the gender effect vanished.

The final analysis concerned the interaction between aptitude and learning condition (or treatment). Our hypothesis was that high aptitude subjects would benefit more from the Stat Lady condition while low aptitude subjects would profit from the traditional Lecture condition. As predicted, high apti-

tude subjects (i.e., those individuals one standard deviation above the SAT mean) did learn significantly more from Stat Lady than from the Lecture condition. But for low aptitude subjects, there was very little difference among the three conditions. Why would such results be obtained? One explanation is that low SAT students had difficulty adjusting to a new learning environment. As a consequence, they may have found it difficult to identify, and attend to, those components that were differentially important for successful acquisition. Thus, the potential benefit of Stat Lady would be attenuated for this group that struggled to simply endure within the novel environment. In contrast, the more intelligent (i.e., high SAT) students may have adjusted more rapidly to the new learning situation and found it to be an interesting challenge. As a consequence, they may have been able to attend to those components of the new task which were important and tested at a later time. Furthermore, the lecture was (as most usually are) pitched to the average student – one who learned more slowly than the high SAT students. Thus, the high SAT students may have found that being required to learn something from a lecture (albeit presented in an interesting manner) was tedious. As a consequence, the brighter students performed relatively worse if they were required to learn from the Lecture condition than from Stat Lady.

In general, the Lecture condition provided a familiar environment for all students, both high and low aptitude. Given such a familiar environment, the low SAT students probably found the lecture an easily understood method of learning the presented material. What is surprising about our findings is that the low aptitude students did not show a greater Lecture condition advantage. We have recently completed another study with Stat Lady (compared to a Workbook treatment), where we collected a wide range of cognitive measures from subjects (Shute & Gawlick-Grendell, 1994). In that study, cognitive ability was assessed by performance on a battery of computer-administered tests (measuring working-memory capacity, information processing speed, inductive reasoning skill, general knowledge, and associative learning skill). An aptitude factor score was computed from the cognitive measures, and we found the same ATI pattern as reported herein; namely, that high-aptitude subjects learned significantly better from the Stat Lady, than the Workbook condition, and for low-aptitude subjects, there was no difference in their relative learning gains by condition.

In conclusion, many factors in this study strongly favored the Lecture condition: (1) Lectures constitute a very familiar learning environment for all subjects, while considerably fewer subjects are familiar with computer-administered instruction; (2) The professor delivering the lecture was both popular and had over 20 years experience teaching the subject matter, compared to Stat Lady who was on her 'first teaching assignment;' (3) Subjects in

the Lecture group were informed in advance about the posttest, giving them an opportunity to study for it (which many reportedly did), while the other treatment group did not know about it, (4) Subjects in the Lecture condition had higher quantitative SAT scores compared to the Stat Lady condition (albeit, not statistically significant) providing them with an incoming math advantage, and (5) While subjects in the Lecture condition were exposed to all curricular elements during the three one-hour classes, not all subjects in the Stat Lady condition were able to complete each of the one-hour curriculum units in the designated time. In other words, slower Stat Lady subjects were not exposed to all of the material, while slower subjects in the Lecture condition were exposed to all material. So, despite these aforementioned obstacles, we were very encouraged to see the Stat Lady group perform as well as the Lecture group. Moreover, subjects learning from Stat Lady provided us with unsolicited comments of enjoyment, while there were no comments offered by subjects in the Lecture condition. These can be seen in Appendix II.

Future Research. While the obtained ATI reported within this paper was replicated in a different study, the findings in regard to low-ability students remain troubling; they represent precisely the population that need help the most. Whereas Stat Lady instruction results in greater performance for high aptitude subjects, perhaps what is needed to boost performance for low-aptitude subjects is more explicit real-time cognitive diagnosis and remediation. This issue is currently being examined with the inclusion of intelligence in a new version of Stat Lady teaching descriptive statistics (Shute & Gluck, 1994). In this tutor, we've embedded a novel approach to student modeling called SMART (Student Modeling Approach for Responsive Tutoring) (Shute, 1994). Information about curriculum elements, derived from a cognitive task analysis, arranged in an inheritance hierarchy, and computed in relation to a series of regression equations, provides the basis for inferences about what knowledge and skills have been acquired, and to what degree. It is our hope that the new version of Stat Lady (using robust cognitive diagnosis, mastery learning criterion, and effective remediation) will promote learning, especially for the lower-ability students.

