AIED 2013 Workshops Proceedings
Volume 8

Formative Feedback in Interactive Learning Environments (FFILE)

Workshop Co-Chairs:

Ilya Goldin
Carnegie Mellon University, USA

Taylor Martin
Utah State University, USA

Ryan Baker
Teachers College Columbia University, USA

Vincent Aleven
Carnegie Mellon University, USA

Tiffany Barnes
North Carolina State University, USA

https://sites.google.com/site/ffileworkshop/

Preface

Educators and researchers have long recognized the importance of formative feedback for learning. Formative feedback helps learners understand where they are in a learning process, what the goal is, and how to reach that goal. While experimental and observational research has illuminated many aspects of feedback, modern interactive learning environments provide new tools to understand feedback and its relation to various learning outcomes.

Specifically, as learners use tutoring systems, educational games, simulations, and other interactive learning environments, these systems store extensive data that record the learner's usage traces. The data can be modeled, mined and analyzed to address questions including when is feedback effective, what kinds of feedback are effective, and whether there are individual differences in seeking and using feedback. Such an empirical approach can be valuable on its own, and it may be especially powerful when combined with theory, experimentation or design-based research. The findings create an opportunity to improve feedback in educational technologies and to advance the learning sciences.

The FFILE workshop aims to advance and encourage research on using data to understand and improve feedback and interactive learning environments. The organizers hope to facilitate the exchange of ideas and the growth of the community of researchers who are interested in these topics. As evidenced by the publications in this volume, using data to understand and improve feedback is important and timely. The papers cover a variety of topics, including rubric-based automated assessment of student drawings of chemical reactions (Rafferty et al.), IRT-based modeling of the effect of feedback on analogical reasoning in children (Stevenson et al.), and an assessment technique for student responses that relies on student participation (Jordan et al.).

Each submission to the workshop was reviewed by three members of a Program Committee, which included the co-chairs and representatives of academia, industry and independent research institutions. The co-chairs thank the Program Committee for diligent reviewing and service.

The co-chairs also thank Erin Walker and Chee-Kit Looi, the AIED 2013 Tutorial and Workshop Chairs, and Andrew Olney and Phil Pavlik, the AIED 2013 Local Arrangements Chairs, for their tireless assistance in helping us organize the workshop.

The workshop will include talks, posters, demos, and interactive activities. The organizers hope that the workshop will be of interest to the wider AIED community.

June, 2013

Ilya Goldin, Taylor Martin, Ryan Baker, Vincent Aleven, Tiffany Barnes

Program Committee

Table of Contents

Automating Guidance for Students' Chemistry Drawings

Anna N. Rafferty
Computer Science Division
University of California
Berkeley, CA 94720
rafferty@cs.berkeley.edu

Libby Gerard
Graduate School of Education
University of California
Berkeley, CA 94720
libby.gerard@gmail.com

Kevin McElhaney
Graduate School of Education
University of California
Berkeley, CA 94720
kevin777@berkeley.edu

Marcia C. Linn
Graduate School of Education
University of California
Berkeley, CA 94720
mclinn@berkeley.edu

ABSTRACT

Generative educational assessments such as essays or drawings allow students to express their ideas. They provide more insight into student knowledge than most multiple-choice items. Formative guidance on generative items can help students engage deeply with material by encouraging students to effectively revise their work. Generative items promote scientific inquiry by eliciting a variety of responses and allowing for multiple correct answers, but they can be difficult to automatically evaluate. We explore how to design and deliver automated formative guidance on generative items requiring precollege students to draw the arrangement of atoms before and after a chemical reaction. The automated guidance is based on a rubric that captures increasing complexity in student ideas. Findings suggest that the automated guidance is as effective at promoting learning as teacher-generated guidance, measured both by immediate improvement on the revised item and pre- to post-test improvement on a near-transfer item. Immediate and delayed delivery of automated guidance are equally effective for promoting learning. These studies demonstrate that embedding automated guidance for chemistry drawings in online curricula can help students refine their understanding. Providing automated guidance can also reduce the time teachers spend evaluating student work, creating more time for facilitating inquiry or attending to the needs of individual students.

Keywords

formative feedback | automatic assessment | chemistry education

1. INTRODUCTION

One of the promises of computer assisted education is the ability to provide timely guidance to students that is adapted to their particular mistakes. Such adaptive formative feedback is provided by human tutors [18], and has been shown to be an important principle in designing computerized tutors [1, 2]. This guidance can scaffold student understanding and address common errors that lead different students to express the same incorrect response. While the majority of computerized tutors provide formative feedback in some form [11, 26], this guidance is often limited to selection tasks or numeric answers. These kinds of answers are easy to evaluate yet may encourage students to recall facts rather than distinguish and integrate ideas.

Generative tasks, in contrast, elicit students' range of ideas and encourage them to use evidence to sort out ideas in order to create a coherent explanation. Mintzes, Wandersee, and Novak point to the fact that generative assessments can provide a fuller picture of students' conceptual understanding and drive students towards "making meaning" rather than memorizing facts [19]. Generative tasks are difficult to evaluate due to the variety of responses and possibilities for multiple ways to express the correct answer. Evaluating student work is time consuming and requires content expertise. Subsequently it is often not possible for teachers to provide detailed guidance to all students [5].

In this paper, we explore how automated formative guidance on student-generated drawings can improve students' conceptual understanding of chemical reactions. By constraining students to use virtual atom stamps, rather than drawing the atoms themselves, we limited the degree to which student drawings could vary while still allowing for expression of different conceptual views. We designed an algorithm to automatically evaluate students' conceptual views, and provided targeted guidance to improve understanding.

We begin by reviewing some of the relevant literature on formative feedback as well as the theoretical framework, knowledge integration, in which our work is grounded. We then describe the drawing tasks that students completed as part of an inquiry-based activity concerning global climate change and the highly accurate automated scoring system we developed. We demonstrate how the automated guidance affects student learning through two classroom studies: one explores the effect of automated guidance compared

to teacher-generated guidance, and the other investigates whether immediate or delayed automated guidance is more effective.

2. BACKGROUND

There has been a great deal of work on the design and use of formative feedback. We briefly overview some of the most relevant literature on formative feedback for science learning, as well as the knowledge integration framework, which is the pedagogical theory underlying the design of our assessment and guidance.

2.1 Formative Feedback

Formative assessment can help teachers to recognize students' level of understanding and adapt instruction. Ruiz-Primo and Furtak [21] found that teachers' informal use of this type of assessment was related to their students' performance on embedded assessment activities, suggesting that this monitoring can indeed help teachers boost student learning. Guidance based on these assessments provides a way to help students to improve their understanding and recognize gaps or inconsistencies in their ideas [10].

While formative assessment and guidance can be helpful for learning, it is difficult to determine how to design this guidance for generative and open-ended tasks. These tasks facilitate a variety of student responses, and the best form of guidance for promoting learning and conceptual understanding based on students' current knowledge is unclear. Some work has had success at automatically scoring student-generated short answers (e.g., [3],[13]), leading to the potential for conceptual guidance based on these scores. In the science domain, automated feedback has also been effective at driving student learning when creating and revising concept maps [24]. For inquiry learning, there has been significant interest in how to effectively scaffold student learning using technology [20]. While often not aimed directly at guidance, machine learning techniques have been employed to automatically recognize effective inquiry learning skills [22]. Our work adds to this body of literature on formative feedback in open-ended science tasks by demonstrating that drawing tasks in which students pictorially represent scientific ideas are amenable to automatic evaluation. We test how different ways of providing guidance affect student learning.

2.2 Knowledge Integration

The drawing tasks we examine are part of a chemical reactions unit [7] built in the Web-based Science Inquiry Environment (WISE) [16]. This environment is based on the theory of knowledge integration [15]. Knowledge integration is based on constructivist ideas that focus on building on students' prior knowledge and helping them to connect new concepts with this knowledge, even if some of this prior knowledge is non-normative (e.g.,[27]). Knowledge integration consists of four main processes: eliciting existing student ideas, adding new ideas, distinguishing ideas, and sorting ideas into cohesive understandings [14]. Within WISE, these processes are targeted by activities within an inquiry-based learning module. Each module is organized around a central topic, such as understanding climate change, and the activities may include answering multiple choice or short answer questions, watching a visualization, or creating a drawing to illustrate a scientific phenomenon. For instance, the chemical reactions unit contains visualizations of how energy from the sun is reflected by the Earth and transformed into heat energy. This visualization may add to students' existing ideas as well as help them to see cases that are not accounted for by these existing ideas. Later in the unit, students' understanding is challenged through the introduction of new concepts, such as pollution, into both the visualization and the general investigation of why climate change occurs. This adds new ideas to the student's existing model and prompts revision of the student's ideas to form a more complete understanding. The knowledge integration framework has been the building block for a number of WISE units, and has also been revised and used for pedagogical design in other settings [8, 25].

In the context of knowledge integration, generative tasks elicit students' existing ideas and help them to clarify and distinguish their ideas from one another. Through this process, they may form more cohesive conceptual understandings. For example, a student might make a drawing or write a textual explanation of the visualization she observed. This prompts her to pull out individual ideas and consider how to connect what she saw in the visualization with her prior knowledge. Formative guidance can assist students by prompting them to revise their ideas and evaluate their consistency with normative scientific ideas, which may be articulated or referred to in the feedback [17]. When this guidance is based on students' own ideas, as articulated in their initial response to the activity, it can directly help students to develop criteria for distinguishing between normative and non-normative ideas and push students to integrate ideas rather than holding separate, conflicting conceptions [16].

3. DRAWING CHEMICAL REACTIONS

We focus our investigation of formative feedback on students' drawings of chemical reactions. These drawings show students' particulate understanding of how atoms are rearranged in a reaction. Past work has shown that learning multiple models of chemical reactions and providing students with ways of visualizing the particles involved in the reactions can help to strengthen student understanding [9, 23]. The drawing tasks are part of a WISE unit entitled *Chemical Reactions: How Can We Help Slow Climate Change?*, which focuses on students' understanding of chemical reactions [7]. As shown in Figure 1(a), these drawing tasks ask students to draw the arrangement of atoms before and after a chemical reaction; one of the tasks focuses on the combustion of methane while the other involves the combustion of ethane. The WISE Draw screen provides students with "stamps" for each atom; for instance, the methane reaction problem includes stamps for oxygen, carbon, and hydrogen. Students must choose how many of each atom to add to their drawing and arrange the atoms to reflect how they are grouped into molecules. Students then create a new frame in their drawing to show the products of the reaction. The drawings enable students to articulate their ideas about chemical reactions and to work with a different model of these reactions than the typical equation based format.

Both the methane and ethane tasks ask the student to show the combustion of oxygen and a hydrocarbon, resulting in the products carbon dioxide and water. In the methane drawing, students are asked to draw two methane molecules

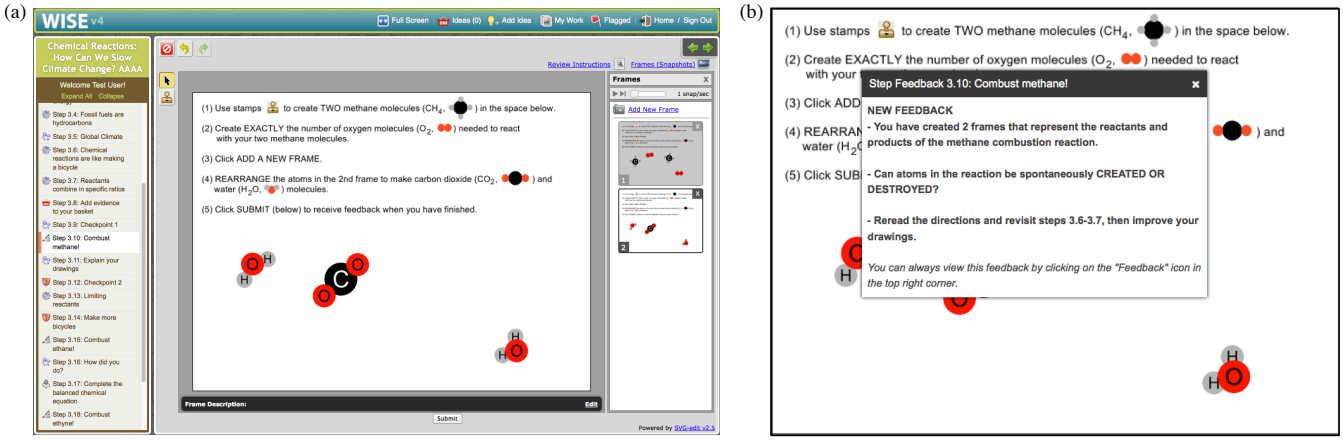


Figure 1: The WISE drawing environment. (a) A screenshot of a student drawing. Students place atom stamps on the central canvas to show the molecules at the beginning and end of a chemical reaction. On the right side of the screen, the two frames that the student has created are shown. (b) The student drawing canvas with automated guidance. The student has submitted her drawing, and a pop up box appears with adaptive textual feedback to help her develop her conceptual understanding of chemical reactions.

and as many oxygen molecules as are required for complete combustion of the methane. This item thus requires students to reason about how many oxygen molecules each methane molecule reacts with. For the ethane drawing, students are told to illustrate ten oxygen molecules and two ethane molecules as the reactants, and then to rearrange them to form the products. This leaves three oxygen molecules that are unchanged by the reaction.

4. PROVIDING GUIDANCE ON STUDENT DRAWINGS

Since the drawing tasks assess important conceptual ideas about chemical reactions and students frequently make errors on these tasks, they are a natural target for providing students with formative feedback. Our goal is to provide conceptual guidance that targets errors that the student has made. This requires detecting errors in the drawing and creating guidance for each category of conceptual errors.

4.1 Evaluating Student Drawings

To evaluate student drawings, we created an algorithm that processes each drawing and assigns it a score. We used a development set of 98 drawings from past students, half from each item, to determine the most common errors and to tune the parameters of the scoring algorithm. Of these 98 drawings, 45% were correct, as marked by a human evaluator.

Examination of the student drawings showed many similar errors across students. We grouped these errors into conceptual categories, shown in Table 1. Category 0 includes drawings that do not have two frames, one for the reactants and one for the products. In some cases, this may be due to difficulties using the drawing interface. Category 1 corresponds to lack of conservation of mass. Student drawings with this error have different atoms in the reactant and product frames. Category 2 corresponds to drawings that conserve mass, but have incorrect reactants. This may be due to having the wrong number of molecules, or to having atoms incorrectly arranged into molecules. Category 3 refers to drawings that have correct reactants, but incorrect products. For instance, a student might combust only one methane molecule, incorrectly leaving one methane and two oxygen molecules in the products. Category 4 includes drawings that are nearly correct, but where molecules are overlapping; for example, four oxygen atoms might be arranged in a square, rather than arranged in two distinct groups. Finally, Category 5 includes correct drawings.

In order to facilitate feedback across a variety of chemical reaction drawings, we separated the scorer into a scoring algorithm and a specification file. The scoring algorithm maps the drawing into one of the six categories described above, drawing information from the specification file to determine the correct configuration of atoms into molecules and what molecules are correct for each frame. In the methane case, for example, the specification file lists four allowed molecules: oxygen, methane, carbon dioxide, and water. Each molecule is defined by the atoms that it includes and how these atoms touch one another. For instance, the specification file indicates that carbon dioxide includes one carbon and two oxygen atoms, and each oxygen atom must touch the carbon atom. The specification file also lists the correct reactants and products for the given reaction. While this level of expressivity was sufficient for our tasks, which have a single correct set of molecules that should be present in each frame, the specification file and scorer could easily be extended to specify non-unique correct answers, such as requiring that the products should have twice as many of one molecule as another.

Student drawings are saved as SVG strings, an XML-based vector image format, which facilitates automatic processing. Each string indicates how many frames exist, what stamps are in each frame, and the location of each stamp. The specification file lists how stamps (image files) correspond to atoms, so the string effectively indicates the location of each atom in the drawing. The automated scoring algorithm has three stages: pre-processing, identifying molecule groupings,

Criteria	0	1	2	3	4	5
Two frames		✓	✓	✓	✓	✓
Conserves atoms			✓	✓	✓	✓
Correct reactants				✓	✓	✓
Correct products					✓	✓
Groupings clear						✓
Rate in dev. set	11%	19%	16%	5%	3%	45%

Table 1: The scoring rubric. Each level adds an additional criterion that must be met. The bottom row indicates the proportion of drawings in the development set with each score.

and assigning a numerical score. Pre-processing removes stamps that are outside of the viewable image area, often due to a student dragging a stamp offscreen rather than deleting it. This stage also removes duplicate stamps that have identical or almost identical center locations; this can occur when a student double-clicks to place a stamp. The pre-processing steps thus makes the SVG string correspond more closely to the image as a viewer would perceive it.

After pre-processing, atom stamps are grouped into molecules, and the frames are annotated with the atoms and molecules that they contain. Atoms are part of the same molecule if they are visually grouped. This is indicated by the atoms directly touching, with atoms in one molecule not touching atoms in another molecule. Small spaces between the atoms in a molecule and small amounts of overlap are ignored by our algorithm due to our focus on conceptual errors; these issues are more likely to be due to the constraints of the medium than evidence of student misunderstanding.

Algorithmically, the grouping of atoms into molecules is computed via depth-first search and by solving a constraint satisfaction problem [28]. Depth-first search computes the connected components of the drawing, where a component is connected if all images in that component are within ϵ of at least one other image in the component; given small $\epsilon > 0$, atoms can be in the same molecule but not directly touch. Components are then matched to molecules, where a match is valid if the identity of the atoms in the specification and in the drawing are the same and if the touching relations given in the problem specification are satisfied; this is implemented as constraint satisfaction. If one connected component can only be recognized as consisting of several molecules, the drawing is marked as having overlapping molecules unless the overlap is less than some constant. Again, this constant allows us to ignore small amounts of overlap.

Based on the annotations of the molecules and atoms in each frame, the numerical score for the drawing is computed based on the rubric in Table 1. For instance, if the number of atoms of each type changes between the first and second frames, the drawing is given a score of 1. If the drawing conserves mass but reactants are not correct, the drawing is given a score of 2, regardless of whether the products are correct. A score of 4 is given only if all atoms in the frames are correct, and the scorer recognized that the correct molecules were present but overlapping.

We evaluated the accuracy of the algorithm on both the development set and on pilot data from 251 student drawings. In both cases, the drawings were scored by a trained human scorer, and these scores were compared to the automated scores. On the development set, the automated score matched the human score on 97% of the drawings. Accuracy was very similar for the pilot data, which was not used in the creation of the scorer: automated scores matched the human score on 96% of the drawings.

4.2 Creating Guidance from Scores

Given that the scoring algorithm is quite accurate, we can provide guidance based on the conceptual understanding that the student has displayed in the drawing. For each of the six possible scores, we designed a textual feedback message to help students revise their drawing. We chose to use textual feedback to facilitate a comparison between automated and teacher-generated guidance. The WISE platform supports teacher guidance by allowing teachers to view student work and type comments to each student group.

The textual feedback was designed to promote knowledge integration by recognizing students' normative ideas and helping them to refine and revise their non-normative ideas [16]. Drawings that were scored as having some conceptual error (scores 0-4) all received textual feedback of a similar format. First, a correct feature of the drawing was recognized, anchoring the guidance with students' prior knowledge. For example, a student who received a score of 2 would be praised for conserving mass, since this is the conceptual feature that bumped the student from a score of 1 to 2. The textual feedback then posed a question targeting the student's conceptual difficulty, such as identifying what molecules should be present in the reactant frame; this elicits student ideas about the topic of difficulty. Finally, the feedback directed students to a relevant step earlier in the unit, and encouraged them to review the material in that step and then to revise their drawing. This promotes adding new ideas and distinguishing normative and non-normative ideas. The feedback for a score of 1 is shown in Figure 1(b).

