Can Automated Emotional Scaffolding Affect Student Persistence? A Baseline Experiment

Jack Mostow mostow@cs.cmu.edu
Joseph E. Beck joseph.beck@cs.cmu.edu
Joe Valeri jmv@andrew.cmu.edu
Project LISTEN, Carnegie Mellon University, RI-NSH 4213, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

Abstract

A 2002 Wizard of Oz study showed that emotional scaffolding provided by a human significantly increased children's persistence in an automated Reading Tutor, as measured by the number of tasks they chose to undertake. We report a 5,965-trial experiment to test a simple automated form of such scaffolding, compared to a control condition without it. 348 children in grades K-4 spent significantly longer per task in the experimental condition due to a design flaw, yet still averaged equal numbers of tasks in both conditions. We theorize that they subjectively gauged effort in terms of number of tasks rather than number or duration of solution attempts.

1. Introduction

Motivational concerns affect student learning (Schunk, 2003), inform expert human tutoring (Lepper et al., 1993), and might help automated tutors too (Soldato & Boulay, 1995), given sufficient ability to model and manipulate students' motivation. Motivation itself is not directly observable, and is problematic to measure using students' self-reports. Inferring motivation from student behavior depends on whatever theory the inference process assumes. To avoid these various pitfalls, we prefer to directly measure the student behavior we seek to motivate. In particular, we measure student persistence on tasks to evaluate tutors' motivational effects on students. Persistence is sensible to measure because it is both externally observable and essential to learning.

1.1. Pilot study

A Wizard of Oz study (Aist *et al.*, 2002a, b) tested the hypothesis that emotional scaffolding could improve student persistence in an automated tutor. 14 children in grades 2-5 (which correspond to ages 7-11) completed the experiment. The tutor was Project LISTEN's Reading Tutor, which listens to children read aloud (Mostow & Aist, 1999, 2001), and responds with spoken and graphical assistance. The tasks were word lists for students in grades 2-3 and limericks for students in grades 4-5. After each task the student chose whether to quit or do another.

Each student used the Reading Tutor under two different conditions, with a week between the two sessions. In the experimental condition, a human wizard added spoken emotional scaffolding over the same audio channel used by the Reading Tutor. The wizard could see and hear the child, but the child could not see the wizard. In the control condition, the wizard observed the interaction but did not intervene.

Scaffolding significantly boosted persistence: children did more tasks with scaffolding than without (8.6±4.8 versus 6.4±5.1, p=.04). This experiment confirmed that human emotional scaffolding helped. The difference was much larger for boys than for girls, but the sample size was too small to be sure.

1.2. Purpose of current experiment

To test whether automated emotional scaffolding can also improve students' persistence, we embedded a randomized experiment in the 2002-2003 version of the Reading Tutor. We designed and user-tested this experiment in late summer 2002, keeping it simple to have it ready in time to include in a larger controlled study of the Reading Tutor's overall effectiveness.

The automated experiment let us scale up sample size *N* and duration *T*. The 2002-2003 Reading Tutor was used by hundreds of children in nine elementary schools for several months. The larger *N* gave us the statistical power to test for gender and other effects. The longer *T* let us test if scaffolding continued to affect persistence once its novelty wore off.

In the Wizard of Oz study, the human wizard could see and hear the student, and used his expertise in tutoring to provide effective emotional scaffolding. In our experiment, we wanted to see whether a much simpler form of emotional scaffolding might work. It made sense to try something simple before deciding whether to invest in complex sensing and modeling. Thus the resulting experiment serves as a baseline.

To avoid speech recognition errors in classifying students' answers as correct or incorrect, we used type-in spelling tasks instead of oral reading tasks. These tasks also served the educational goal of providing spelling practice and the research goal of collecting data on children's spelling attempts.

2. Experimental design

The experiment worked as follows. The Reading Tutor took turns picking with the student picking which story to read next (Aist & Mostow, 2003). Occasionally, just before a story, the Reading Tutor would insert an experimental trial, randomly assigned to one of two conditions. The experimental condition included a simple form of spoken emotional

scaffolding throughout, such as congratulations after correct attempts and sympathy after incorrect attempts, using recorded expressive human speech. The control condition included no such scaffolding.