Appendix I: Global curriculum elements

Section 1 (15 elements)

sample space	elementary event
probability formula	event class
estimating probability	sampling with replacement
probability notation	exhaustive rule
sure events	impossible events

range of probability	addition rule
mutually exclusive events	not mutually exclusive events
complementary rule	

Section 2 (4 elements)

multiplication rule	sampling without replacement
conditional prob., related events	conditional prob., unrelated events

Section 3 (6 elements)

number of sequences, N trials	for sequences
permutations	ordered combinations
combinations	binomial expansion

Appendix II: Summary of affective comments about Stat Lady

Comments

The program was excellent!

The computerized method was very useful because it used realistic situations.

This tutor improved my understanding of probabilities and statistics even though I've never been any good at math and formulas.

I enjoyed the computer a lot, but may have learned even more from an instructor.

This program was educational and fun. I even think I learned something!

In general, students enjoyed learning from Stat Lady and found the computerized method of instruction useful in terms of its realistic situations and feedback features. Similar results were found in a study comparing learning from Stat Lady with learning from a Workbook version of Stat Lady (Shute & Gawlick-Grendell, 1994). Students learning from the computerized version enjoyed their experience significantly more than students learning from a paper and pencil version of the same curriculum and presentation format (Shute & Gawlick-Grendell, 1994). However, in the current study, some students admitted to not being very good in mathematics at the outset, and felt they would have performed better had they been learning from an instructor (or at least had someone available to answer questions about the curriculum). Complaints about Stat Lady concerned not having enough practice in working out formulas, which made it difficult for subjects to remember them at test time. An updated version of the Stat Lady Probability Module has incorporated a practice option, where students have the choice of working out more problems on a given topic if they desire.

References

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology.* Cambridge: Cambridge University Press.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence* 47: 139–159.

Brown, J. S., Collins, A. & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher* 18(1): 32–42.

Clancey, W. J. (1992). Representation of knowing: In defense of cognitive apprenticeship. *Journal of Artificial Intelligence in Education* 3: 139–168.

Collins, A. Brown, J. S. & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics, in L. B. Resnick, ed., *Cognition and Instruction: Issues and Agendas.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Friedman, P. G. & Yarbrough, E. A. (1985). *Training Strategies from Start to Finish.* Englewood Cliffs, NJ: Prentice-Hall.

Hyde, J. S., Fennema, E. & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin* 107(2): 139–155.

Lave J. & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation.* Cambridge: Cambridge University Press.

Piaget, J. (1954). *The Construction of Reality in the Child.* New York: Ballentine Books.

Shute, V. J. (under review). SMART: Student Modeling Approach for Responsive Tutoring. To appear in a special issue of: *User Modeling and User-Adapted Interaction: An International Journal.*

Shute, V. J. & Gawlick-Grendell, L. A. (1992). *Stat Lady : Probability Module* [Unpublished computer program]. Brooks Air Force Base, TX: Armstrong Laboratory.

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

Shute, V. J. & Glaser, R. (1991). An intelligent tutoring system for exploring principles of economics, in R. E. Snow & D. Wiley, eds., *Improving Inquiry in Social Science: A Volume in Honor of Lee J. Cronbach* (pp. 333–336). Hillsdale, NJ: Lawrence Erlbaum Associates

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

Shute, V. J. & Psotka, J. (in press). Intelligent tutoring systems: Past, present, and future. To appear in D. Jonassen, ed., *Handbook of Research on Educational Communications and Technology*, Scholastic.

Suchman, L. A. (1987). *Plans and Situated Actions.* Cambridge, MA: Cambridge University Press.