5. STUDY 1: EFFECTIVENESS OF AUTOMATED GUIDANCE

To test the effectiveness of our automated guidance system, we compared student learning when given automated or teacher-generated guidance. In this study, automated guidance was provided to students upon request, taking advantage of the fact that automation facilitates immediate feedback. Based on evaluation of the existing student drawings, we believed the automated scorer would have relatively high accuracy, but the guidance it can provide is still less specific than that which teachers can provide. The teachers could adjust guidance for individual students, while there were only six different automated feedback messages that a student might receive. Since prior work has had mixed results concerning whether specific or general feedback is more helpful(e.g., [6],[12]), it is not clear whether the lack of specificity in the automated guidance will be a disadvantage.

5.1 Methods

5.1.1 Participants

A total of 263 students used the WISE unit and completed both the pre- and post-tests.

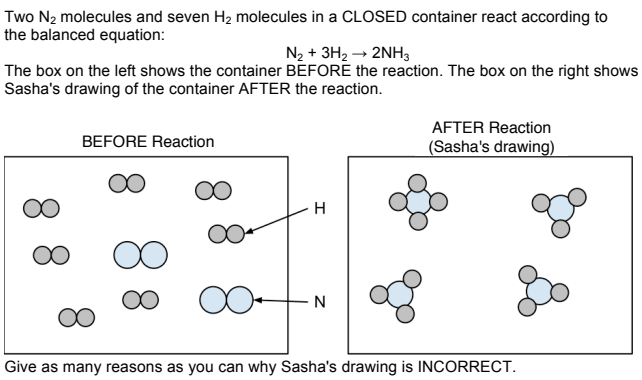


Figure 2: Item from the pre- and post-test related to drawing chemical reactions. Students are asked to examine Sasha's drawing and explain why the drawing is incorrect. The drawing task is similar to those in the unit, but asks students to evaluate rather than generate the drawing and requires integrating the equation and the drawing.

5.1.2 Study design

Students were assigned on a full-class basis to receive either automated or teacher-generated guidance. Two teachers from the same public middle school participated in the study, using the WISE activity in their eighth grade physical science classes. The activity took approximately five hours, spread over multiple class periods. The first teacher had 139 students in five classes; three of these classes received automated guidance and two received teacher guidance. The second teacher had 124 students, also spread over five classes; again, three of the classes were assigned automated guidance and two were assigned teacher guidance. This led to 155 students in the automated condition and 108 students in the teacher guidance condition. Students used WISE in groups of between one and three students; there were 71 groups in the automated condition and 58 in the teacher condition, although a small number of students in these groups did not complete the pre-test or the post-test.

All students experienced the same activities in the WISE unit except for the draw steps. On the two draw steps, all students received the same instructions, except that students in the automated condition were told to click the "Submit" button when they wished to receive feedback. When students clicked this button, they were warned that they only had two chances to receive feedback and to confirm that they wanted to proceed. After confirming, a pop-up box with the textual feedback appeared, as in Figure 1(b). Students could close the feedback or re-open it to view their existing feedback at any time.

Students in the teacher-generated guidance condition did submit their work. Instead, teachers provided feedback to these students using the WISE Grading Tool after the students made a drawing. When students signed in to the activity the following day, they were informed that they had received feedback, and teachers also reminded the students to revise their drawings based on the comments. This condition was intended to mirror how teachers usually give feedback to student work in WISE. Due to time constraints, students in this condition received only one round of feedback.

Students in all conditions completed a pre- and post-test assessment. Both assessments contained the same items. As shown in Figure 2, one of these items asked students to examine a drawing of a chemical reaction and to explain why the drawing was incorrect. This item addresses some of the same conceptual skills as the drawing tasks in the unit, and thus can be used as a transfer measure of student learning from the draw activities. Unlike the WISE unit, these assessments were completed by students individually.

5.2 Results

Overall, students improved their drawings by 0.9 points after receiving guidance, as computed via the automated algorithm. An analysis of variance of student scores on the drawing items with factors for revision that received feedback versus final revision and feedback condition, as well as a random factor for student group, showed that there was a main effect of revision ($F(1, 142) = 68.8$, $p < .001$), indicating the improvement was significant. However, there was not a main effect of condition: improvement was nearly identical for students who received automated guidance and those who received teacher guidance, and both groups had similar initial scores.

While amount of improvement on the drawing items is similar for both conditions, one might be concerned that students in the automated guidance condition have an advantage on this metric since their feedback is directly based on the scoring rubric. Comparison of the proportion of groups revised an incorrect drawing to be correct suggests that this is unlikely to be the case: 27% of groups who were initially incorrect revised their drawing to be correct in the automated condition, compared to 30% in the teacher-feedback condition. Thus, comparable number of students were able to completely correct their work in both conditions.

The improvement from pre- to post-test of student answers on the item concerning evaluation of another student's drawing provides another way of comparing student learning across conditions (see Figure 2). Student answers on this item were evaluated using the rubric in Table 2. This rubric gives higher scores to student answers that include more correct ideas and that connect conceptual ideas with features from the drawing, consistent with the knowledge integration focus on creating a cohesive conceptual understanding. While some of these concepts, such as conservation of mass, were addressed in the drawing items in the unit, the item asks students to go beyond the initial drawing tasks by articulating the connections between the drawing and the equation for the chemical reaction. Students in both conditions improved significantly on this item from pre- to post-test: an average of 0.37 points for students in the automated condition ($t(154) = 4.63$, $p < .005$) and an average of 0.27 points for students in the teacher-feedback condition ($t(107) = 2.93$, $p < .01$). An analysis of variance showed that there was no main effect of feedback type on amount of improvement. Like the results of the improvement in drawings, this suggests that the automated feedback is as helpful for student learning as the teacher-generated feedback.

Inspection of the teacher comments revealed that one teacher gave substantially more detailed and conceptually focused comments than the other. This teacher used a relatively

Score	Criteria
1	Blank or no scientific ideas.
2	Invalid scientific ideas or only correct ideas about products, failing to explain why the products are incorrect.
3	Incomplete scientific ideas: isolated ideas about too few hydrogen in Sasha's drawing or about product identity, without connecting to concepts.
4	One complete statement linking a feature of Sasha's drawing with why it is incorrect.
5	Identification of at least two errors, with complete statements linking the features of Sasha's drawing with why they are incorrect.

Table 2: The knowledge integration scoring rubric for the pre- and post-test item.

small number of comments for all students, customizing these comments slightly on a case by case base, and each one tended to focus on a particular conceptual issue. For example, one comment was *"You have only made one frame to represent the products and reactants. Your first frame should be for the reactants. A second frame should be made for the products. Follow the directions on the top of the page."* This comment combines procedural elements connecting to the student drawing with conceptual ideas. In contrast, the second teacher tended to give short comments that were solely procedural or solely conceptual. These comments commonly directed students to read the directions or stated a concept in isolation, such as the comment *'Conservation of mass?"*. These comments may have been too terse to help students connect concepts with their drawings.

Due to these differences in comments, we analyzed how effective the feedback was at helping students based on what type of feedback they received as well as which teacher they had in the teacher-feedback condition. An analysis of variance on the amount of improvement in drawing scores from initial feedback to final revision, with a factor for feedback type (automated, Teacher 1, or Teacher 2) and a random factor for student group, showed that feedback condition did have an effect on amount of improvement ($F(2, 127) = 4.4$, $p < .05$). As shown in Figure 3, students who received more cohesive guidance (Teacher 1) improved more than students in the other conditions, and students who received automated guidance improved more than students who received terse guidance (Teacher 2). Note that this is not an overall difference between response to guidance based on whether students were in a class with Teacher 1 versus Teacher 2: students in the automated condition showed similar improvement across teachers. While this interaction was not significant for the pre- to post-test improvement, the same trend held: students who received feedback from Teacher 1 improved an average of 0.37 points, students in the automated condition improved 0.35 points, and students who received feedback from Teacher 2 improved 0.12 points.

6. STUDY 2: TIMING OF GUIDANCE

The previous study showed that automated guidance is comparable to teacher-generated guidance in helping students to revise their drawing and improving post-test scores. However, the two types of guidance were not administered under the same timing schedule: automated guidance was given to students when they asked for it, while teacher guidance was given to students at a fixed delay. We hypothesized that immediate guidance would be more engaging and motivating to students, but delayed guidance might boost retention by allowing students to space their studying of the concepts. Students who are frustrated with the problem may also benefit from a chance to do other activities before receiving guidance. To explore these issues, we conducted a new study in which all students received automated guidance, but some were given the guidance immediately, just as in the automated condition in Study 1, while others received the guidance at a delay, following the same pattern as the teacher guidance in Study 1.

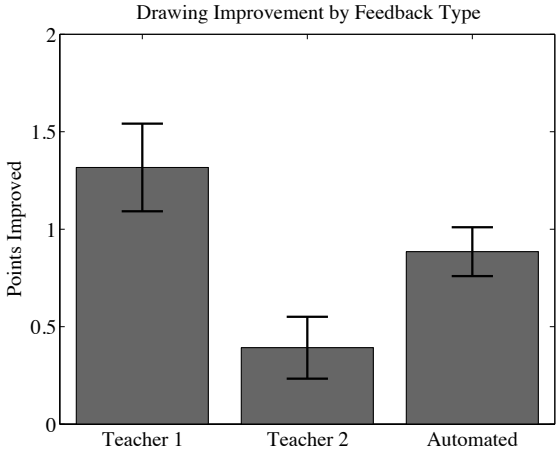


Figure 3: Improvement on drawing scores based on type of feedback received. Error bars indicate one standard error.

6.1 Methods

6.1.1 Participants

A total of 88 students used the WISE unit and completed both the pre- and post-tests.

6.1.2 Study design

Students were assigned to the immediate or delayed guidance conditions on a full-class basis. All classes were taught by the same teacher in a public high school. He used the activity in his four ninth grade basic chemistry classes. Two classes were assigned to the immediate guidance condition, and two were assigned to the delayed guidance condition. As in Study 1, students completed the activity in groups of one to three students; there were 30 groups in the immediate condition and 27 groups in the delayed condition.

The immediate guidance condition in this study was identical to the automated condition in Study 1. The delayed guidance was provided to students after they had completed their initial drawings, and was added to the grading tool overnight. When students signed into the activity the following day, they were informed that they had new feedback and shown the textual comments. In both cases, the comments students received were based on the score of their drawing,

and the text was identical to that of Study 1. Students in the immediate guidance condition could submit their drawing up to two times; due to time constraints, students in the delayed condition received only a single round of feedback.

The pre- and post-test had the same items as in Study 1 and were again completed by students individually.

6.2 Results

Students showed similar improvements in their drawings across conditions. Students in the immediate condition improved their drawing scores by an average of 0.65 points, while students in the delayed condition improved their drawing scores by an average of 0.81 points. A repeated-measures analysis of variance including factors for revision (initial versus final) and guidance condition showed that there was a main effect of revision ($F(1, 65) = 25.2$, $p < .001$), but no significant effect of condition.

In Study 1, we collapsed across the two drawing items as students showed similar improvements across items. However, in this study, there was a trend towards greater improvement on the ethane item for students in the delayed condition versus the immediate condition, while both types of guidance resulted in similar improvement on the methane item. A repeated measures analysis of variance on the amount of improvement with factors for guidance condition and item showed that the interaction between the two factors was marginally significant ($F(1, 52) = 3.44$, $p = .0695$). One reason for this interaction may simply be the placement of these items in the unit: ethane occurs after methane, late in Activity 3 of the WISE unit. Students in the immediate condition may be rushing through the ethane item in order to finish, while students in the delayed condition come back to the items on a later day. Yet, other factors could also contribute to this difference, such as frustration in low-performing students due to the repeated interactive sequences or some item-specific factor.

On the post-test item asking students to evaluate Sasha's drawing, students showed small improvements from their pre-test scores, with an average improvement of 0.19 points. A repeated measures analysis of variance with factors for pre- versus post-test and feedback condition showed that both main effects were significant (pre- versus post test: $F(1, 86) = 4.58$, $p < .05$; condition: $F(1, 86) = 4.12$, $p < .05$). Closer examination revealed relatively little improvement for students in the delayed condition (an average of 0.073 points) compared to an improvement of 0.30 points for students in the immediate condition; by chance, students in the delayed condition also began with higher pre-test scores, although their initial drawing scores were similar.

Overall, this study suggests that immediate and delayed guidance have similar effects on student revision, and immediate guidance may be more helpful for retention and transfer based on the pre- to post-test improvement. Given the difference in effectiveness between the two conditions for improvement on the methane and ethane items, we plan to investigate whether changing the placement of the items within the activities reduces the differences between immediate and delayed guidance. More broadly, we will explore whether students might be helped by different guidance timing for some types of drawing items versus others.

7. DISCUSSION

Formative guidance can help students to improve their understanding of a topic and focus their efforts on the material that is most critical given their current knowledge. We investigated how to provide this guidance in the context of constrained drawing tasks. These tasks allow students to articulate their ideas, including misunderstandings, more fully than multiple choice questions, but are harder to evaluate automatically and too time consuming for teachers to evaluate in many classrooms. We found that by constraining the space of feedback to target six levels of conceptual understanding, we could classify the drawings automatically and help students to improve their understanding. We now turn to some possible next steps for providing formative guidance on drawing items using our automated scoring algorithm.

In our initial studies, we focused on textual feedback in order to compare automated and teacher-generated guidance. However, one of the benefits of a computer-based system is the ability to give other types of guidance, such as interactive activities or guidance that combines text and images. These types of guidance might be more engaging for students, and provide more help for those students who are less motivated or struggle to understand the text-based conceptual feedback. We are currently exploring guidance in the form of interactive activities that place students in the role of evaluating a drawing rather than generating it, just as in the post-test assessment item. The specific activity provided is based on the score of the student's initial drawing.

Another area that we would like to explore in future work is whether more specific or detailed guidance might be helpful for some students. We have observed that some students find it challenging to connect the text-based conceptual feedback with their own drawings. While some level of difficulty is desirable in order to push students to make connections and revise their understanding [4], guidance that is incomprehensible to students is unlikely to help them learn. The automated scoring algorithm provides the potential to scaffold students in their attempt to uncover what is wrong. For instance, if the student has incorrectly grouped some atoms, the algorithm could show the student only the relevant portion of the screen and ask them to explain why that portion was incorrect. This would still prompt students to reflect on their drawings and understanding, but would more closely connect the guidance to their own work. Creating connections between the drawings and the chemistry concepts was common in the guidance of the more effective teacher, suggesting that strengthening these connections in the automated guidance would promote student learning.

The issue of timing and agency when giving feedback remains another useful area for exploration. In Study 2, we compared immediate feedback versus delayed feedback for students, where feedback timing was independent of drawing quality. To better understand how timing of guidance affects learning, we hope to conduct experiments in which timing is based on the score of the current drawing or particular characteristics of students' previous work. These customizations may also allow some students to choose when they would like guidance (as in the immediate condition in

Study 2) while automatically providing guidance to others.

Automatically scoring generative tasks in computerized tutors can be difficult, but is usually a prerequisite of providing adaptive formative feedback on the tasks. In this work, we created an automated scorer for a particular type of constrained yet generative drawing task. This scorer is easily customized to evaluate new drawing items that follow the same pattern as those in the unit, and is able to detect common conceptual errors that students make. Drawing on the knowledge integration pattern, we developed textual guidance for these conceptual errors. Our studies show that that this automated guidance results in comparable learning as guidance given by a teacher. The automated scorer facilitates experimentation with different types of formative feedback, allowing us to test hypotheses about what types of guidance are most effective for promoting understanding in open-ended science activities.

Acknowledgements. This research was supported by a DoD NDSEG fellowship to ANR and by NSF grant number DRL-1119670 to MCL.

8. REFERENCES

[1] J. Anderson, C. Boyle, R. Farrell, and B. Reiser. Cognitive principles in the design of computer tutors. *Modelling Cognition*, pages 93–133, 1987.

[2] J. R. Anderson, A. T. Corbett, K. R. Koedinger, and R. Pelletier. Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4(2):167–207, 1995.

[3] Y. Attali and J. Burstein. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3), 2006.

[4] R. A. Bjork. Memory and metamemory considerations in the training of human beings. In *Metacognition: Knowing about knowing*, pages 185–205. The MIT Press, Cambridge, MA, 1994.

[5] P. Black and D. William. Assessment and classroom learning. *Assessment in Education*, 5(1):7–74, 1998.

[6] D. L. Butler and P. H. Winne. Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65(3):245–281, 1995.

[7] J. Chiu and M. Linn. Knowledge integration and wise engineering. *Journal of Pre-College Engineering Education Research (J-PEER)*, 1(1):1–14, 2011.

[8] E. A. Davis and J. S. Krajcik. Designing educative curriculum materials to promote teacher learning. *Educational Researcher*, 34(3):3–14, 2005.

[9] A. G. Harrison and D. F. Treagust. Learning about atoms, molecules, and chemical bonds: A case study of multiple-model use in grade 11 chemistry. *Science Education*, 84(3):352–381, 2000.

[10] J. Hattie and H. Timperley. The power of feedback. *Review of educational research*, 77(1):81–112, 2007.

[11] K. Koedinger, J. Anderson, W. Hadley, and M. Mark. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8(1):30–43, 1997.

[12] K. R. Koedinger and V. Aleven. Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, 19(3):239–264, 2007.

[13] C. Leacock and M. Chodorow. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405, 2003.

[14] M. Linn, B. Eylon, and E. Davis. The knowledge integration perspective on learning. *Internet Environments for Science Education*, pages 29–46, 2004.

[15] M. C. Linn. Chapter 15: The knowledge integration perspective on learning and instruction. In *The Cambridge handbook of the learning sciences*, pages 243–264. Cambridge University Press, New York, NY, 2004.

[16] M. C. Linn and B. Eylon. *Science Learning and Instruction: Taking Advantage of Technology to Promote Knowledge Integration.* Routledge, 2011.

[17] M. C. Linn and B. S. Eylon. *Science Education: Integrating Views of Learning and Instruction*, pages 511–544. Lawrence Erlbaum Associates, Mahwah, NJ, 2006.

[18] D. Merrill, B. Reiser, M. Ranney, and J. Trafton. Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems. *The Journal of the Learning Sciences*, 2(3):277–305, 1992.

[19] J. J. Mintzes, J. H. Wandersee, and J. D. Novak. *Assessing science understanding: A human constructivist view.* Academic Press, 2005.

[20] C. Quintana, B. J. Reiser, E. A. Davis, J. Krajcik, E. Fretz, R. G. Duncan, E. Kyza, D. Edelson, and E. Soloway. A scaffolding design framework for software to support science inquiry. *The Journal of the Learning Sciences*, 13(3):337–386, 2004.

[21] M. A. Ruiz-Primo and E. M. Furtak. Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching*, 44(1):57–84, 2007.

[22] M. A. Sao Pedro, R. S. de Baker, J. D. Gobert, O. Montalvo, and A. Nakama. Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. *User Modeling and User-Adapted Interaction*, 23(1):1–39, 2013.

[23] P. Schank and R. Kozma. Learning chemistry through the use of a representation-based knowledge building environment. *Journal of Computers in Mathematics and Science Teaching*, 21(3):253–279, 2002.

[24] J. R. Segedy, J. S. Kinnebrew, and G. Biswas. The effect of contextualized conversational feedback in a complex open-ended learning environment. *Educational Technology Research and Development*, 61(1):71–89, 2013.

[25] S. Sisk-Hilton. *Teaching and Learning in Public: Professional Development through Shared Inquiry.* Teachers College Press, 2009.