Either way the Reading Tutor picked a word at random from the story and prompted the student to spell it by typing it in, giving unlimited opportunities to repair incorrect attempts. After each word, it asked if the student wanted to try another. If so, the process repeated, up to a maximum of five words. The outcomes for each trial were the number of tasks (spelling words) the student chose to do, and the number and duration of attempts at each task.

2.1. Experimental condition

To illustrate, we use *italics* for what the Reading Tutor said aloud, **bold** for the text it displayed, and ***bold italics*** for text it both displayed and read aloud:

Before you read this story, let's practice spelling.

Try your best, and I'll help if you can't get it.

The fish goes splish splash! *Please type FISH.*

(Note that *fish* was spoken but not displayed.)

The student typed in the word (here, as `fissssh`).

If it was correct, the Tutor congratulated the student:

Alright! You got it!

If a first try was incorrect, it encouraged another try:

Not quite, so click back and try again.

I know you can do it!

The student had to try again (and typed `fish`).

If a subsequent try was incorrect, it sympathized:

No... What a tough question!

F I SH spells FISH.

The student could go on or click Back to try again (on a screen where the spelling word was not shown).

Next the Reading Tutor asked whether to persist:

Do you want to try more words?

Yes

No

The student clicked on ***Yes*** or ***No***, presented in randomized order to avoid bias towards either choice.

If ***No***, the Reading Tutor gave final congratulations:

Wow!

You did a great job on that word!

If Yes, the Reading Tutor presented the next word:

And the dog sat up and roared? *Please type ROARED.*

The student typed in an answer (`roard`).

If it was correct, the Reading Tutor gave praise:

Excellent! Way to go!

If a first attempt was wrong, it gave encouragement:

Not that one, click Back to try again.

You can get it!

The student had to try again (and typed `roared`).

If a subsequent try was incorrect, it gave consolation:

Nope... You gave it a good try!

R O A R E D spells ROARED.

This process continued up to at most five words per trial. As the example illustrates, the structure was the same for each word but changed phrasing for variety.

2.2. Model of affective state and scaffolding

The Reading Tutor's response distinguished only three "affective states," corresponding to presumed effects on perceived self-efficacy (Schunk, 2003) depending on whether the student had just spelled the word correctly, misspelled it on a first attempt, or misspelled it twice in a row. In the first state, congratulating the student should provide positive reinforcement. In the second state, encouragement to try again might bolster student confidence. In the third state, sympathetically phrased corrective feedback might reduce student frustration.

This model of the student's affective state, the tutor's corresponding responses, and their phrasing was loosely based on advice from Barry Kort, who served as the wizard in the pilot study. Although we used many of his phrases, time was short so we adopted a much simpler model for when to use them.

2.3. Control condition

The control condition gave choice but no scaffolding:

Before you read this story, let's practice spelling.

Scientists say that they may be extinct soon.

Please type EXTINCT.

The student tried to spell the word (as `anstankt`).

After a first attempt, whether correct or incorrect:

Click Back if you want to try again or Go to move on.

The student chose to try again (typing `aanstant`).

After a subsequent attempt, the Reading Tutor said:

Let's move on now.

Next the Reading Tutor asked whether to persist:

Do you want to try more words?

Yes

No

The process continued up to at most 5 words per trial.

2.4. Flaws in experimental design

In analyzing our data we noticed two flaws in this implementation of the experimental design. The control condition gave no feedback on correctness, leaving the student little reason to try again. Thus the experiment conflated emotional scaffolding with the information value of performance feedback.

Moreover, if a student's first attempt to spell a word was incorrect, the experimental condition required a second attempt, but the control condition did not. Thus it is hardly surprising if students made more attempts per word in the control condition.

A modified design could eliminate both confounds by changing the control condition to report whether an attempt is correct, and by relaxing the scaffolded condition to let the student choose whether

to correct mistakes. However, even without fixing these flaws the experiment yielded interesting results.

3. Evaluation

3.1. Data set

A database server at each school sends back each day's Reading Tutor transactions that night via Internet to update an aggregated database in our lab.

The data set for this paper is for 348 students (162 girls and 186 boys) using 167 Reading Tutors at eight schools. The 2,962 experimental and 3,003 control trials ranged from January 2 to April 23, 2003, with mean 8.6 and maximum 36 per student per condition.