[26] J. Slotta and M. Linn. *WISE science: Web-based inquiry in the classroom.* Teachers College Press, 2009.

[27] J. P. Smith III, A. A. Disessa, and J. Roschelle. Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences*, 3(2):115–163, 1994.

[28] E. Tsang. *Foundations of Constraint Satisfaction.* Academic Press London, 1993.

Estimating the Effect of Web-Based Homework

Kim Kelly, Neil Heffernan, Cristina Heffernan
Worcester Polytechnic Institute
kkelly@wpi.edu

Susan Goldman, James Pellegrino, Deena Soffer-Goldstein
University of Illinois-Chicago

ABSTRACT

Traditional studies of intelligent tutoring systems have focused on their use in the classroom. Few have explored the advantage of using ITS as a web-based homework (WBH) system, providing correctness-only feedback to students. A second underappreciated aspect of WBH is that teachers can use the data to more efficiently review homework. Universities across the world are employing these WBH systems but there are no known comparisons of this in K12. In this work we randomly assigned 63 thirteen and fourteen year olds to either a traditional homework condition (TH) involving practice without feedback or a WBH condition that added correctness feedback at the end of a problem and the ability to try again. All students used ASSISTments, an ITS, to do their homework but we ablated all of the intelligent tutoring aspects of hints, feedback messages, and mastery learning as appropriate to the two practice conditions. We found that students learned reliably more in the WBH condition with an effect size of 0.56. Additionally, teacher use of the homework data lead to a more robust and systematic review of the homework. While the resulting increase in learning was not significantly different than the TH review, the combination of immediate feedback and teacher use of the data provided by WBH resulted in increased learning compared to traditional homework practices. Future work will further examine modifications to WBH to further improve learning from homework and the role of WBH in formative assessment.

Keywords

Intelligent tutoring system, immediate feedback, homework, effect size, formative assessment

1. INTRODUCTION

Several studies have shown the effectiveness of intelligent tutoring systems when used in the classroom [9 & 11], reporting effect sizes up to 0.78. However, very few studies have explored the effectiveness of ITS when used as homework. Cooper et al. [3] highlight the point that poorly conceived homework does not help learning. Therefore it was very encouraging when Van Lehn et al. [12] presented favorable results when ANDES, an ITS, was used in this fashion. Yet, most systems are not currently designed to be used for nightly homework. Computer aided instruction (CAI), which gives all students the same questions with immediate end-of-question feedback is more applicable than complex ITS for nightly homework as teachers can easily build the content from textbook questions or worksheets. Kulik and Kulik's [5] meta-analysis reviewed CAI and reported an effect size of 0.3 for simple computer based immediate feedback systems. However, these studies were not in the context of homework use and did not focus on how teachers use the data to respond to student performance. Web-based homework systems (WBH) like WebAssign (www.webassign.com) are commonly used in higher ed. These systems are similar to web based computer aided instruction (CAI), providing students immediate feedback and reports to teachers. While VanLehn et al. [12] reported on three such systems used at the higher ed level for physics, there are no studies that we know of at the K12 level that allow this contrast.

Despite the relatively low effect sizes reported in Kulik and Kulik [5], WBH holds promise for improving learning from homework by tailoring practice to individual performance. Doing so enables individuals to get corrective feedback so they can focus on areas where they are not successful. Shute [8] reviews the plethora of studies and theoretical frameworks developed around understanding the role of feedback for student learning. However, teacher use of the feedback was not a focus. Black and William [1] have focused on formative assessments, with an eye on informing the teacher and giving feedback to students. The cognitive science literature suggests that letting students practice the wrong skill repeatedly on their homework is detrimental to learning. In this study we look to measure the effect on learning by comparing simple WBH to a traditional homework (TH) condition representing the type of practice that millions of students perform every night in America and probably around the world. Additionally, we explore how the teacher can use the data to modify and improve instruction.

The current study employed ASSISTments.org, an intelligent tutoring system that is capable of scaffolding questions, mastery learning, and hint and feedback messages [9]. However, for this study, we ablated those features creating an "end-of-problem-correctness-only" feedback system for homework in the WBH condition. The system was also used for the TH condition by further removing the correctness feedback thus emulating traditional paper and pencil homework assignments. ASSISTments is currently used by thousands of middle and high school students for nightly homework. Many teachers enter the textbook homework problems and answers into ASSISTments so their students can receive immediate feedback on the homework and the teachers can then access item reports detailing student performance. This allows for focused classroom review. In the current study we were also interested in examining the effects of teacher review of homework performance based on information derived from the ASSISTments system under each of the two different homework conditions. The goal was to estimate the additional effects of teacher-mediated homework review and feedback following each of the two homework practice conditions – TH and WBH – and also study differences in how teachers might approach homework review given variation in student performance following each type of homework practice.

2. EXPERIMENTAL DESIGN

Participants were 63 seventh grade students, who were currently enrolled in an eighth grade math class, in a suburban middle school in Massachusetts. They completed the activities included in the study as part of their regular math class and homework. Students were assigned to conditions by blocking on prior

knowledge. This was done by ranking students based on their overall performance in ASSISTments prior to the start of the study. Matched pairs of students were randomly assigned to either the TH (n=33) or WBH (n=30) condition.

The study began with a pre-test that was administered at the start of class. This pretest and all the rest of the materials for this study are archived via WebCite so others can see the exact materials, videos and anonymous data at tinyurl.com/AIED2013 [4]. This test consisted of five questions, each referring to a specific concept relating to negative exponents. Students were then given instruction on the current topic. That night, all students completed their homework using ASSISTments (see Kelly, 2012 to experience exactly what students did). The assignment was designed with three similar questions in a row or triplets. There were five triplets and five additional challenge questions that were added to maintain ecological validity for a total of twenty questions. Each triplet was morphologically similar to the questions on the pre-test.

Students in the WBH condition were given correctness-only feedback at the end of the problem. Specifically, they were told if their answer was correct or incorrect. See Kelly [4] to see what these materials looked like and to be able to "play student" in either condition. If a student answered a question incorrectly, he/she was given unlimited opportunities to self-correct, or he/she could press the "show me the last hint" button to be given the answer. It is important to emphasize that this button did **not** provide a hint; instead it provided the correct response, which was required to proceed to the next question.

Students in the TH condition completed their homework using ASSISTments but were simply told that their answer was recorded but were not told if it was correct of not (it says "Answer recorded"). It is important to note that students in both conditions saw the exact same questions and both groups had to access a computer outside of school hours. The difference was the feedback received and the ability for students in the WBH condition to try multiple times before requesting the answer.

The following day all students took PostTest1. This test consisted of five questions that were morphologically similar to the pre-test. The purpose of this post-test was to determine the benefit of feedback while doing their homework. At that point, students in the WBH condition left the room and completed an unrelated assignment. To mimic a common homework review practice, students in the TH condition were given the answers to the homework, time to check their work and the opportunity to ask questions. This process was videotaped and can be seen in Kelly (2012). After all of the questions were answered (approximately seven minutes) students in the TH condition left the room to complete the unrelated assignment and students in the WBH condition returned to class. The teacher used the item report, generated by ASSISTments to review the homework. Common wrong answers and obvious misconceptions guided the discussion. This process was videoed and can be seen at Kelly [4]. The next day, all students took PostTest2. This test was very similar to the other assessments as it consisted of five morphologically similar questions. This post-test can be found at Kelly [4]. The purpose of this test was to measure the value-added by the different in-class review methods.

3. RESULTS

Several scores were derived from the data collected by the ASSISTments system. Student's HW Average was calculated based on the number of questions answered correctly on the first attempt divided by the total number of questions on the assignment (20). Partial Credit HW Score accounted for the multiple attempts allowed in the WBH condition. Students were given full credit for answers, provided they did not ask the system for the response. The score was calculated by dividing the number of questions answered without being given the answer by the number of total questions on the homework assignment (20). Time Spent was calculated using the problem log data generated in ASSISTments and is reported in minutes. Times per action are truncated at five minutes. Recall that the homework assignment was constructed using triplets. Learning Gains within the triplets were computed by adding the points earned on the third question in each triplet and subtracting the sum of the points earned on the first question in each triplet.

3.1 Learning Gains From Homework

One student, who was absent for the lesson, was excluded from the analysis (n=63). A t-test comparing the pre-test scores revealed that students were balanced at the start of the study ($t(61)=0.29$, $p=0.78$). However, an ANCOVA showed that students in the WBH condition reliably outperformed those in the TH condition on both PostTest1 ($F(1,60)=4.14$, $p=0.046$) and PostTest2 ($F(1,60)=5.92$, $p=0.018$) when controlling for pre-test score. See Table 1 for means and standard deviations. If the difference was reliable a Hedge corrected effect size was computed using CEM [2]. The effect sizes do not take into account pretest. The key result for posttest2 of 0.56 effect size had a confidence interval of between 0.07 and 1.08.

A comparison of HW Average shows that students scored similarly ($F(1,60)=0.004$, $p=0.95$). An ANCOVA reveled that when calculating homework performance using the Partial Credit HW Score, students in the WBH condition performed reliably better than those in the TH condition ($F(1,60)=17.58$, $p<0.0001$). This suggests that with unlimited attempts, students are able to self-correct, allowing them to outperform their counterparts. Similarly, comparing Learning Gains revealed that students with correctness feedback and unlimited attempts to self-correct learned reliably more while doing their homework ($F(1,60)=45.72$, $p<0.0001$).

Table 1: Means, standard deviations (in parenthesis), and effect size for each measure by condition. *Notes a reliable difference.

	TH	WBH	*p*-value	Effect Size
Pre-Test	9% (17)	7% (14)	0.78	NA
PostTest1	58% (27)	69% (21)	0.046*	0.52
PostTest2	68% (26)	81% (22)	0.018*	0.56
HW Average	61% (20)	60% (15)	0.95	NA
Partial Credit HW Score	61% (20)	81% (18)	0.0001*	1.04
Time Spent (mins)	22.7 (9.6)	23.2(6.2)	0.96	NA
Learning Gains	0.03 (0.9)	1.73(1.1)	0.0001*	2.21

A review of the item report further describes this difference in learning gains. As expected, students in the TH condition continued to repeat the same mistake each time the question was encountered resulting in three consecutive wrong responses. Conversely, students in the WBH condition may have repeated the

mistake once or twice but rarely three times in a row, accounting for the learning. While this behavior appears in four out of the five triplets, triplet 1 was analyzed in depth to explain this finding. See Table 2 for an in depth review of Triplet 1 and Figure 1 to see how the teacher observed this finding using the item report.

2: An in depth review of Triplet 1.

	WBH	TH
Got the first correct and the last one correct (already knew)	8	17
Got the first one wrong and last one correct (learned)	18	4
Got the first one correct and the last one wrong (unlearned?)	1	2
Got both the first one and the last one wrong (Failed to Learn)	4	9
Total	**31**	**32**

The first thing that we want to point out is that students in the WBH condition had a significantly lower percentage correct on the first item. To demonstrate this finding an in depth review of triplet 1 is provided. Eight of these students requested the answer on the first question in triplet 1. Presumably students in the WBH condition would use the hint button when they were not sure of the answer. However, in the TH condition, there was no such button, therefore perhaps students were more likely to take other steps when they were confused. These steps might have included looking at class notes, asking a parent or calling a friend for help. While there is no data to explain

Additionally, when looking at students in the WBH condition that could demonstrate learning (they got the first one wrong), 18 out of 22 students (80% of students) demonstrated learning. In one sense this learning benefit might be overestimated, as there were some interesting differences in response behavior between the conditions. Specifically, response time for the initial response shows that perhaps students' approach the problems differently. We analyzed the time it took students to type in their first response on question 4, and found that students in the TH condition took longer (121 seconds) than students in the WBH condition (89 seconds). In fact, the TH condition had 34% of students take over two minutes to generate their first response while the WBH condition only had 17% of students take that long. This difference was not statistically significant. We speculate that this is due to the fact that students in this condition knew they would have multiple attempts to correctly answer the question and that there was no penalty for answering incorrectly on the first attempt. This indicates that students in the WBH condition may have a higher percentage of incorrect first responses due to less thorough processing and would account for the higher number of students who seemingly already knew the material in the TH condition.

The ability to attempt each question multiple times is unique to students in the WBH condition. We suggest that this feature may play an important role in the presented learning gains. While this specific feature was not empirically tested in this study, we can only speculate on its effect. However, it is important to note that students in the WBH condition had on average 49 attempts (standard deviation=24) to answer the 20-question homework assignment. The fewest attempts made by any student was 25 and the most was 140. The average number of times the answer was requested was 4 was a standard deviation of 3.5. This suggests that students in the WBH condition took advantage of the ability to try questions multiple times to learn the material without requesting the correct answer.

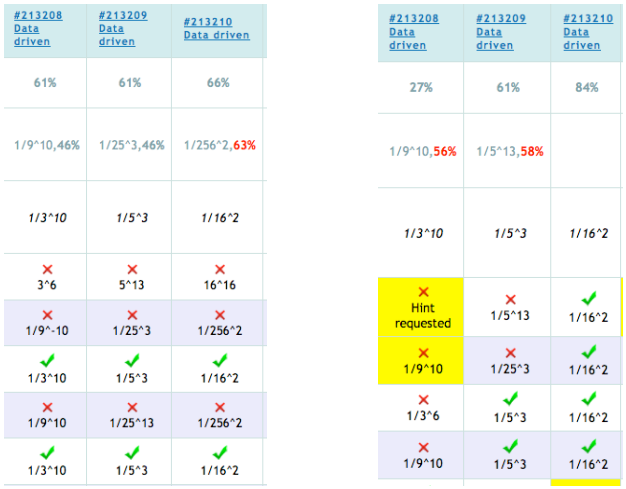


Figure 1: The item report for the control condition (on the left) and experimental condition (on the right) for triplet 1, showing the percent of students answering each question correctly, common wrong answers, the correct answer and several rows of student data.

We were not expecting that correctness only feedback was going to be time efficient. But in fact, students in both conditions spent the same amount of time to complete their homework ($F(1,60)=0.002$, $p=0.96$). However, it appears that the time spent was apportioned differently in the conditions. Specifically, the TH condition took longer to generate a first response, but the WBH condition took time making multiple attempts as well as requesting the answer. It seems that students in the TH group spend more time thinking about the problem but the WBH group can get the problem wrong, and then use their time to learn the content.

3.2 Learning Gains from Homework Review

To address the second research question of the effectiveness of using the data to support homework review, a paired t-test revealed that students in both conditions did reliably better on PostTest2 than on PostTest1 ($t(62)=3.87$, $p<0.0001$). However, an ANCOVA revealed that when accounting for PostTest1 scores, there is not a reliable difference by condition in the gains from PostTest1 to PostTest2 ($F(1,60)=2.18$, $p=0.15$). This suggests that both methods of reviewing the homework lead to substantially improved learning. Interestingly, the results indicate that TH feedback, while students complete homework (69% PostTest1), is as effective as receiving no feedback and then having the teacher review of the homework (68% PostTest2). This suggests that to save time, teachers may not even need to review the homework if students have access to web-based homework systems.

3.3 Observational Results

In addition to examining the effects of immediate feedback on learning, this study explored the potential changes to the homework review process the following day in class. In the traditional format of homework review, time must be spent first on checking answers and then the teacher responds to students'

questions. However, we hypothesized that when teachers have access to the item report they are able to identify common misconceptions and address those ensuring that the time spent reviewing homework is meaningful.

Remember, that when reviewing the homework, students were separated by condition. The teacher recorded herself as she reviewed the homework with each group. In the following section we attempt to characterize what happened in the video segments.

As usual, the teacher reviewed the item report in the morning to determine which questions needed to be reviewed in class. The item report (see Figure 1) shows individual student performance as well as class performance at the question level. Common wrong answers are also displayed for each question. Using this information, the teacher noted that triplet 1 showed a common misconception when multiplying powers with like bases. While the item report shows that students learned from the feedback, the teacher still felt it was important to highlight and discuss the error in multiplying the bases of the powers together. Therefore the teacher highlighted question 4.

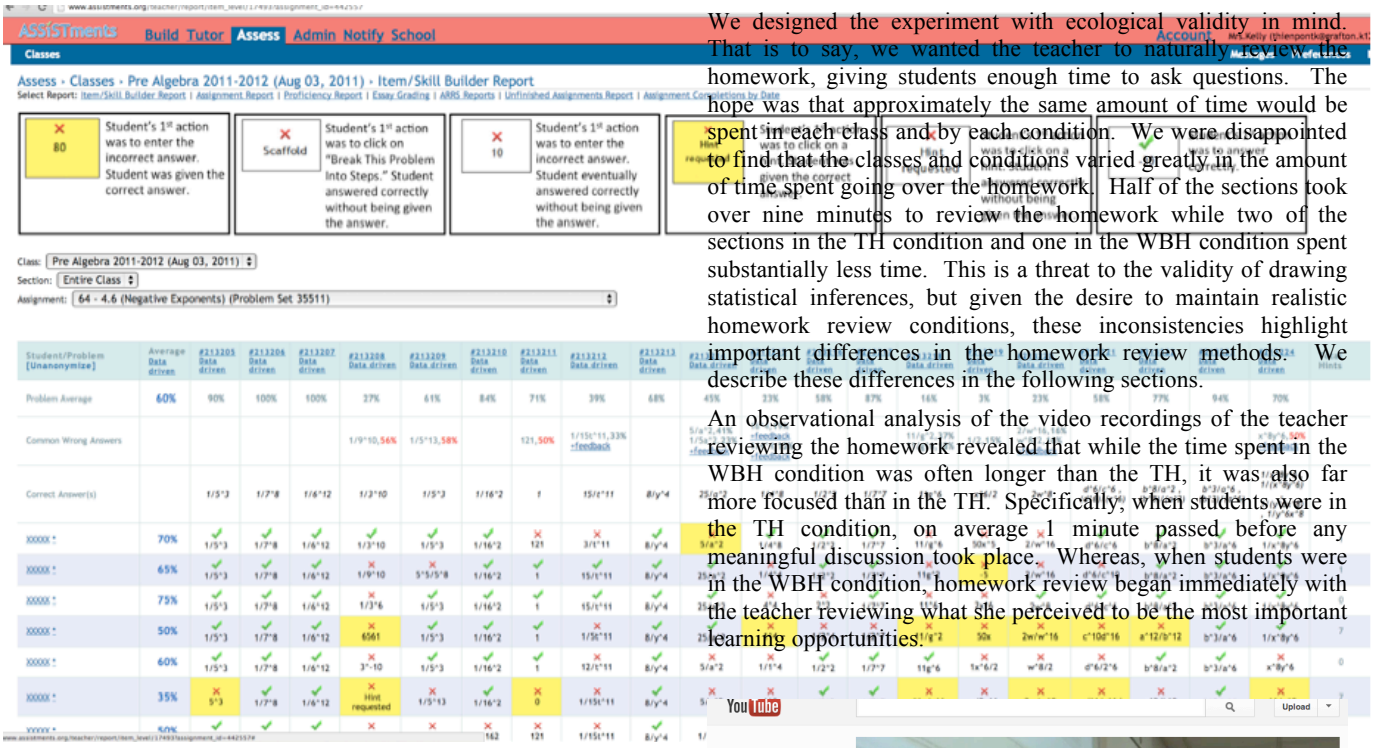


Figure 2: The item report for the WBH condition as viewed by the teacher. Note that class performance for each question and common wrong answers are provided along with individual student performance.

This was especially important because in triplet 2, students incorrectly applied this concept. Specifically, 39% of students initially got this type of question right (multiplying powers with coefficients and variables). However, learning took place as 68% got the next similar question right. It was therefore puzzling to see that on the third question in that triplet (question number 10), only 45% got the question right. Upon investigating the question, the teacher was able to identify the misconception and therefore addressed it with the class. Students learned in the prior triplet not to multiply the bases together. However, in this problem $(5a^3)(5a^{-5})$ students didn't realize that they should multiply the coefficients, 5 and 5 together. You can see in the video that the teacher highlights the difference between these types of problems.

The third and fifth triplet showed adequate learning. Additionally, questions 1, 2, and 3 were introductory questions and performance was above 90% on each question, therefore the teacher did not feel the need to address any of these questions. Similarly, questions 7 and 20 were challenge questions and were therefore not discussed in class.

However, the 4th triplet proved to be the most challenging and showed little learning. Therefore, the teacher chose to review the first question of the triplet (question number 14.) The teacher was able to identify the common mistakes, which were improperly subtracting the negative exponents as well as dividing the base. Because the next question had the poorest performance on the assignment, the teacher also chose to review question number 15 and highlight the importance of subtracting negative exponents carefully. Performance on this triplet suggests that feedback alone wasn't enough to cause learning. Teacher input and clarification was required.

We designed the experiment with ecological validity in mind. That is to say, we wanted the teacher to naturally review the homework, giving students enough time to ask questions. The hope was that approximately the same amount of time would be spent in each class and by each condition. We were disappointed to find that the classes and conditions varied greatly in the amount of time spent going over the homework. Half of the sections took over nine minutes to review the homework while two of the sections in the TH condition and one in the WBH condition spent substantially less time. This is a threat to the validity of drawing statistical inferences, but given the desire to maintain realistic homework review conditions, these inconsistencies highlight important differences in the homework review methods. We describe these differences in the following sections.

An observational analysis of the video recordings of the teacher reviewing the homework revealed that while the time spent in the WBH condition was often longer than the TH, it was also far more focused than in the TH. Specifically, when students were in the TH condition, on average 1 minute passed before any meaningful discussion took place. Whereas, when students were in the WBH condition, homework review began immediately with the teacher reviewing what she perceived to be the most important learning opportunities.

Figure 3: Video of homework review for experimental condition. To watch the full video, go to: http://www.youtube.com/watch?feature=player_embedded&v=Jb6Szy4fZ2w

Other notable differences in the type of review include the number of questions answered. In the TH condition, 2 classes saw 3 questions each and one saw 7. However, in the WBH condition each class saw 4 targeted questions and 2 classes requested 1 additional question. The variation in question types also is important to note. The teacher was able to ensure that a variety of question types and mistakes were addressed whereas in the TH condition students tended to ask the same types of questions or even the same exact question that was already reviewed. Additionally, students in the TH condition also asked more general questions like "I think I may have gotten some of the multiplying ones wrong." In one TH condition only multiplication questions were addressed when clearly division was also a weakness and similarly, another TH condition only asked questions about division. This accounts for much of the variability in overall review time.

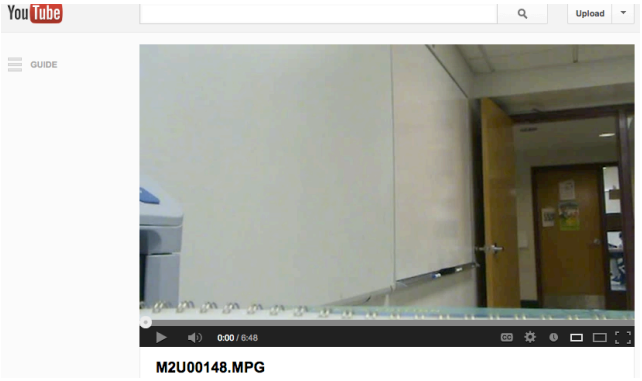


Figure 4: Video of homework review for the control condition. To watch the video go to: http://www.youtube.com/watch?feature=player_embedded&v=tBhcuCnKVCY

In listening to the comments made by students it appears that the discussion in the TH condition was not as structured as the WBH condition. Not all students had their work and therefore couldn't participate in the review. One student said, "I forgot to write it down." Another said, "I left my work at home." Because students were asking questions and the teacher was answering them, we suspect that only the student who asked the question was truly engaged. In fact, one student said, "I was still checking and couldn't hear" which led to the teacher reviewing a question twice. In the WBH condition, the teacher used the information in the report, such as percent correct and common wrong answers to engage the entire class in a discussion around misconceptions and the essential concepts from the previous question.

Other notable differences include the completeness of the review. In the TH condition, the review was dominated by student directed questions. This means that each class experienced a different review and the quality of that review was directly dependent on the engagement of the students. Conversely, in the WBH condition, all 3 classes were presented with the same 4 troublesome questions and common mistakes. Additional questions were reviewed when asked (as in two sections) but the essential questions as determined by the data in the item report were covered in all three sections.

3.4 Student Survey Results

Following participation in this study, students were questioned about their opinions. We want to acknowledge that students might have been telling the teacher what she wanted to hear: the whole classroom of students had been using ASSISTments for months and the teacher had told them on multiple occasions why it's good for them to get immediate feedback. So with that caveat, we share the following results. 86% of students answered ASSISTments to the question "Do you prefer to do your homework on ASSISTments or a worksheet?". 66% mistakenly think that it takes longer to complete their homework when using ASSISTments (we showed in this study that that was not the case) and 44% feel that they get frustrated when using ASSISTments to complete their homework. However 73% say that their time is better spent using ASSISTments for their homework than a worksheet. When asked what students like best about ASSISTments, student responses included:

"Being able to try again."

"That if you get stuck on a problem that it will give you the answer."

"You can redo your answer if you get it wrong and learn from your mistakes."

"How it tells you immediately that you are right or wrong."

"I like how I know if I'm right or wrong. This helps because often times when I get things wrong I just go back to my work and I see what I'm doing wrong which helps me when doing other problems."

"I like knowing if your right or wrong. it helps me learn from my mistakes because it makes me go back and keep trying until I get it right. I cant just move on when I feel like it. normally I would just try it a 1st time, and not go back and check, but assistmsnt makes me double Check my work."

"My favorite thing about ASSISTments is that it will tell you if you get the question wrong. PS--it doesn't help when it just says you get it wrong, it's helpful to see the steps so you can compare it to what your answer looked like."

"I like that you can tell what you did wrong and learn from it. That's it though. otherwise I would prefer a wkst [worksheet]."

"I like how it is online and easy to access."

While the learning benefits are profound and students prefer a web-based system, there is a sense of frustration that must still be addressed. Specifically, when asked what should be changed about ASSISTments, student responses included:

"I would make the hint button give a hint and not just the answer."

"I would make it so the hints maybe give you another example or helpful information so instead of just getting the answer and not knowing how you got it you could actually learn from it."

"If you get it wrong more than 4 times you have to move on to the next question."

"I would change how long it takes you to type it in. it would be cool if you could just say the answer and it would enter it in. that probably won't happen, but it would be awesome."

"I would change it to having hints to tell you if you have a little mistake when you hit submit answer so you don't get it wrong because of that little mistake."

This feedback suggests that students appreciate the features of intelligent tutoring systems, including hints, worked examples and scaffolding. Therefore, future studies should explore adding additional feedback to determine if added AIED features improve learning or if maybe learning requires some levels of frustration. All of the survey results are made available without names, including students' comments at http://www.webcitation.org/6DzciCGXm.

4. DISCUSSION

This papers' contribution to the literature is exploring the potential use of ITS for homework support. Used as designed, ITS are somewhat cumbersome for teachers to use for homework as the content is not customizable. However, if ITS are simplified they could be used like web-based homework systems, providing correctness feedback to students and reports to teachers. This begs the question, is correctness only feedback enough to improve the efficacy of homework and what effect does teacher access to reports have on homework review? This randomized controlled study suggests that simple correctness-only feedback for homework substantially improves learning from homework. The benefit of teachers having the data to do a more effective homework review was in the expected direction (but not reliable). But taken together (immediate feedback at night and an arguably smarter homework review driven by the data) the effect size of 0.56 seems much closer to the effect of complex ITS. Of course the large 95% confidence interval of [0.07 to 1.08] tells us we need more studies.

Future studies can explore features of other web-based homework systems like Kahn Academy to determine which aspects of the systems are particularly effective. Incrementally adding tutoring features to determine the effectiveness of each feature would also be valuable. Finally, the role of data in formative assessment should be further explored. In what way can teachers use the data to improve homework and review and instruction?

Caveats: the participants in the current study were all advanced middle school students. Therefore it would be necessary to replicate this study across a broader range of student abilities to determine if these effects are generalizable. Additionally, the correctness feedback is confounded with the unlimited attempts provided on the homework assignment. Therefore, it would be interesting to see if it's simply the correctness feedback that contributes to learning or if the impact stems from the unlimited attempts to self-correct. Finally, to address the secondary research question of the effectiveness of using that data and item report to enhance homework review, a more complicated research design would be required. Specifically, in the present study, the effect of the homework review was confounded with already improved learning that resulted from having correctness feedback. A two-by-two design where both immediate feedback and the factor of going over the homework with the data varies would be necessary.

In this fast-paced educational world, it is important to ensure that time spent in class and on homework is as beneficial as possible. This study provides some strong evidence that web-based homework systems that provides correctness-only feedback are useful tools to improve learning without additional time.

5. ACKNOWLEDGMENTS

The authors would like to acknowledge support from the Bill and Melinda Foundation via EDUCAUSE as well as IES grants R305C100024 and R305A120125.

6. REFERENCES

[1] Black, P., & Wiliam, D. (2006). Inside the black box: Raising standards through classroom assessment. Granada Learning.

[2] CEM (2013). Accessed 1/28/13 at http://www.cemcentre.org/evidence-based-education/effect-size-calculator.

[3] Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does Homework Improve Academic Achievement? A Synthesis of Research, 1987–2003. Review of Educational Research. Spring 2006, Vol. 76, No. 1, pp. 1–62.

[4] Kelly, K. (2012). Study Materials http://www.webcitation.org/6E03PhjrP. To browse, see http://web.cs.wpi.edu/~nth/PublicScienceArchive/Kelly.htm.

[5] Kulik, C. C., & Kulik, J. A. (1991). Effectiveness of computer-based instruction: An updated analysis. Computers in Human Behavior, 7, 75–94.

[6] Rochelle (2013). The IES Grant. Accessed 1/28/13 at http://ies.ed.gov/funding/grantsearch/details.asp?ID=1273.

[7] Schneider, S (2012). Accessed 1/28/13 at http://www.iesmathcenter.org/home/index.php.

[8] Shute, V. (2008). Focus on Formative Feedback. Review of Educational Research, 78(1), 153 -189. http://www.ets.org/Media/Research/pdf/RR-07-11.pdf

[9] Singh, R., Saleem, M., Pradhan, P., Heffernan, C., Heffernan, N., Razzaq, L. Dailey, M. O'Connor, C. & Mulchay, C. (2011). Feedback during Web-Based Homework: The Role of Hints In Biswas et al (Eds). Proceedings of the Artificial Intelligence in Education Conference 2011. Springer. LNAI 6738, Pages. 328–336.

[10] VanLehn, Kurt (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197-221.

[11] VanLehn, K., Lynch, C., Schulze, K. Shapiro, J. A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., & Wintersgill, M. (2005). The Andes physics tutoring system: Lessons Learned. In *International Journal of Artificial Intelligence and Education*, 15 (3), 1-47

Target the controls during the problem solving activity, a process to produce adapted epistemic feedbacks in ill-defined domains.

The case of a TEL system for orthopaedic surgery

Vanda Luengo
Université Joseph Fourier
Grenoble
Vanda.Luengo@imag.fr

Dima Mufti-Alchawafa
Université Joseph Fourier
Grenoble
dima.mufti@gmail.com

ABSTRACT

In this paper we study one feedback process which is adapted to ill-defined domains. Indeed, this process use others aspects than expected solutions to propose a feedback. The feedback process is based in a set of didactical aspects. In particular, the feedback targets the control element of knowledge, i.e. the knowledge that allows to validate one step in the problem solving process. The paper describes the feedback process and its implementation in the framework of one TEL system in orthopedic surgery.

Keywords

Control knowledge, feedback process, ill defined domain, and didactical decision.

1. INTRODUCTION

In ill defined domain one of the challenges is to continue to develop new tutoring strategies and seek out ways to combine existing strategies [13]. This challenge still open in particular when the domain has multiple and controversial solutions or ill-defined task structures [4]. In this framework our research question is how to design a tutoring feedback system which is not only based in defined solutions but in the known characteristics of knowledge and learning situations.

We study one kind of feedback which is adapted and epistemic. It is adapted because it takes into account the individual differences in relation to incoming knowledge and skills among students [18]. It is epistemic because it is specific to the piece of knowledge at stake and its learning characteristics. Compute an epistemic feedback involves knowledge from the learner, the learning situation and the learning domain [11].

We design a process to produce adapted epistemic feedbacks which includes one decisional model based in a set of didactical hypothesis. The process was implemented and tested in the case of orthopedic surgery.

The research discussed in this paper is developed in the framework of the TELEOS[1] platform [9] which is a Technological Enhanced Learning environment for orthopaedic surgery. This platform proposes a set of resources for the student (haptic simulator, online course, clinical case database) and a diagnosis system able to analyse the student productions and make a knowledge diagnosis based in identified controls.

[1] http://teleos .imag.fr

Based in the model presented in this paper we add a feedback system in the TELEOS environment. This implementation proposes a formative feedback which is delayed, i.e. at the end of the exercise in the simulator. The model is presented in the section 4 and the TELEOS example is presented in the section 5.

2. RELATED WORKS

In some domains (like percutaneous screw fixations in orthopaedic surgery) the *knowledge* obtained by experience plays an important role for both the expert teacher and the novice learner during a problem-solving process. This kind of knowledge, often tacit, refers to "work-related, practical know-how that typically is acquired informally as a result of on-the-job experience, as opposed to formal instruction." [22]. This kind of knowledge is pragmatic, obtained by experience. Moreover a skillful learner, even a domain expert, often makes several attempts before arriving at an acceptable solution: the person makes an error and then tries to correct the error several times. Also there are multiple solutions and because some parts of the knowledge are tacit the strategic to a good solution are unclear. This kind of problem is ill structured. Indeed, an ill-structured problem as one that is complex, with indefinite starting points, multiple and arguable solutions, or unclear strategies for finding solutions [19].

Several works address the problem to model ill defined knowledge and build feedback from these models ([13] and [20]). Based in this previous works, Fournier-Vigier et al. [5] pointed the design feedback difficulties in ill defined domains, in particular the difficulties to provide domain knowledge in ill structure problems. All studied paradigms (cognitive task analysis, constraint-based modeling, expert system, data mining algorithms) propose to describe task models using different techniques. The task models could be complete or partial. In all cases the model is used to offer assistance to the learner (ibid. 234). Most of the feedback systems in these approaches try to guide the student to the intended solution, even if it is described partially and beside most of the feedback are goal oriented.

We aim to study a model of feedback that is not only based in calculated solutions. We explore another feedback paradigm which is centered in the validation process more than the attended solution. In others words the feedback will be related to the characteristics of the controls brought into play during the problem solving process: it was brought into play in the right moment? It was valid or invalid? What is its nature ?

We would like to investigate how to produce an adapted epistemic feedback that takes into account these knowledge characteristics and is able to handle the uncertainty coming from

the diagnosis results. Indeed, like more and more intelligent tutoring systems, we chose to use Bayesian network for our diagnostic knowledge.

From adaptive point of view, Shute & Zapata-Rivera [18] propose a four-process adaptive cycle connecting the learner to appropriate educational materials and resources. This four process cycle include (ibid. p 9) *capture* of the information about the learner, *analyze* the information in relation to the learner model, *select* the information for a particular learner and *present* specific content to the student.

In relation to the selection step of the feedback, few systems propose a computer model which describes the decision of a pedagogical feedback following an uncertain diagnosis. Mayo and Metrovic [14] introduce the idea of Pedagogical Action Selection (***PAS***) and identified three general approaches to produce them in intelligent tutoring systems that use Bayesian networks: alternative strategies, diagnostic strategies, and decision-theoretic pedagogical strategies (ibid., p 132).

For us a didactical decision is to propose the best feedback and depending on the diagnosis results. This decision means a choice between different possible hypotheses based on didactical analysis. We use a decision-theoretic approach in order to model this process. The decision-theoretic strategy is used in some ITSs to select tutorial actions that maximize the expected utility. The systems CAPIT [14] and DT tutor [16] use this strategy.

CAPIT is a system for learning capitalization and punctuation in English. To decide two kinds of next feedback (next problem selection, error message selection) this system uses the utility function, which is based on the number of errors that the student made [14]. DT tutor also uses a decisional model: "*For each tutorial action alternative, the tutor computes (1) the probability of every possible outcome of that tutorial action, (2) the utility of each possible outcome in relation to the tutor's objectives, and then (3) the alternative's expected utility by weighing the utility of each possible outcome by the probability that it will occur. The tutor then selects the action with maximum expected utility with utility function*" [16]. In DT tutor, many factors related to the student (their morale, behaviour, etc) have an influence on expected utility. To propose the next feedback, DT tutor chooses first the theme where the feedback is focused and second the type of feedback (help, hint, positive or negative feedback). DT tutor is implemented in two learning systems, calculus and elementary reading.

A further difference between these previous works and our approach is that the decision feedback models proposed previously are not based on the nature of the control knowledge; in our case we would like to center the feedback on the knowledge control dimension (knowledge that allows the users to validate their actions during the process) and to take into account the knowledge control specificities (pragmatic, declarative and perceptive-gestural). Another difference is that, in our learning environment, there are no well defined solutions and thus it is not possible to define a priori, a list of actions as expected feedback. Because we have some characterised resources in our environment, the feedback is built in several steps; it has a target, an objective, a form and content. It is created with a decision-making process based on several PAS (Pedagogical Action Selection). In each step of the process, the chosen strategy corresponds to the degree of dependency of the step in relation to the domain knowledge.

Finally the factors considered in our system must be the parameters that can be established by researcher. Indeed, this is multidisciplinary research that evolves and the system must adapt to the evolution of didactic and medical analysis. Different feedback hypotheses must be able to be tested.

3. THEORETICAL FRAMEWORK AND DIDACTICAL HYPOTHESIS

According to research in cognitive psychology and didactics, the learner/situation interaction can be modelled as a problem-solving process that engages itself different processes, tightly linked and recurrent: identification of the situation, planning, action, control of actions' effects, regulation. The crucial role of control elements in this process has been pointed ([1],[17]), allowing the subject to decide whether an action is relevant or not, or to decide that a problem or sub-problem is solved.

This framework has some important consequences on our work for our objectives related to the design of a feedback system:

- It is necessary to *choose characteristics of problems* that will conduct to the right processes of learning according to professional objectives and to learner's state of knowledge. This, in turn, leads to the necessity to diagnose learner's knowledge, and interpret this diagnosis to be able to provoke targeted learning through learners' actions and controls on problems. Thus, one objective of *the feedback system is to consider* is not only the actions but also *the controls brought into play by the learner during the problem solving activity.*
- It is necessary to *distinguish* and consider both, *the result* (a punctual state of the problem, intermediate or final) *and the problem solving process*. We thus adopt a continuous approach of diagnosis and learning process.

Besides, we argue that is necessary to distinguish the controls characteristics. These categories are related to the way that knowledge can be validated, and therefore, built. In the case of orthopedic surgery we identify three categories: declarative, pragmatic and perceptive/gestural. The first category, declarative knowledge, corresponds to shared knowledge, constituting a common reference for professionals. It can be expressed, formally, and serves communication, discussion, exchanges. The second, pragmatic, is partly expressible, and is linked to individual experiences and situations. The third concern the perceptive and gestural (technical gesture like surgical gesture) part, hardly expressible and embedded in particular situations.