3.2. Analysis

We formulate research questions in terms of database queries (Mostow *et al.*, 2002a, b) and analyze their results using SPSS (SPSS, 2000).

For this paper, the basic research question is whether the intervention affects student persistence. But how should we measure persistence? Mean trial duration measures overall persistence as time on task.

However, it is more informative to factor time into (tasks/trial) × (attempts/task) × (time/attempt). The number of tasks (spelling words) per trial measures the student's willingness to undertake more tasks. The number of attempts per task measures willingness to keep trying to do the task. Finally, the time per attempt may reflect the effort invested in it.

We averaged each of these components on a per-student basis, so as not to skew results toward faster readers or more prolific users who had more trials.

For each component, we defined the scaffolding effect for each student as the mean value in the student's experimental trials minus the mean value in that student's control trials. Comparing the scaffolding effect against zero is equivalent to comparing the two conditions, paired by student. The *N* for such comparisons is the 348 students rather than the 5,965 trials, which were not independent.

3.3. Results

Students averaged 2.12 tasks per trial with scaffolding versus 2.16 without. This difference was not significant ($p > .50$). The number of tasks (words spelled) per trial showed identical bimodal distributions of values, mostly 1 or 5. Over 50% of the trials had 1 task. About 20% of the trials had 5 tasks. 15% had 2, 7% had 3, and 4% had 4.

The number of attempts per task was the only persistence measure affected by treatment, averaging 2.22 with scaffolding versus 1.48 without. This difference was significant at $p < 4.2E\text{-}42$ (2-tailed), but may be due to the confounds discussed earlier.

Response time averaged 35.67 seconds with scaffolding versus 34.21 seconds without, but this difference was not significant ($p > .24$).

We used MANOVA to analyze how each component (and its treatment effect) depended on other variables likely to predict outcomes: gender, grade, and the student's total number of trials.

Different factors affected different components. The number of tasks per trial differed by gender ($p = .031$), averaging 2.34 for girls versus 1.98 for boys, and by grade ($p = .017$), averaging 2.2 in grades 1, 2, and 3, versus 1.3 in kindergarten and 1.7 in grade 4. The number of attempts per task differed only by treatment, as already discussed. Time per attempt decreased significantly ($p < 9.4E\text{-}16$) with grade (especially from K to 1), and also with number of trials ($R = -0.17$, $p < 2.2E\text{-}5$).

We found no strong interactions with treatment. Boys averaged .063 fewer questions ($p = .033$ for gender * treatment) in the experimental condition than in the control condition, and the scaffolding effect correlated at –0.17 with the student's number of trials ($p = .043$ for trials * treatment). However, these interactions have low significance for such a large sample, and are not corrected for multiple comparisons, so we consider them negligible.

4. Conclusion

We have described a goal, a method, and a result.

The goal is to understand how emotional scaffolding can motivate student persistence. We argued for measuring persistence directly as more reliable than trying to infer hidden "motivation."

The method is to embed automated, randomized within-subject experiments in an interactive tutor, so as to control for individual differences and harness the power of large data samples. In particular, we detailed an experimental design to measure the effects of emotional scaffolding on persistence, and we identified two flaws in the design. We factored time on task into number of tasks attempted, number of attempts per task, and amount of time per attempt.

The result analyzes effects of treatment and other variables on these three components of persistence. Different variables affected different components. Treatment affected only the number of attempts per task. Unfortunately, our experimental design conflated emotional scaffolding with performance feedback and an artificial difference in task structure. Nevertheless, we salvaged an interesting observation: students apparently gauged persistence by the number of tasks undertaken (words to spell) rather than the total number of tries (attempted spellings).

The current results suggest that fixing the flaws in the experimental design would not cause a difference in the number of words attempted, because they were already the same, though it might be interesting to see if the difference in the number of attempts survives the elimination of the confounds.

Why wasn't there a treatment effect on the number of words, as there was in the pilot study? Various hypotheses suggest possible future work.

Did the scaffolding help but its novelty wore off? Relating individual trial outcomes to calendar time could tease apart novelty effects (early versus later trials) from proficiency effects (number of trials). But any novelty effects were swamped: students tried slightly fewer words with scaffolding than without.