These are not the same that the classical division of knowledge between declarative and procedural. For example, part of procedural knowledge is validated in a declarative manner (is a reference for professionals and transmitted in a declarative way), part is validated in a pragmatic manner (by experience) and can also be validated in a perceptive-gestural manner (what is seen, felt). This second kind of activity is ill defined task, i.e. there are not clear strategies for finding solutions at each step of the problem solving process.

3.1 Characterization of didactical hypothesis' factors

Based in previous framework our objective is to propose a feedback system able to take into account the didactical hypothesis.

First of all and as explained above, each control element of knowledge is labelled according to its nature: declarative,

pragmatic, or perceptual/gestural. Then, concerning knowledge related to the user's action, it is labelled according to the moment it appears in the resolution process and according to its possible validity.

This last element necessitates some clarification: knowledge elements are diagnosed by the environment, according to user's set of actions and knowledge domain of validity, as being mobilized (brought into play) in a valid situation state (inside its validity domain), not mobilized or mobilized in an invalid situation state (outside its validity domain).

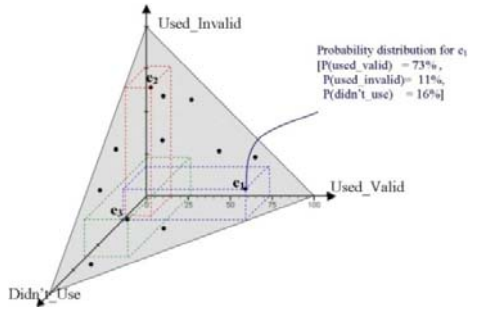


Figure 1. Result of knowledge elements diagnosed

The output can be considered like a tri-dimensional space (shown in Figure 1), where each knowledge element (e_i), in our case controls, has a probability distribution according to their state (invalid, valid, or not-used). In the given example, the knowledge element e_1 is brought into play in a valid manner with a probability of 73%.

Based in this result we made choices concerning the best relevant type of feedback to be provided to the user, according to previous diagnosed elements.

Thus, to produce epistemic feedback, the didactical analysis is based on the characteristics (state, order, type, etc.) of the control knowledge element and the classes of situations available. Also, to integrate the adapted dimension the feedback process has to take into account the student knowledge (the diagnosis result) and the characteristics of the learning environment (resources manipulated by the student and the characteristics of the problem).

4. THE PROCESS TO PRODUCE AN ADAPTIVE EPISTEMIC FEEDBACK

This process has four related steps. First, our decision model chooses the knowledge element(s), proposed by the diagnosis system, which will target the feedback. Second, it determines the feedback's apprenticeship objective for the chosen target. Third, according to the target and the objective, it determines the relevant form of feedback from the existing forms in the learning environment. Finally, according to the form, the decision model formulates the feedback by defining its content. The process is conceived from objectives and didactical hypothesis, summed up in §3, which are represented like parameters in the system.

In the next paragraphs, we describe each step in detail.

4.1 Chose the target of the feedback

This step can be shown as the selection of knowledge elements which are target by the feedback. The selection is influenced essentially by the knowledge diagnosis results and the controls' characteristics. In our case the knowledge elements are the controls which are brought into play during the problem solving activity. At each student action a list of controls were diagnosed. The results of one step can be seen like in the Figure 1. This diagnosis system is described in Chieu et al. [4].

We use influence diagrams to represent this step of decision. It is used to represent and to calculate the decision-making in several applications [6], [7]. In the influence diagrams there are decision nodes and utility nodes as well as chance nodes.

We have chosen this approach because it allows making decisions under uncertainty. Indeed, the learner's state of knowledge, produced by the diagnosis, will be deduced from learner actions with a degree of uncertainty, so our model has to generate the best feedback according to this input.

In our model (Figure 2) there are knowledge nodes (the oval nodes that represent the result of the diagnosis), an apprenticeship utility node (hexagonal node) and target decision node (rectangular node with the list of candidate elements or knowledge to be targeted). The inference in this diagram allows selecting a knowledge element as target. Indeed, the result of the inference gives the values of the utilities for each knowledge element, the highest one will be the targeted element for the feeback.

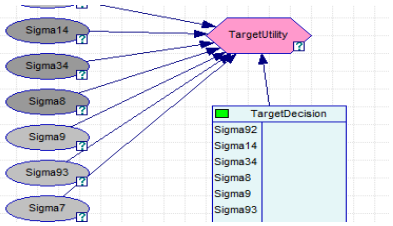


Figure 2. The influence diagram for target decision

To apply the inference in the diagram, we defined a function that models the preferences from an apprenticeship point of view, which is the utility function. The preferences will be described numerically under the notion of utility U, where U(a1)> U(a2) means the decision-maker prefer action a2 compared to the action a1.

In our case the apprenticeship utility function, Uapp(c, E), allows us to calculate the a priori utility to focus feedback on an element knowledge of a candidate (c) by taking into account the set of knowledge elements (E). Then, the inference in the influence diagram calculates the estimated utility for each candidate according to the diagnosis results. In other words, the utility function initializes the calculation in the influence diagram and then the inference algorithm deduces the decisions.

As we can see, in the previous figure the diagram is very simple; our contribution is basically in the definition of the apprenticeship utility function that takes into account the didactical hypothesis, which we explain in the next paragraphs

4.1.1 Apprenticeship utility function

This utility function allows initializing a priori utilities according to the factors that influence the target decision. We identified some factors as the element state and the element characteristic:

1. Element State is the diagnosis result. It represents the manner of using the knowledge element in the problem-solving process: Used-valid, Used-invalid, not-used.
2. Element Type, it is linked to the validation criteria for each identified knowledge, like explained after, in our

current Teleos example it can be "declarative", "pragmatic" or "perceptive-gestural";

3. Element Order represents the step of problem-solving in which this element intervenes. An element can intervene in several problem-solving steps, for example the control knowledge related to the profile X-ray can intervene in several steps of the activity;
4. Element Context indicates the context of problem-solving in which this element intervenes. It can be 'general' or 'particular'. For example, in the case of surgical domain, some steps and knowledge control elements could be especially for the scoliosis intervention.

From all of these factors we define in the equation (1) $U_{app}(c, E)$ the utility to choose a candidate element, ***c,*** as feedback target in taking into account the set of knowledge elements, ***E***, as the sum of all the utilities related to each factor.

$$U_{app}(c,E) = \alpha \cdot U_{state}(c,E) + \beta \cdot U_{Type}(c) + \gamma \cdot U_{order}(c) + \delta \cdot U_{context}(c) \quad (1)$$

In our didactical hypotheses, these factors do not have the same weight in influencing the choice of the target. Thus, we attribute to each factor a priority variable (α, β, γ, and δ), which represents its weight in the utility calculation.

We define in the equation (2) the utility of choosing a candidate ***c*** as a target according to its state $U_{state}(c, E)$, as the sum of utilities for each pair of candidates ***c*** and element $\boldsymbol{e_j}$ in ***E***; n is the number of knowledge elements of the set E.

$$U_{state}(c, e_1, e_2, ..., e_n) = \sum_{j=1}^{n} U_{State}(c, e_j) \quad (2)$$

In addition, we define the state utility in the table for each pair $U_{state}(c, e_j)$. The values are defined according to didactical hypotheses and the domain of knowledge.

For example, the didactical hypothesis "*it is more important to focus the feedback on an element that is used in an invalid way than to focus it on an element that it didn't use"* is represented by a value where Ustate(c = "used_valid", e) ≥ Ustate(c = "not-used", e). In other words, we propose one utility state table that allows selecting between two elements situated in the diagnosis results space (shown in Figure 1) according to the chosen didactical hypothesis.

The definition of the type utility $U_{type}(c)$ from didactical hypothesis can be "*it is more important to focus the feedback on a declarative element than to focus it on a pragmatic one*". We express this by giving to declarative elements the higher value of utility. In this example, the $U_{type}(c) = 3$ if c is declarative and 2 if it is pragmatic. In the present implemented version, the system doesn't take into account the perceptive-gestural knowledge because the didactical analysis is ongoing, but it is modelled to integrate it in an easy and modular manner.

We define the utility order: Uorder(c), from the didactical hypothesis "it is more important to focus the feedback on an element appearing in a primary stage of the solving than to focus it on an element appearing in later stages". Thus, it is possible that an element appears in several stages. We define the utility order in equation (3); m is the number of steps where this element appears and O(c) is its order. The first time of the control *i* is identified Oi(c) = 1.

$$U_{order}(c) = \sum_{j=1}^{m} \frac{1}{O_j(c)} \quad (3)$$

We define the nature utility $U_{nature}(c)$ from the didactical hypothesis as follows: "*it is more important to focus the feedback on an element appearing in the solving of a general problem than to focus it on an element appearing in a particular context*". Like the Utility type function case, we express this by giving a higher value of utility to the nature target chosen (in this case if c is general $U_{context}(c) = 2$).

According to these considerations, we have defined an algorithm that calculates the apprenticeship utility function and initializes the utility table from a set of knowledge elements with their characteristics. In this algorithm we create, first of all, the coefficients' matrix « *coeff* » in relation to the number of knowledge elements (k), and then we calculate the state utility table for each candidate. It is calculated based in formula 4, where *k* is the number of the column, j is the possible state of the knowledge element (used-valid, invalid or not-used) and *Hypo* is one of the didactic vectors A,B or C related to the state of the targeted candidate in column *k*

$$ValeurUtilitéEtat[k] = \sum_{j=1}^{3} Coeff[j,k] * Hypo[j] \quad (4)$$

This algorithm needs to be to running only once, after settle the didactical hypothesis. The inference in the influence diagram then uses probabilities resulting from the diagnosis and then calculates utility values to infer the estimated utility for each element. Finally, the target for the feedback is the element that has the maximal estimated utility value (Figure 3) calculated. It is possible to have some elements with the same maximal utility.

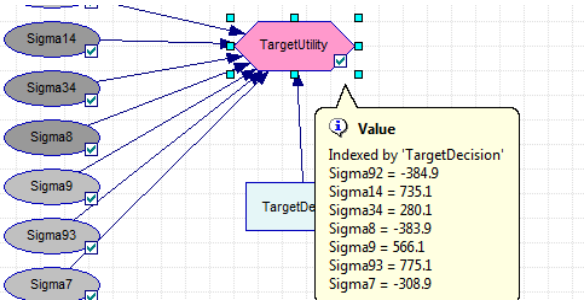


Figure 3. Inference Diagram decision result

As we presented before, we have chosen to represent all didactical hypotheses as parameters in the utility function. This choice makes our model flexible to add or modify didactical hypotheses. For example, for the factor "Type of Knowledge" if the didactical hypothesis is "*it is more important to focus the feedback on an pragmatic element than to focus it on a declarative one*", then it is sufficient to give the parameter that represents the utility for an pragmatic element a value higher than the utility for a declarative element $U_{type}(c=\text{pragmatic}) > U_{type}(c=\text{declarative})$.

4.2 Choose the objective of the feedback

After choosing the target, the decision model determines its feedback objective in order to give, from the learning point of view, a semantic to the feedback intention. In our model we distinguish several feedbacks. Indeed, if the target knowledge is diagnosed (with a higher probability) as 'brought into play in an

invalid manner' (BPI) the feedback is not the same than if this target knowledge is diagnosed as 'not brought into play' (NBP).

We have defined a procedure that determines the feedback objective by applying an analysis on the target element state. The principle of this procedure is that it segments the diagnosis space into several zones, and it attributes an objective to each zone. Then, the feedback objective corresponds to the zone in which the target element is situated. This step permits to pass from an uncertain state of knowledge to fixed objectives of learning. The number of segmented zones and the objective for each zone is customizable in our model.

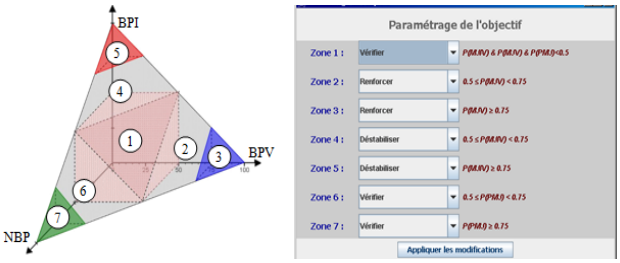


Figure 4. Example of segmentation of the knowledge elements space to determine the feedback objective

In this example, if the knowledge element is in zone 1 ("*if P(NBP) – P(BPI) > 0.25 and P(NBP) – P(BPV) > 0.25*") then the feedback objective is to "verify" if the targeted knowledge is understood by the learner. The possibilities proposed for the feedback objective are: verify, reinforce and destabilize. The meaning of *verify* feedback is to propose a type of feedback to improve the diagnosis related to a set of knowledge targeted elements (for example, proposing another problem where specific targeted knowledge has to be mobilized). The idea of the *reinforce* feedback is to support the user in relation to the targeted knowledge elements (for example a positive feedback, a closer clinical case that was studied or solve another problem where the targeted knowledge could be used). Finally, *destabilize* feedback has the objective to show that the targeted knowledge is used in an invalid manner in these kinds of problems (by explaining the right way in the course, proposing a counter example from the clinical case database or proposing another problem where if the knowledge is used, the result could be wrong)

4.3 Determine the feedback form

In this step, the decision model chooses the most relevant form of feedback linked to the type of the target knowledge element and the feedback objective (*reinforce, verify, destabilize*).

Here the idea is to associate one kind of feedback form to the feedback objective and the type of the targeted knowledge element. In this step we need to consider the resources proposed to the student. Indeed more and more TEL system proposes several resources to the student. For example if the environment has a wiki with concepts we can associate it to a form of feedback when the targeted knowledge element is declarative and the feedback objective is to reinforce.

This association is a simple table where we can match the resources with a pair <type of knowledge, feedback objective>.

4.4 Determine the feedback content

The content is essentially related to the form of feedback. Here the objective is to determine the content of the feedback in relation to the feedback form. For example if the feedback form is a wiki with concepts the content has to be related with the targeted knowledge element.

This step is not generic, it depends on the kind of feedback forms that the TEL system has. For this reason this step will be more detailed in the next section where we explain the case study where we implemented the feedback process.

5. THE TELEOS SYSTEM EXAMPLE

The analyzed procedure is about surgical orthopaedic percutaneous (without incision) operation. It is developed in [21]. It could be summarized as follows: The surgeon first inserts a pin in the bone through the skin. S/he makes the pin progress in the bone, taking several X-rays to validate the pin's course at different steps of its progression. The X-rays allow him or her to "reconstruct" a complete vision of the position of the pin, in relation to the bone. If s/he recognizes any problems in those views, s/he restarts the operation process, taking another pin and correcting its entry point and/or direction. Until now we have analyzed the sacroiliac screw operation and the vertebroplasty. The description procedure does not have to be complete and well-defined but the goal is to extract from the diversity of each particular situation, the significant controls elements, from a learning point of view, of the surgical procedure.

The analysis, made in [21], allows us to identify crucial aspects of the surgical procedure. We identified primarily that the pin's positioning is the most important part of the procedure, the definitive screw being placed along this pin. Secondly, we notice the crucial role of X-ray controls. As the surgeon cannot directly visualize the operating area, he has to interpret his gesture through these controls. This necessitates two levels of interpretation. On the first level, the surgeon has to ensure that the X-ray is valid (i.e. being oriented in order to represent what it is intended to represent); on the second level, the surgeon can look at the validity of the pin's position according to anatomical criteria on the X-ray.

Table 1. Examples of knowledge controls for sacroiliac screw

Control Type	Control elements of knowledge	Domains of validity
declarative	The pin's trajectory must be completely intra-osseous	all
declarative	If the pin is well positioned then the pin appears as a point on the profile X-ray	PB, PC, PE
Pragmatic	If the pin would become extra osseous by being pushed in S1, 1cm after the median line, then it can be stopped at the median line	PC, PD
Pragmatic	If the pin would become extra osseous, then it can be stopped just 1cm after having reached S1	PA,PD,PF
Perceptive-gestural	If the pin was in the sacroiliac and the resistance force decrease then the pin would become extra osseous	All

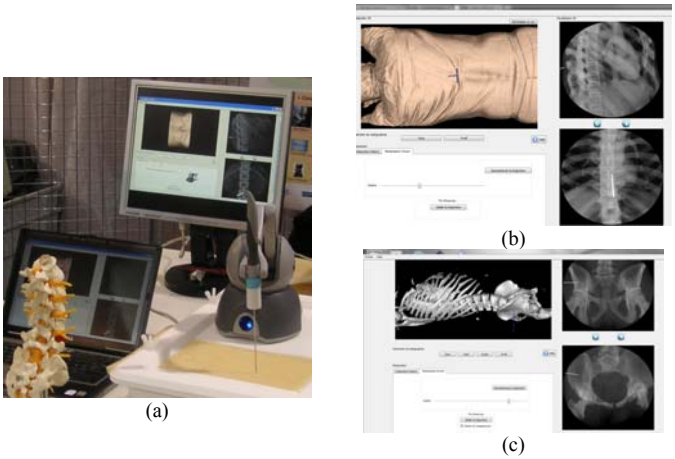
Thus, we identified the control knowledge elements, which are related to surgeons' actions during the intervention, they allow surgeons to validate their actions; some examples are shown in Table 1. The controls have a domain if validity, i.e. they are valid for a set of problems. The control type is also identified: it could be declarative, pragmatic, or perceptive gestural.

5.1 TELEOS SYSTEM

We have developed a modular architecture. Each module is built in relation to the knowledge learning constraints [10]. The learner interacts with the following modules: Semantic Web Courses, Simulator, and Clinical Cases. We introduce briefly these three modules in the next section. The decision-making model uses these modules and the result of the diagnosis to build the feedback. The diagnosis model will not be described in this paper. The result will show in the Figure 1.

5.1.1 Simulator for orthopaedic surgery

The last implementation version is explained in a previous paper [12]. Two surgeries were implemented in this last version: the vertebroplasty and the sacroiliac screw.



(a) (b) (c)

Figure 5. Haptic interface (a). Graphical interface during the pin trajectory (b). Graphical interface when the trajectory is validated by the user (c).

Regardless of the simulated operation, the TEL system gives to the learner the opportunity to train himself and practise a surgical operation thanks to several functionalities: Choosing the type of patient and the type of operation; visualizing in 3D the tool and the patient model; Adjusting the position and the incidence of the fluoroscopic image intensifier; Drawing the cutaneous marking on the body of the patient model; Producing and visualizing radiographies; Manipulating the surgical tool through a mouse or through haptic interface; Verifying the trajectory in a 3D bone model when it has been validated (Figure 5). In this paper we are focused on the pelvis operation.

In the previous figure we can see on the right of the graphical interfaces (b and c), two 2D images representing the last two radiographies produced by the user. In the top left hand corner, there is the 3D model of the patient, and the surgical tool, the user is able to see the 3D bone model only when the trajectory is validated.

5.1.2 Clinical cases database

The role of the Clinical Case agent is to illustrate the consequences of a proposed trajectory. It is a database where we can find pertinent information related to different phases (before, during and after the operation).

Figure 6. Clinical Case with data from one operation

For example, one clinical case may have some x-rays before the operation (**Figure 6**, right side), some films of the gesture during the operation and some x-rays and data describing the post-operatory information (the position of the bone, the state of the bone, etc… left side **Figure 6**). This Clinical Case Database could be useful to show, for example, trajectories that have consequences in the post-operatory period (there may be a problem with the fracture reduction because the trajectory with the pin is too short, for Instance).

5.1.3 Online Courses

We have an online course (at http://www-sante.ujf-grenoble.fr) that explains the declarative knowledge (anatomy, surgical procedure, tools, etc.) about sacroiliac percutaneous screw placement. It is based on online courses and academic documentation, and is improved by interaction between the didactical expert and the surgeons.

For this part we use ontology with a set of rules based in OWL language. We have developed a semantic web module, with more than eighteen web pages, which have metadata based on ontology. This module proposes not only syntactic links, but also semantic ones; it allows the redirection to precise and relevant chapters of the online course. The implementation of this module is explained in previous work [8].

5.2 ADAPTIVE AND EPISTEMIC FEEDBACK PROCESS

Like introduce in the paper the implemented feedback process is a delayed feedback, i.e. the TELEOS system propose a feedback at the end of the activity. The result of the process can be to solve another problem on the simulator, to consult a particular webpage on the online course or to consult one specific clinical case in the database.

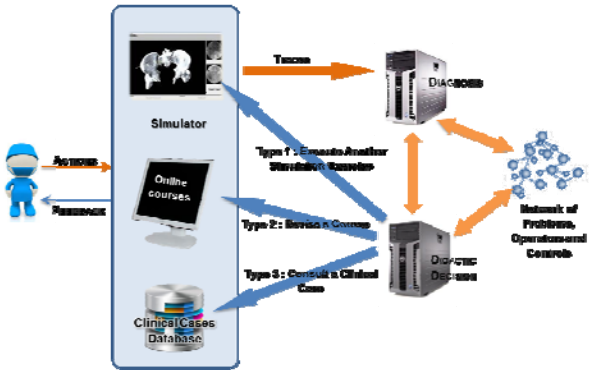


Figure 7. Kinds of feedback in TELEOS system.

Because the two first steps described previously are generic, we don't explain them in detail here. In the step three we propose a simple table interface where the didactical or pedagogical user can propose the match between the resource (simulator, clinical

case and web course) and the pair type of knowledge (declarative, pragmatic, perceptivo-gestural) and feedback objective ((*reinforce, verify, destabilize*). We can choose one or several forms for the same pair <type, objective>. For example, the pedagogical user was proposed the clinical case and the simulator to destabilize the pragmatic knowledge.

For the step 4 we need to consider the specific form of the feedbacks. In our case we have three forms of feedback (online web course, simulator, and clinical database) and to find the inquired form we do not apply the same process. One example of possible feedback is shown in the next figure:

In the case of the form 'consult part of web courses', the content represents the links to the appropriate pages. It is made by sending keywords related to the target element to a semantic web model [8]. This feedback receives the knowledge elements to be considered, which are analyzed by the java program, using the ontology, and finally it produces a web page with a set of links to the online course, which are related to the targeted knowledge. The Java Engine code uses the open source tool Jena which offers libraries to work with OWL files. In the case of the sacro-illiac surgical operation, our system is based on two ontologies, one related to the pelvis anatomy, which is built on Standford university anatomy ontology [2], and the other one is related to the screw placement procedures, which we built and validated with our experts.

For example if we give the knowledge element 'Outlet radio control', which is in relation with anatomical ontology, the java engine finds the classes related to these knowledge elements and produces a set of links which come from the online course.

The calculation of the content for the forms 'clinical cases' and 'simulator' is made according to the target and to the feedback objective. For the form 'consult a clinical case', it represents the relevant case like a query in a database.

Finally, for the form 'solve another problem with the simulator', it represents the relevant problem to solve. The design can be made by applying inference algorithms in the Bayesian Network (that represents the knowledge domain) or by a decisional theoretical approach to select a closer problem [15]. In the present version we find the problem that has the most common didactical variables (kind of fracture, the hardness of the spongy bone) with the solved problem.

5.3 Evaluation and discussion

The evaluation of the didactical decision process was achieved in several steps. Because the utility function is additive, we evaluated first the dominance between different modelled factors and second we made a sensitive analysis to study the adaptability of the model. Moreover, we made an evaluation to study the behaviour of the system in relation to the expert's propositions. Here we present this last evaluation. The others evaluations show that, firstly, small changes in the assigned probabilities lead to different decisions of feedback target. It means that if there is one small change then the result of the calculus of the target feedback could be radically different. Secondly, the sensitivity level can be adjusted according to the weight given to the element state factor.

The aim of the comparison with expert proposition is to verify and refine the model in relation to the human didactical feedbacks. Here the input is the simulated diagnosis of learner's state of knowledge (e1 [BPI 0.7, BPV 0.17, NBP 0.13], e 92 [BPI 0.65, BPV 0.23, NBP 0.12], etc) and the output is the feedback proposed (Consult the parts of the course 'entry point related to skin marks', propose a problem, with a disjunction, to solve in the simulator, etc.). These scenarios are run by an expert in didactics and by our didactical decision system, afterwards they are compared.

Because in our model the didactical hypotheses are customizable, the parameters have to be calibrated by an expert (in didactics for example) before using it. To make the adjustment of these parameters easier, we developed some interfaces and we also proposed a questionnaire that contains multiple-choice questions, (associate to didactical hypothesis) and we associate with each choice a possible value of the parameter. Therefore, the answers to this questionnaire allow initializing the calculation in the model.

One example of scenario given to the experts is "after radio outlet, a student does not takes Inlet radio and modifies its trajectory in the wrong direction (the pin was placed a little low on the outlet, it starts and moves the point of entry down). The declarative control e93 (coupling outlet / inlet) comes NBP 30%, the declarative control e19 (risk of passing through the hole of the sacrum because too low on outlet) is BPV 50% and the pragmatic control e18 (link outlet position / position of patient) is 75% BPI". One expert proposition was: "propose the web page linked to the inlet/outlet coupling and propose an exercise related to the 2D and 3D association".

In relation to the configuration of the system, the answer of the questionnaire shows us a dependent relationship between the state of the knowledge elements and its characteristics while in our model these factors are independent (it is an additive function). For example, the question about what is more important to target a "not-valid knowledge" or a "not used knowledge", the expert answer depends on the type of the knowledge (declarative, pragmatic, etc.).

In addition, regarding the output proposed by the expert, the results show that the system is able to produce relevant feedbacks for each scenario. Furthermore, some feedbacks are not exactly the same as the expert feedbacks. We identify two reasons for these differences. Firstly, the present model selects as target one (or some) element(s) that has(have) the maximal value of estimated utilities but in the expert propositions, the feedbacks can be related to some elements with positive values of estimated utilities and related as well to the elements with the maximal value. Secondly, the present model is not able to propose a sequential set of feedbacks (for instance, the expert proposes that feedback 1 follows feedback 2). In fact, the present model is able to take the historical dimension with the evolution of the probabilities, but it does not yet treat the historical dimension related to the previous feedback

6. DISCUSSION

This system had to support an explicit representation of pedagogical and didactical hypotheses and, from a computer architecture point of view; the system had to be separated from the other modules. These choices are related to the idea of proposing a normative system, able to evaluate separately and also to allow the investigation of some didactical hypothesis to generate the feedback.

The decision model thus integrates didactical hypotheses in order to represent the decision-maker's preferences. These didactical hypotheses are customizable; this choice makes our model dynamic and partially generic. Also, this kind of model intends to

allow multidisciplinary work in order to investigate pedagogical feedback.

From the epistemic dimension of the feedback point of view, the system cannot be completely generic but the design allows identifying the generic steps from the knowledge analysis dependant steps.

In relation to the adaptive dimension of the feedback, the system is able to adapt the feedback to some epistemic considerations about the user and the available resources. Indeed, this adaptive dimension takes into account only the knowledge factors. It doesn't take into account other factors like the morale or attention. Also, as pointed out by Woolf ([23] p. 133), it is necessary to integrate different teachers' strategies: *A single teaching strategy was implemented within each tutor with the thought that this strategy was effective for all students. However, students learn at different rates and in different ways, and knowing which teaching strategy (…) is useful for which student would be helpful. This section suggests the need for multiple teaching strategies within a single tutor so that an appropriate strategy might be selected for a given student"*.

The reliability of our model depends on the accuracy of diagnosis results and the best set of parameters. Here it is also necessary to refine the model using real data in order to improve its structure, the conditional probability and the decision factors by using a method of automatic learning from data.

Moreover, the evaluation indicates that it seems necessary to consider not only the history of the student activity but also the dynamic aspect linked to the decisions. Indeed, in the classical approach the decision is in relation to the predictive aspect of the student model ([16], [2]) i.e. it calculates the consequences of the feedback on the predictive student model. However, it appears that the dynamic aspects concern not only the student factors but also the resources or the decision itself.

The data collection seems to be the perspective's keystone in order to improve the present model but also to go forward in this kind of research. However, the data to be collected it is not only the classical data in the domain of learning systems, i.e. the data from the student, but also the data linked to the feedback decision. This kind of collection will be more centred on the analysis of the decision process for the feedback production.

7. REFERENCES

[1] Balacheff N., Gaudin N. (2010) Modeling students' conceptions: The case of function. *Research in Collegiate Mathematics Education*, Volume 16, 183-211.

[2] Chieu, V., Luengo, V., Vadcard, L., & Tonetti, J. (2010). Student modeling in complex domains:Exploiting symbiosis between temporal Bayesian networks and fine-grained didactical analysis. *Journal of Artificial Intelligence in Education.*.

[3] Foundational Model of Anatomy, S. (2006). *Research Projects, Foundational Model of Anatomy*. Retrieved 2008, from http://www.smi.stanford.edu/ (Research Projects, Foundational Model of Anatomy).

[4] Fournier-Viger, P., Nkambou, R. & Mephu Nguifo, E. (2010). Building Intelligent Tutoring Systems for Ill-Defined Domains. In Nkambou, R., Mizoguchi, R. & Bourdeau, J. (Eds.). *Advances in Intelligent Tutoring Systems*, Springer, p.81-101.

[5] Fournier-Viger, P., Nkambou, R., Mayers, A., Mephu Nguifo, E & Faghihi, U. (2012). Multi-Paradigm Generation of Tutoring Feedback in Robotic Arm Manipulation Training. *Proceedings of the 11th Intern. Conf. on Intelligent Tutoring Systems* Springer, pp.233-242.

[6] Horvitz, E., Kadie, C., Paek, T., & Hovel, D. (2003). Models of attention in computing and communication: from principles to applications. *Commun. ACM* , 52-59.

[7] Kabanza, F., Bisson, G., Charneau, A., & Jang, T. (2006). Implementing tutoring strategies into a patient simulator for clinical reasoning learning. *Artificial Intelligence in Medicine* , 79-96.

[8] Luengo, V., & Vadcard, L. (2005). Design of adaptive feedback in a web educational system. *Workshop Adaptive Systems for Web-Based Education: Tools and Reusability*. In International Conference on Artificial Intelligence in Education. Springer-Verlag.

[9] Luengo, V., Vadcard, L., Dubois, M., & Mufti-Alchawafa, D. (2006). TELEOS : de l'analyse de l'activité professionnelle à la formalisation des connaissances pour un environnement d'apprentissage. *IC 2006.*

[10] Luengo V. (2008) Take into account knowledge constraints for TEL environments design in medical education. *ICALT* 2008.

[11] Luengo V. (2009). *Les rétroactions épistémiques dans les Environnements Informatiques pour l'Apprentissage Humain. Habilitation à diriger des recherches*. Habilitation à diriger des recherches. Université Joseph Fourier., 2009.

[12] Luengo V., Larcher A., Tonetti J.. (2011) Design and Implementation of a Visual and Haptic Simulator in a Platform for a TEL System in Percutaneuos Orthopedic Surgery. *Medicine Meets Virtual Reality* MMVR, 2011.

[13] Lynch, C., Ashley, K., Aleven, V., & Pinkwart, N. (2006). "Defining Ill-Defined Domains; A literature survey." *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 8th International Conference on Intelligent Tutoring Systems* (p. 1-10).

[14] Mayo, M., & Mitrovic, A. (2000). Using a Probabilistic Student Model to Control Problem Difficulty. *Intelligent Tutoring Systems* (pp. 524-533). Biarritz: Springer-Verlag.

[15] Muldner, K., & Conati, C. (2007). Evaluating a Decision-Theoretic Approach to Tailored Example Selection. *IJCAI*, (pp. 483-488).

[16] Murray, R., VanLehn, K., & Mostow, K. (2004). Looking ahead to select tutorial actions: A decision-theoretic approach". *International Journal of Artificial Intelligence and Education* , 235-278.

[17] Schoenfeld, A. (1985). *Mathematical Problem Solving*. New York, NY: Academic Press.

[18] Shute, V. J., & Zapata-Rivera, D. (2012). Adaptive educational systems. In P. Durlach (Ed.), *Adaptive technologies for training and education* (pp. 7-27). New York, NY:Cambridge University Press.

[19] Simon, H. A., "The Structure of Ill Structured Problems," *Artificial Intelligence* **4**, (1973), 181-201.

[20] Stamper, J., Barnes, T., and Croy, M. (2010) Enhancing the Automatic Generation of Hints with Expert Seeding. Proceeding of ITS2010. vol. II, pp. 31-40. Berlin, Germany: Springer Verlag.

[21] Tonetti J., Vadcard L., Girard P., Dubois M., Merloz P., Troccaz J. (2009). Assessment of a percutaneous iliosacral screw insertion simulator. *Orthopaedics & Traumatology: Surgery & Research, Vol 95* n°7, 471-477.

[22] Wagner, R., Sujan, H., Sujan, M., Rashotte, C., & Sternberg, R. (1999). Tacit Knowledge in Sales. In S. R. J., & H. J. A., *Tacit Knowledge in Professional Practice: Researcher and Practitioner Perspectives* (pp. 155–182). Mahwah, NJ: Lawre.

[23] Woolf, B. P. (2009). Building Intelligent Interactive tutors : student-centered strategies for revolutionizing e-learning. Burlington, MA, USA: Elsevier

Eliciting student explanations during tutorial dialogue for the purpose of providing formative feedback

Pamela Jordan
Learning Research and Development Center
University of Pittsburgh
Pittsburgh PA, USA, 15260
pjordan@pitt.edu

Patricia Albacete
Learning Research and Development Center
University of Pittsburgh
Pittsburgh PA, USA, 15260
palbacet@pitt.edu

Michael J. Ford
School of Education
University of Pittsburgh
Pittsburgh PA, USA, 15260
mjford@pitt.edu

Sandra Katz
Learning Research and Development Center
University of Pittsburgh
Pittsburgh PA, USA, 15260
katz@pitt.edu

Michael Lipschultz
Department of Computer Science
University of Pittsburgh
Pittsburgh PA, USA, 15260
mil28@pitt.edu

ABSTRACT

In this paper we explore the question of whether additional benefits can be derived from providing formative feedback on students' explanations given the difficulties of accurately assessing them automatically. We provide a preliminary evaluation of an approach in which students assist in interpreting their own explanations and we lay out our plans for evaluating the effectiveness of a natural-language intelligent tutoring system's feedback to that interpretation effort. The preliminary evaluation suggests that students respond well to the approach. While their interpretation assistance may be similar to an automated explanation matcher, they continue to provide explanations throughout their interactions.

Keywords

student explanations, tutorial dialogue, formative feedback

1. INTRODUCTION

Numerous studies suggest that self-explanations can be more beneficial to students than explanations from others (e.g. [3]). In the context of an automated learning environment this raises the question of whether additional benefit can be derived from providing formative feedback on any explanations the student enters when the automated understanding of those explanations remains a major obstacle. Must we be satisfied with the self-explanation effect or can and should we do more?

Previous work has attempted to recognize natural language explanations and then engage in a natural language dialogue with the student to refine and improve those explanations (e.g. [11]). And more recent work has attempted to field dialogue systems that incorporate more knowledge intensive automated recognition of students' elaborations during dialogue [4]. But so far, recognizing what the student meant is still very limited. And even if we step away from attempts at actual understanding, the performance for matching to canonical sets of answers is still relatively low (e.g. [5, 12]) compared to what can be achieved with short answer responses (e.g. [13]). Perhaps even more troubling is how sensitive students are to a system's failure to understand them [4]. Although a system can recover and move forward in a coherent manner, the students notice the lack of understanding. One possibility for this sensitivity may be that the errors are often quite different from those a human makes (e.g. the system fails to recognize a response as correct but a human clearly would).

Related work, which studied the impact of decisions about dialogue tactics [2], seems to have avoided some of these issues by substituting a human interpreter (wizard) for the automated interpreter. One goal of this substitution was to reduce the confounds of misunderstandings so that the system could focus on evaluating decision policies regarding whether to elicit or tell the explanations and justifications for statements made either by the system or the student. The human interpreter was presented with a list of canonical answers and was asked to find the best match for the student's response or to select "none of the above". There were significant differences in learning based just on varying decision policies about whether to elicit or tell the same content. This result suggests that being able to request explanations and justifications and being able to reduce the confounds of errors in matching to canonical answers has potential. But is there a practical way to include a human interpreter in a classroom setting? And how sensitive are students to problems that arise if their answer is close to correct but not a good match for any of the canonical answers?

First we will introduce the Rimac[1] system and its experimental setting and our approach for eliciting and assessing students' responses to requests for explanations/justifications. Next we will describe the data we have collected and provide a preliminary evaluation of the success of our approach for eliciting explanations/justifications. Finally, we will lay out our plans for exploring if there is value added to providing feedback on students' explanations.

2. THE RIMAC SYSTEM

Rimac is a natural-language intelligent tutoring system that engages students in dialogues that address physics concepts and principles, after they have solved quantitative physics problems. Much research has been devoted to identifying features of tutorial dialogue that can explain its effectiveness (e.g., [1]), so that these features can be simulated in natural-language tutoring systems. One hypothesis is that the highly interactive nature of tutoring itself promotes learning. Several studies indicate that our understanding of interactivity needs refinement because it cannot be defined simply by the amount of interaction nor the granularity of the interaction but must also take into consideration how well the interaction is carried out (e.g., [2]).

This need for refinement suggests that we should more closely examine the linguistic mechanisms evident in tutorial dialogue. Towards this end, we first identified which of a subset of co-constructed discourse relations correlate with learning and operationalized our findings with a set of nine decision rules which we implemented in Rimac [9]. To test for causality, we created two versions of Rimac: an experimental version that deliberately executes the nine decision rules within a Knowledge Construction Dialogue (KCD) framework, and a control KCD system that does not intentionally execute these rules. One of the rules in the experimental version of Rimac that is of particular relevance for this paper is: If the student gives no explanation for an answer to a complex question then ask for one.

Rimac was built using the TuTalk tutorial dialogue toolkit [6] but has been enhanced with additional dialogue features such as reformulation of student input (e.g., [7]). The dialogues are tutor-initiative only and are primarily short answer questions in order to keep the accuracy of automatic recognition high. However, the system does request student explanations at a few key points in the dialogues. It does not attempt automatic recognition of these responses but instead always follows-up with multiple choice answers for the explanation question and a request that the student select the best match for the explanation she just provided. In effect, we are asking the student to fulfill the role of a human interpreter for the system. We give the students brief verbal instructions for what is expected of them and why as we will explain in more detail below.

Figure 1 is an excerpt of an interaction between one of the students in our evaluation set and Rimac. This excerpt shows the system requesting an explanation at line 3. The student provides her response and then the system asks the student to select the best match for her response at line 5. The system then follows-up in lines 7-9 with formative feedback that is appropriate for what the student selected as the best match for her response.

1. T: We talked about the acceleration on the way up and on the way down, but what about at the very top of the jump? What was the direction of the acceleration at the top?
2. S: zero
3. T: Why do you think there wasn't any acceleration?
4. S: it had a zero velocity
5. T: Thanks! So what below is the best match for what you just entered? [menu choices: 'Since the kangaroo is not moving, there isn't any acceleration | The kangaroo is changing direction, so it doesn't have an acceleration | I guessed | None of the above']
6. S: Since the kangaroo is not moving, there isn't any acceleration
7. T: There actually was acceleration at the top of the jump, even though the kangaroo wasn't moving.
8. T: The kangaroo wasn't moving for just a very brief moment and it was like that because its velocity was switching from upward to downward. To switch from an upward velocity to a downward velocity, the kangaroo's velocity must reach a point where it is 0 m/s.
9. T: Since acceleration represents the change in velocity and the velocity was changing in this situation, the kangaroo's acceleration must be non-zero.