Did scaffolding affect reading but not spelling? Perhaps students' preferences for how many words to spell were too strong to be affected by scaffolding. A Wizard of Oz study could test whether human scaffolding can increase persistence on spelling. Conversely, an embedded experiment could use word reading tasks as the pilot study did, but at the cost of reintroducing the complexities that spelling avoided.

Did the context invalidate the experiment? Perhaps any motivational effects of scaffolding were swamped by whatever the Reading Tutor had done just before the trial, or by the students' attitudes toward the reading tasks they expected to follow it.

Did the scaffolding have poor content or triggers? If the pilot study results transfer to spelling tasks embedded in a larger tutorial interaction, then the Reading Tutor failed to say the right things at the right times, or if it did say the right things (that the human wizard did), then did so at the wrong times. This possibility would justify the need for a richer model of student affect. The simple model described here gives a baseline for such a model to improve on.

Acknowledgements

We thank the students and educators who participated, Greg Aist and other LISTENers for suggestions on experiment design, and members of MIT's Affective Learning Companion project. They discussed Barry Kort's emotional scaffolding in the Wizard of Oz study and suggested ways to automate such scaffolding, but are not to blame for the simple version implemented in this baseline experiment.

This work was supported in part by the National Science Foundation under Grant No. REC-9979894. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation or the official policies, either expressed or implied, of the sponsors or of the United States Government.

References (also www.cs.cmu.edu/~listen)

Aist, G., Kort, B., Reilly, R., Mostow, J., & Picard, R. (2002a, June 4). Experimentally Augmenting an Intelligent Tutoring System with Human-Supplied Capabilities: Adding Human-Provided Emotional Scaffolding to an Automated Reading Tutor that Listens. *Proceedings of the ITS 2002 Workshop on Empirical Methods for Tutorial Dialogue Systems*, San Sebastian, Spain, 16-28.

---. (2002b, October 14-16). Experimentally Augmenting an Intelligent Tutoring System with Human-Supplied Capabilities: Adding Human-Provided Emotional Scaffolding to an Automated Reading Tutor that Listens. *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces (ICMI 2002)*, Pittsburgh, PA, 483-490.

Aist, G., & Mostow, J. (2003). Faster, better task choice in a reading tutor that listens. In V. M. Holland & F. N. Fisher (Eds.), *Speech Technology for Language Learning*. Lisse, The Netherlands: Swets & Zeitlinger Publishers.

Lepper, M. R., Woolverton, M., Mumme, D. L., & Gurtner, J. (1993). Motivational Techniques of Expert Human Tutors: Lessons for the Design of Computer-Based Tutors. In S. P. Lajoie & S. J. Derry (Eds.), *Computers as Cognitive Tools* (pp. 75-105). Hillsdale, NJ: Erlbaum.

Mostow, J., & Aist, G. (1999). Giving help and praise in a reading tutor with imperfect listening -- because automated speech recognition means never being able to say you're certain. *CALICO Journal, 16*(3), 407-424.

---. (2001). Evaluating tutors that listen: An overview of Project LISTEN. In K. Forbus & P. Feltovich (Eds.), *Smart Machines in Education* (pp. 169-234). Menlo Park, CA: MIT/AAAI Press.

Mostow, J., Beck, J., Chalasani, R., Cuneo, A., & Jia, P. (2002a, June 4). Viewing and Analyzing Multimodal Human-computer Tutorial Dialogue: A Database Approach. *Proceedings of the ITS 2002 Workshop on Empirical Methods for Tutorial Dialogue Systems*, San Sebastian, Spain, 75-84.

---. (2002b, October 14-16). Viewing and Analyzing Multimodal Human-computer Tutorial Dialogue: A Database Approach. *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces (ICMI 2002)*, Pittsburgh, PA, 129-134.

Schunk, D. H. (2003). Self-efficacy for reading and writing: Influence of modeling, goal setting, and self-evaluation. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 19*(2), 159-172.

Soldato, T. d., & Boulay, B. d. (1995). Implementation of motivational tactics in tutoring systems. *Journal of Artificial Intelligence in Education, 6*(4), 337-378.

SPSS. (2000). SPSS for Windows (Version 10.1.0). Chicago, IL: SPSS Inc.