Figure 1: Excerpt of actual student interaction that requests student's assistance with interpretation along with the resulting formative feedback

3. THE EXPERIMENTAL DESIGN

Students in five Pittsburgh area high schools interacted with one of the two versions of Rimac during two course units (kinematics and dynamics). They used the system for one to two class periods per unit. In this paper, we examine the dialogues from the kinematics unit only.

A day or two prior to using the system, students first took a pre-test, and then completed a homework assignment in which they solved four quantitative physics problems. In a subsequent class, they used the Rimac system and finally during the next class meeting took a post-test.

Just before students began using Rimac, we introduced them to the system and read the following to them regarding requests for explanations:

> "Sometimes it will ask you to explain your response. This is regardless of whether it thinks you were right or wrong.

[1]Rimac is the name of a river whose source is in the Andes. Its name is a Quechua word meaning "talking;" hence the nickname for Rimac: "talking river." We thus considered Rimac to be well suited to a dialogue system that we intended to embed in the Andes tutoring system.

When it asks you to explain, please be sure to type in what you were thinking that lead you to your answer. You may have to think a bit about it. If you realize that you guessed or used your intuition, that's fine; just type that.

It will then follow-up with a multiple choice question and ask you to pick what is the best match for what you just wrote. It is important that you pick the best match for the explanation you just wrote and not what looks like the best explanation. Rimac needs to know what your thought process was so it can do a better job of helping you understand the physics concepts involved in solving the problem.

It asks you to do this matching for explanation questions because it cannot understand explanations accurately enough. However, for all the other answers you type in it is fairly accurate."

As the student and system begin the review of an assigned homework problem, Rimac first instructs her to view a brief video that describes how to solve the homework problem and then they engage in a reflective dialogue about that problem. See [10] for a more detailed description of the pilot study and planned analyses for testing the nine interactivity decision rules. Note that in this paper we are focusing on just one of these nine rules.

4. EVALUATING THE APPROACH FOR ELICITING EXPLANATIONS

We visited the high schools at two different points during their physics courses, once when the students were ready to learn the material covered in Rimac's kinematics problems and later when they were ready for Rimac's dynamics problems. The students who used the experimental version of Rimac, in which the interactivity rules are implemented, had the opportunity to discuss three kinematics problems and three dynamics problems. In this preliminary evaluation of our elicitation approach, due to time constraints, we have examined the explanation request data for the three kinematics problems only and for just a subset of the students who used this version of the system (93 of 193 students).

From this subset of the data, we extracted and reviewed the dialogue data for 16 of the 18 system turns where students were asked to interpret their own responses. Two of the 18 turns were requests to provide equations as justifications so we excluded those for this analysis. Although there are 93 students in this data sample, not all questions are asked of all students. The questions that are asked depend on the student's responses during the dialogue. In this sampling of students there were 251 instances of requests for explanations/justifications to examine. We reviewed each response and categorized the student's initial typed response as either relevant to the question asked, an acknowledgement that she had guessed or didn't know, or irrelevant to the physics domain. We listed this analysis per problem in order to see if there is any indication that the problem, the order in which the problems were discussed or the types of questions asked might influence the students' behaviors.

1. T: Why do you think that acceleration is upward?
2. S: nevermind
3. T: Which of the following is the best match for what you just entered? [menu choices: 'Because the kangaroo changed direction | Acceleration is always in the opposite direction as velocity | I guessed | None of the above']
4. S: Acceleration is always in the opposite direction as velocity

Figure 2: Example of an actual student reaction when asked to explain an incorrect answer

	Problem 1	Problem 2	Problem 3
relevant content	88% (123)	68% (50)	72% (28)
guessed/idk	5% (9)	15% (11)	23% (9)
irrelevant content	3% (7)	16% (12)	5% (2)

Table 1: A categorization of students' responses to requests for explanations

Table 1 shows that students rarely refused to respond (3-5%) for problems 1 and 3. Refusal to answer increased noticeably for problem 2 (16%) so it does appear that the questions asked for this problem may have impacted students' willingness to respond.

While most student replies that we counted as refusals to answer were simply replies of "because", we noted in problem 2 some instances of "ok" and "never mind" as in the excerpt shown in Figure 2. Note that in this case the system is asking the student to explain her incorrect answer. After saying "nevermind" the student may be selecting what she was thinking or something that looked like a true statement. Of the 16 system turns evaluated, 8 were in response to wrong answers. However, problem 2 accounted for 6 of those requests and problem 3 for the rest. The increased refusal to answer for problem 2 may be because there are more requests to explain incorrect responses. When students self-explain and recognize they have made an error, it appears they sometimes try to figure out how to retract their answer or evade the expected follow-up. However, if the student does not realize her error, it does seem that the formative feedback could be beneficial. For example, consider the dialogue excerpt shown in Figure 3 where the student is asked to explain a wrong answer in line 6 and the formative feedback is provided in lines 8-13.

Next we evaluated how well the students performed in their role as human interpreter. We reviewed each response by following the training given to human interpreters for the experiments described in [2] which used the Cordillera system (i.e. how would I match this response to the choices provided). One of the reviewers of the students' responses (the main author) was also one of the human interpreters (wizard) for the first two versions of the Cordillera system. So it could be instructive to informally compare and contrast the interpretation tasks for the two systems and populations from the perspective of a wizard.

1. T: If the stone is thrown with a larger horizontal initial velocity, what happens to the time spent in the air (i.e. will it be the same or larger?)
2. S: larger
3. T: Hmm...So why did you think this?
4. S: more time
5. T: Thanks for explaining your reasoning.
6. T: So which of the reasons below best matches your explanation above for why you think the stone will stay longer in the air? [menu choices: 'I used my intuition but I am not sure how to explain it in physics terms | Throwing it harder means the stone will not fall right away but will continue horizontally for a longer time before beginning to fall | Throwing it harder makes the stone spin and float on the air better | I've seen this happen before | Because it is going farther, it will take longer to get there | Another reason']
7. S: Because it is going farther, it will take longer to get there
8. T: That doesn't sound right to me. What equation would you use to reason that out? [menu choices: '$d_x = v_x * t | d_x = v_x * t + 0.5 * a_x * t^2 | vf_x^2 = vi_x^2 + 2 * a_x * d_x$']
9. S: $d_x = v_x * t + 0.5 * a_x * t^2$
10. T: Almost. What is the value of a_x?
11. S: 0 m/s^2
12. T: Right, so then the equation becomes $d_x = v_x * t$.
13. T: The horizontal component of the motion is constant velocity (not accelerating), so $d_x = v_x * t$ applies. In this case, because the velocity is bigger, the displacement will be bigger, but the time will not be.

Figure 3: Excerpt of actual student interaction where system requests explanation of wrong response

The Cordillera students were all undergraduates and their explanations were longer and required more effort to interpret and match. However, there was usually one clear candidate for the match and when matching to a correct response the criteria were that the necessary and sufficient details were present or could be easily inferred and no additional details signalled an error in thinking. The choices were authored to provide the minimum that would be needed to qualify as a complete answer. While wizards did not have to be physics experts, they did need to understand the physics concepts being discussed.

In contrast, the Rimac students were all in high school and their explanations were relatively short. We did not expect students to do well with a set of minimal match choices since we assume you need to understand the physics concepts to determine whether an answer actually matches. So instead the Rimac dialogue authors provided responses for matching

Context: Problem solved for homework "A red colored stone is thrown horizontally at a velocity of 5.0 m/s from the roof of a 35.0 m building and later hits the ground below. What is the red stone's horizontal displacement? Ignore the effects of air friction."
Question: Why did we need to find the time first?
Choices:

1. time is the same in both directions
2. d = vt
3. we don't have enough information to solve for displacement in the horizontal direction
4. we can find the displacement if we know how long it is moving at the given velocity
5. another reason

Figure 4: An example of where some choices offered to students for matching are related to the same underlying explanation (as in choices 1,3 and 4)

that were intended to be closer to what a student might say and were based on input from teachers and responses collected during pilot testing. As a result some of the choices offered to students for matching varied only in the detail provided or how it was expressed. But these similar choices present the same formative feedback when selected. For example, in Figure 4, choice 2 is close to a good explanation but requires more detail to be complete while choices 1,3 and 4 are all related to the same underlying explanation. If the student selects 1,3 or 4 as a match then the underlying explanation is presented as an acknowledgement and may be interpreted by the student as a reformulation. If the student selects choice 2 then the system provides scaffolding that elicits the missing details.

So during our review of students' response matching, we selected all that we considered to be potential matches and not just the best match. The rationale was that if a student selected one of a similar set of responses that had details that were missing in her response, a wizard cannot know whether the student's self-explanation included these details and she chose not to express them or whether she thought more detail was necessary and was trying to avoid formative feedback.

After reviewing the student responses we counted the number of times we disagreed with their match choices. Again we present the results per problem. Table 2 shows that students' performance may be similar to that of an automated explanation matcher. The larger disagreement for problem 2 could be due to students possibly trying to evade further feedback when they were asked to explain an incorrect answer or could be related to the questions or answer choices offered. If deserves a closer look in future work to see if a reason can be identified.

However, overall the students seem less perturbed by the results of their matching behaviors. They still continued to respond to the requests for explanations as shown by the

	Problem 1	Problem2	Problem 3
agree	78% (108)	59% (43)	74% (29)
disagree	22% (31)	41% (30)	25% (10)

Table 2: Reviewer agreement with students' matches of their responses

small increase in irrelevant content in Table 1, which remains low with an increase from 3 to 5% when moving from the first to last problem. The increase from problem 1 to problem 3 in "guessed/idk" could be due to fatigue, the explanations requested or more specifically asking for more explanations for incorrect answers in problems 2 and 3. Although the number of "guessed/idk" decreased from problem 2 (11) to problem 3 (9), recall that some students completed problems in two class sessions and some in one. This was because of differences in the length of classes across schools.

To give an idea of an upper bound for agreement, we do not expect 100% agreement between the reviewer and a trained human interpreter (wizard). When offline reviewers examined the selection choices made by the real-time human interpreters for the Cordillera system for just the most difficult student responses (i.e. those that fell into the "none of the above" category), the reviewer disagreed with 1% of the assignments to this category [8]. However, the lower bound that is allowable for matching when students are acting as the interpreter is still an open question. It will depend on whether formative feedback on the explanation related to their match choice is beneficial.

By the time of the workshop, we expect to have completed the above analyses for all students for the kinematics problems.

5. PLANS FOR EVALUATING THE FORMATIVE FEEDBACK GIVEN ON EXPLANATIONS

Recall that in the instructions we read to students we asked that they match the response they gave rather than picking what looks like the best response. We offer motivation to do this by pointing out that the system needs to know their thought processes so that it can provide better help for them. We are assuming that the formative feedback of a good match will be better than the "none of the above" feedback. However, this remains to be seen.

But because our experiment was not testing this specific hypothesis, we cannot answer this question directly (e.g. compare to a condition in which the formative feedback is always the "none of the above" feedback). However, we can test for correlations between various match qualities (i.e. trained reviewer agreed or disagreed with student) and learning of the concepts addressed by the requested explanation. This would suggest how important it is for students to receive more adapted formative feedback. In addition, we can test for gains on concepts covered in an explanation when the student's explanation is incorrect and relative to the quality of the match the student provided. This could suggest whether the feedback that followed was beneficial.

This preliminary analysis of the effects of formative feedback is forthcoming. We are currently scoring the pre and post-tests, which (when completed) will allow us to measure learning of particular concepts.

6. ACKNOWLEDGEMENTS

This research is supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A10063 to the University of Pittsburgh. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute or the U.S. Department of Education.

7. REFERENCES

[1] B. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13:4–16, 1984.

[2] M. Chi, K. VanLehn, D. Litman, and P. Jordan. An evaluation of pedagogical tutorial tactics for a natural language tutoring system: A reinforcement learning approach. *International Journal of Artificial Intelligence in Education*, 21:83–113, 2011.

[3] M. T. Chi, N. De Leeuw, M.-H. Chiu, and C. LaVancher. Eliciting self-explanations improves understanding. *Cognitive science*, 18(3):439–477, 1994.

[4] M. O. Dzikovska, P. Bell, A. Isard, and J. D. Moore. Evaluating language understanding accuracy with respect to objective outcomes in a dialogue system. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 471–481. Association for Computational Linguistics, 2012.

[5] A. C. Graesser, P. Wiemer-Hastings, K. Wiemer-Hastings, D. Harter, N. Person, and the TRG. Using latent semantic analysis to evaluate the contributions of students in autotutor. *Interactive Learning Environments*, 8:129–148, 2000.

[6] P. Jordan, B. Hall, M. Ringenberg, Y. Cui, and C. Rosé. Tools for authoring a dialogue agent that participates in learning studies. In *Proceeding of Artificial Intelligence in Education Conference*, pages 43–50, 2007.

[7] P. Jordan, S. Katz, P. Albacete, M. Ford, and C. Wilson. Reformulating student contributions in tutorial dialogue. In *Proceedings of 7th International Natural Language Generation Conference*, pages 95–99, 2012.

[8] P. Jordan, D. Litman, M. Lipschultz, and J. Drummond. Evidence of misunderstandings in tutorial dialogue and their impact on learning. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED), Brighton, UK, July*, 2009.

[9] S. Katz and P. Albacete. A tutoring system that simulates the highly interactive nature of human tutoring. *Educational Psychology (Special Issue on Advanced Learning Technologies)*, in press.

[10] S. Katz, P. Albacete, M. Ford, P. Jordan, M. Lipschultz, D. Litman, S. Silliman, and C. Wilson. Pilot test of a natural-language tutoring system for physics that simulates the highly interactive nature of human tutoring. In K. Yacef and H. Lane, editors,

Proceedings of Artificial Intelligence in Education Conference, 2012.

[11] M. Makatchev and K. VanLehn. Analyzing completeness and correctness of utterances using an atms. In *Proceeding of the 2005 conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology*, pages 403–410, 2005.

[12] V. Rus and A. C. Graesser. Deeper natural language processing for evaluating student answers in intelligent tutoring systems. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1495. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[13] S. Siler, C. P. Rosé, T. Frost, K. Vanlehn, and P. Koehler. Evaluating knowledge construction dialogs (kcds) versus minilessons within andes2 and alone. In *Workshop W6 on Empirical Methods for Tutorial Dialogue Systems*, page 9, 2002.

Must Feedback Disrupt Presence in Serious Games?

Matthew Jensen Hays, H. Chad Lane, Daniel M. Auerbach
University of Southern California
Institute for Creative Technologies
12015 E Waterfront Dr
Playa Vista CA 90094 USA
310-448-5398
{hays,lane,auerbach}@ict.usc.edu

ABSTRACT

Serious games are generally designed with two goals in mind: promoting learning and creating compelling and engaging experiences (sometimes termed *a sense of presence*). Presence itself is believed to promote learning, but serious games often attempt to further increase pedagogical value. One way to do so is to use an intelligent tutoring system (ITS) to provide feedback during gameplay. Some researchers have expressed concern that, because feedback from an ITS is often *extrinsic* (i.e., it operates outside of the primary game mechanic), attending to it disrupts players' sense of presence. As a result, learning may be unintentionally *hindered* by an ITS. However, the most beneficial conditions of instruction are often counterintuitive; in this paper, we challenge the assumption that feedback during learning hinders sense of presence. Across three experiments, we examined how an ITS that provided extrinsic feedback during a serious game affected presence. Across different modalities and conditions, we found that feedback and other ITS features do not always affect presence. Our results suggest that it is possible to provide extrinsic feedback in a serious game without detracting from the immersive power of the game itself.

Keywords

presence, immersion, learning, feedback, serious games, tutoring

1. WHAT'S IN A GAME?

We have all had the experience of being engrossed in an artificial experience, whether it's a good book, an epic movie, a round of golf, or a couple levels of *Angry Birds* on a long elevator ride. Several features of games, especially, can make hours fly by, unnoticed. The interactivity of games draws players' attention from non-game thoughts and stimuli. The rules of the game, too, are designed to add uncertainty and difficulty—and eventual reward—to the pursuit of an objective. Putting a ball into a cup is made fun, for example, by requiring that one use golf clubs to do so—rather than simply picking up the ball, walking over to the cup, and dropping it in. The eventual reward (sinking a putt) compels players to persist and eventually improve.

Real-world games are fun, in part, because they take place in an environment that supports continued play (e.g., a golf course). Digital games, instead, must transport a player to the world of the game. This experience of being in the world of the game is sometimes referred to as a sense of *presence* [1]. Presence can be measured in several ways. The Temple Presence Inventory (TPI), for example, is a robust instrument for estimating the feeling of non-mediation in a multimedia experience [2]. The TPI consists of a series of statements to which participants respond to items such as "How often did you want to or did you make eye contact with a person you saw/heard?" with ratings between 1 (never) and 7 (always). These statements are organized into several subscales, which correspond to various aspects of the experience that contribute to the sense of non-mediation. The two subscales we used were *social* (the experience of direct interaction with an artificial counterpart) and *spatial* (the experience of direct contact with an artificial environment).

2. WHAT'S IN A SERIOUS GAME?

In addition to the standard traits of a digital game (e.g., the difficult pursuit of an in-game objective, creating a sense of presence), *serious games* feature an objective outside the game itself. By "playing" a serious game, one becomes better at a real-world task—or is at least better prepared to learn that task from subsequent instruction or practice [3]. Examples of serious games include CyberCIEGE, which is designed to teach people about the functions of computer network security measures. Another example is Spent, a simple simulation of a U.S. Citizen's experience at the poverty line in a difficult economy with no bootstraps on which to pull. The difficulty, interactivity, and reward structure of serious gameplay can compel students to persist in learning something they would otherwise find dry or boring.

Serious games are also used in part because the sense of presence created by gameplay may improve learning [4, but see 5, 6, 7]. On the other hand, the outside-the-game objective may be in conflict with that intent. Of course, a game-player's sense of presence in a serious (or otherwise overtly educational) game may be disrupted by poorly integrated pedagogical content. For example, some educational games alternate between play and instruction. But even well integrated instructional content may be distracting; the user may occasionally stop to consider how to apply what they are learning to similar real-world tasks. If presence affects learning, this withdrawal may be detrimental.

This potential conflict may be exacerbated when features that are intended to facilitate training are added to a serious game. These

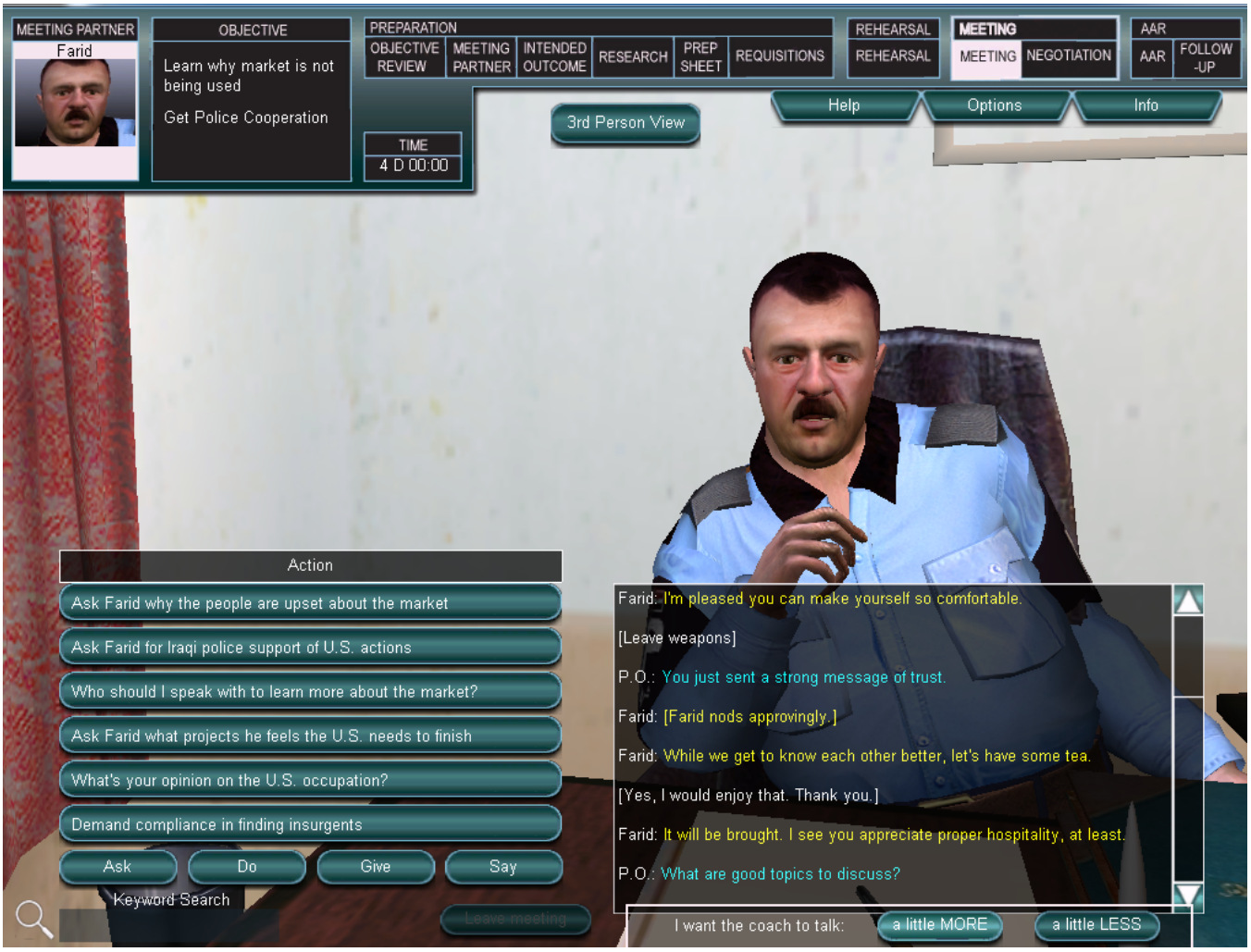


Figure 1. A meeting in BiLAT. In the transcript pane (bottom right), the feedback from the ITS-driven coach appears as blue text. Below that are buttons used to adjust how frequently the coach (P. O., above) decides to intervene (Experiments 2 and 3).

features may directly interfere, or may simply underscore that the player is *using* the game to achieve the external goal, as opposed to *playing* the game because it is fun.

One such feature is an intelligent tutoring system (ITS). An ITS is a computer program or computing device that factors student performance into when and how it generates and provides guidance [8]. The development of ITSs (and other learning-centric game features) is usually guided by principles of cognitive psychology and instructional design [8-10]. However, those principles are often developed in experimental laboratories, in which motivation and fun may not be priorities. Thus, ITSs may provide pedagogically valid feedback, but they may do so in a way that further deepens the rift between gameplay and learning. The goal of the studies reported in this paper was to determine whether extrinsic feedback from an ITS necessarily negatively affects learners' sense of presence when playing a serious game.

3. BILAT: A SERIOUS GAME ABOUT CROSS-CULTURAL NEGOTIATION

The serious game we chose to use for our investigation is the Enhanced Learning Environments with Creative Technologies for Bilateral negotiations (ELECT BiLAT), a screenshot from which is shown in Figure 1. BiLAT provides an environment in which learners can prepare for, execute, and review cross-cultural meetings with virtual characters. The instructional design and underlying structure are focused on knowledge components that relate to culture and negotiation skills.

Before a meeting, players research their meeting partner, learning about his/her interests and experiences. This research provides information that can help the character establish a personal connection with the character during their meeting. Once the meeting begins (shown in Figure 1), players interact with the characters by selecting an action from a menu system of pre-authored actions (e.g., Ask "Who should I speak with to learn more about the market?"). The character responds to the learner with a synthesized voice and physical gestures. The player and the virtual character thus conduct a turn-based interaction, and the transcript of the meeting appears on screen in the panel at the bottom right of Figure 1.

Although dozens of variables govern the actions of the character and the responses that will be chosen, the variable of primary importance is trust. BiLAT characters display a variety of emotions in their responses, but trust is the persistent record of how well players have used their interpersonal and intercultural

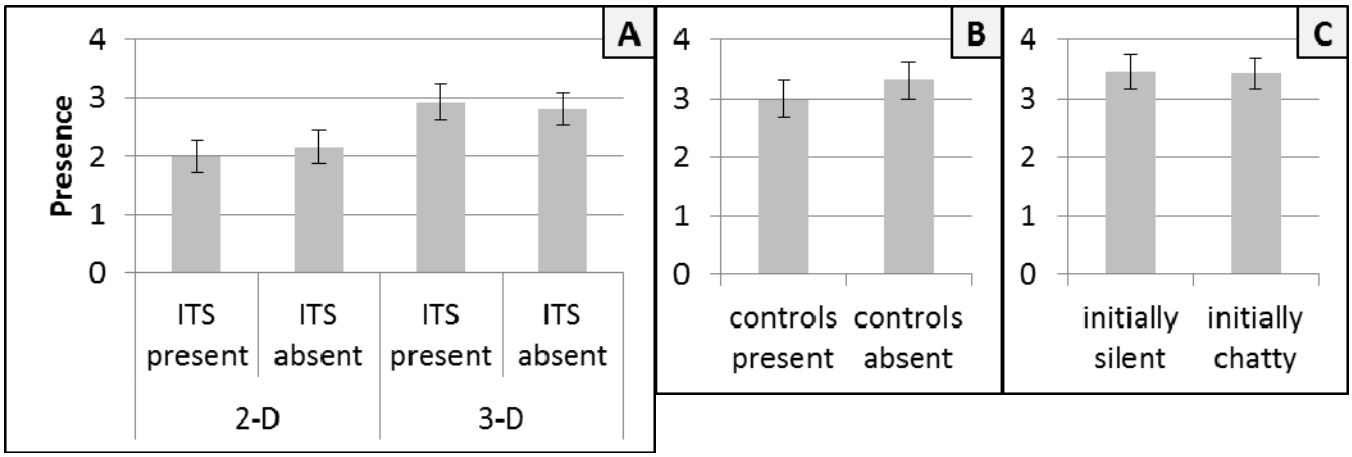


Figure 2. Results from all three experiments. Panel A displays presence as a function of interface richness and ITS activation in Experiment 1. Panel B displays presence as a function of ITS interactivity in Experiment 2. Panel C displays presence as a function of initial ITS feedback frequency in Experiment 3. Error bars represent the standard error of the mean.

skills. In the simulation, trust is a major factor in whether BiLAT characters will agree to negotiate and what deals they will accept. A mistrusting character may demand unfair deals or refuse to negotiate. (For a more detailed description of BiLAT's development and functionality, please see [11, 12].)

The characters' responses and decisions can be considered *internal feedback*. They help the player grasp the knowledge components through the primary interaction that constitutes gameplay. For example, if the player decides to offer the character a bottle of wine as a gift, the character will be offended and say so: "I can't believe you'd even bring that into my home." Depending on what the player has encountered both in and out of BiLAT, the player may conclude that the character does not like wine or that wine is a culturally inappropriate gift.

During BiLAT gameplay, learners can be assisted by an ITS. In meetings with characters, the ITS takes the form of a disembodied, omniscient "coach." The player can read the coach's input in the transcript pane, but the meeting partner is not aware of the coach's presence or input. In other words, the coach is an angel on the player's shoulder. The input the coach provides is outside of the primary interaction that constitutes gameplay; it is *external feedback*.

The coach can provide guidance about past actions ("A bottle of wine probably wasn't the best gift.") or hints about future actions ("What gift can you give Hassan as a gesture of goodwill?"). This advice can be either very general (i.e., focused on the underlying knowledge components) or very specific to something a player has done. For example, the coach could decide to say "Don't give Hassan a bottle of wine" or "Make sure your gifts are culturally appropriate." (For a detailed description of the ITS architecture, please see [13].)

4. EXPERIMENT 1: THE EFFECTS OF EXTERNAL FEEDBACK ON PRESENCE

In Experiment 1, we examined the effects of explicit ITS feedback on learners' sense of presence during BiLAT gameplay. The manipulation was straightforward: whether the ITS was active or inactive during gameplay. We also added another manipulation: whether the sensory experience was rich or poor. Our goal in adding this manipulation was to ensure that we would be able to detect effects on presence with our system, procedure, and participation numbers. Thus, one group of the participants encountered the standard BiLAT experience: a 3-D environment in which a virtual character with realistic body language talks to the player in accented English. The other group of participants encountered a simplified, silent, primarily text-based 2-D environment. We held constant all other aspects of the system for the two groups. Specifically, the BiLAT characters drew from the same sets of utterances and the coach used the same algorithms to decide when to intervene. Only the interface of the two groups' experiences differed. After interacting with the system in one of the four resultant (randomly assigned) conditions, the participants completed the TPI.

Panel A of Figure 2 shows that there was a main effect of interface on presence. A greater sense of presence was created by the 3-D interface (M = 2.88, SE = .21) than by the 2-D interface (M = 2.08, SE = .20): $F(1, 45) = 7.86$, $p = .007$. There was not a main effect of ITS activation on presence. Indeed, presence ratings were similar in the active-ITS condition (M = 2.46, SE = .20) and the inactive-ITS condition (M = 2.49, SE = .20): $F < 1$, *ns*. There was also no interaction between interface and ITS activation on presence: $F < 1$, *ns*. It appears that receiving extrinsic feedback from an ITS does not necessarily affect presence. Thus, any pedagogical benefit provided by the ITS appears not to burden the immersive experience.

5. EXPERIMENT 2: THE EFFECTS OF FEEDBACK CONTROLS ON PRESENCE

In Experiment 1, the activity of the ITS was entirely out of the participants' control. In Experiment 2, we added interactivity to the ITS. We gave the participants the ability to modify the coach's behavior. We thought that this interactivity might cause the participants to attend to the coach (or the external training goal of the serious game) in a way that would disrupt presence.

There were two groups of participants, both of which encountered the standard, 3-D BiLAT system with the coach operating according to its default algorithms. One of the groups was also provided with "coach controls." These controls took the form of the buttons seen in the bottom right corner of Figure 1. These buttons suggested to the participants that they could nudge (up or down) the frequency with which the coach decided to intervene.

The controls, however, were only cosmetic (although they still visually and aurally behaved like other in-game buttons). We chose to display but disable them in order to manipulate the participants' *belief* that they could control the coach without allowing learning, performance, success, or frustration to vary uncontrollably. After interacting with the system in one of the two (randomly assigned) conditions, the participants completed the TPI.

Panel B of Figure 2 shows that there was no main effect of ITS controls on presence: $F(1, 22) < 1$, *ns*. This result provides more evidence that even direct interaction with an ITS outside the primary game mechanic does not necessarily disrupt presence.

6. EXPERIMENT 3: THE EFFECT OF ITS HELPFULNESS ON PRESENCE

Experiment 3 was designed to extend Experiment 2. Our goal was to determine whether the BiLAT ITS could deliver feedback in a way that would disrupt presence. To that end, we modified the coach's feedback-timing algorithms to draw even more attention to the ITS than in Experiment 2. For one group of participants, the coach began the session in complete silence. For the other group of participants, the coach began the session by speaking up on every single turn. We activated the "nudge" controls, which were merely cosmetic in Experiment 2, to encourage the participants to interact with the ITS as much as possible. Each press of "a little more" or "a little less" changed (by 5%) the probability that the coach would speak up on the next turn. After interacting with the system in one of the two (randomly assigned) conditions, the participants completed the TPI.

As can be seen in Panel C of Figure 2, the participants in both conditions provided similar presence ratings: $F(1, 22) < 1$, *ns*. That is, whether the participants' experience began with constant chatter or complete silence from the ITS, their sense of presence remained relatively unaffected. Moreover, in comparing the three panels in Figure 2, it is clear that the participants' overall ratings were similar across all three experiments—despite drastic differences in feedback algorithms and ITS interactivity. It seems that, unless an ITS is designed with the express purpose of disrupting gameplay, it may not interfere with the immersion created by a serious game.

7. GENERAL DISCUSSION

Interpersonal and intercultural skills, to be frank, may not be the most compelling instructional topics. However, when playing BiLAT, players and participants become very engaged. A participant in one study, when meeting with a particularly stubborn character, took off his headphones and threw them across the room, saying "I *know* he wants to agree to it, and he's just trying to give me a headache!"

Our research demonstrates that this sense of presence is not necessarily disrupted when external feedback from an ITS is added to a serious game. Further, learners can even be instructed to directly interact with the ITS, yet still suffer no decrement to self-reported presence. On the other hand, the use of a single, self-report measure of presence is a limitation of the present study. A more compelling case may be presented by including corroborating physiological data. (We did not examine measures of performance or learning because it would have been impossible to disentangle from each other the effects of feedback on presence, feedback on learning, and presence on learning.)

Although these results may seem surprising, external stimuli interrupt engaging experiences quite frequently, often with no negative results. Many people have put down and then resumed an engrossing book—and been able to reinstate their enjoyment of and engagement with the story. Perhaps a compelling narrative or rewarding gameplay may make some serious and educational games robust to interruptions, as well. In these cases, people may be able to suspend and resume their engagement as they wish. If so, it is interesting to consider the extent to which developers can add pedagogically focused game features without sacrificing learners' immersion. It is reasonable to assume there is some limit to the intrusiveness an ITS can exhibit while still being effective—but the present studies suggest that that limit is above zero.

8. ACKNOWLEDGMENTS

The work depicted here was sponsored by the U.S. Army. Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

9. REFERENCES

[1] Lombard, M. and Ditton, T. *At the heart of it all: The concept of presence.*, 1997.

[2] Lombard, M., Ditton, T. and Weinstein, L. *Measuring presence: The Temple Presence Inventory*. 2009.

[3] Dede, C. Immersive interfaces for engagement and learning. *Science*, 323, 2009, 66-69.

[4] Rowe, J. P., Shores, L. R., Mott, B. W. and Lester, J. C. Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education*, in press.

[5] McQuiggan, S., Rowe, J., Lee, S. and Lester, J. *Story-based learning: The impact of narrative on learning experiences and outcomes*. 2008.

[6] Lane, H. C., Hays, M. J., Auerbach, D. and Core, M. *Investigating the relationship between presence and learning in a serious game*. 2010.

[7] Moreno, R. and Mayer, R. E. Personalized messages that promote science learning in virtual environments. *Journal of Educational Psychology*, 96, 2004, 165-173.

[8] Woolf, B. *Building intelligent interactive tutors*. Morgan Kaufmann, 2008.

[9] Mayer, R. E. *Multimedia Learning*. Cambridge University Press, 2001.

[10] Shute, V. J. and Psotka, J. *Intelligent tutoring systems: Past, present, and future*. MacMillan, 1996.

[11] Hill, R. W., Belanich, J., Lane, H. C., Core, M., Dixon, M., Forbell, E., Kim, J. and Hart, J. *Pedagogically structured game-based training: Development of the Elect BiLAT simulation*. 2006.

[12] Kim, J., Hill, R. W., Durlach, P. J., Lane, H. C., Forbell, E., Core, M., Marsella, S., Pynadath, D. V. and Hart, J. BiLAT: A game-based environment for practicing negotiation in a cultural context. *International Journal of Artificial Intelligence in Education*, 2009.

[13] Lane, H. C., Hays, M. J., Auerbach, D., Core, M., Gomboc, D., Forbell, E. and Rosenberg, M. Coaching intercultural communication in a serious game. *Proceedings of the 18th International Conference on Computers in Education*, 2008, 35-42.

Individual differences in the effect of feedback on children's change in analogical reasoning

Claire E. Stevenson
Leiden University, Psychology Methods & Statistics
Wassenaarseweg 52, Postbus 9555
2300 RB Leiden, The Netherlands
+31 71 527 3789
cstevenson@fsw.leidenuniv.nl

Paul A. L. de Boeck
Ohio State University, Dept. of Psychology
232 Lazenby Hall, 1827 Neal Ave,
Columbus, OH, 43210, USA
+1 614 292 4131
deboeck.2@osu.edu

Wilma C. M. Resing
Leiden University, Developmental Psychology
Wassenaarseweg 52, Postbus 9555
2300 RB Leiden, The Netherlands
+31 71 527 3789
resing@fsw.leidenuniv.nl

Willem J. Heiser
Leiden University, Psychology Methods & Statistics
Wassenaarseweg 52, Postbus 9555
2300 RB Leiden, The Netherlands
+31 71 527 3789
heiser@fsw.leidenuniv.nl

ABSTRACT

Various forms of feedback are used in formative assessment and interactive learning environments. The effects of different types of feedback are often examined at a group level. However, effective feedback may differ in learners with different characteristics or between learners at different stages in the learning process. In this paper explanatory item response theory (IRT) models are used to examine individual differences in feedback effects in children's performance on a computerized pretest-training-posttest assessment of analogical reasoning. The role of working memory and strategy-use as well as interactions between these factors were examined in a sample of 1000 children who received either stepwise elaborated feedback, repeated simple feedback or no feedback during the training sessions. The results show that working memory efficiency significantly predicted initial ability and confirm that elaborate feedback is the most effective form of training in this particular interactive learning environment. Furthermore, children with initially less advanced strategy-use benefitted far more from each type of feedback than the children displaying more advanced strategies and this was unrelated to working memory efficiency. In children with advanced strategy-use working memory appears to moderate the effect of training. Explanatory IRT analyses appear useful in disentangling the effects of learner characteristics on performance and change during formative assessment and could possibly be used in optimizing feedback in computerized training and assessment environments.

Keywords

Figural analogies, measuring change, item response theory, formative assessment

1. INTRODUCTION

Computer-based interactive learning environments have enormous potential in optimizing learning by providing feedback tailored to an individual's instructional needs. However, determining what type of feedback best optimizes the learning of a particular task for a particular individual is a complex endeavor. The effectiveness of different types of feedback is not always clear-cut. Furthermore, individual differences may be present in how effective each of these types of feedback is at different stages in the learning process.

In formative assessment different types of feedback can be used. Shute distinguished a range of feedback-types from simple forms such as verification of correct response to elaborated feedback where errors may be flagged, an opportunity to try again is provided and/or strategic prompts are given on how to proceed with the problem [Shutte 2008]. Kluger and DeNisi [1996] argued that although simple feedback, such as information on correctness of response or provision of the correct answer, has the reputation of improving performance on tasks, its effect is not clear-cut and only improves performance or learning in two-thirds of the studies included in their meta-analysis. Furthermore, more recent research demonstrates that elaborate feedback, such as providing scaffolds or an explanation, is generally more effective than simple outcome feedback [Hattie and Gan, 2011; Narciss and Huth 2006; Shutte 2008]. For example, a meta-analysis of effects of different forms of item-based feedback in computer-based environments reports that elaborated feedback shows higher effect sizes than simple outcome feedback, especially in higher-level learning outcomes, where transfer of previous learning to new situations or tasks is required [van der Kleij et al. 2013].

In the case of formative assessment the aim is to optimize learning at an individual level. In this educational setting the assumption is that there are individual differences both in initial ability as well as the effect of different types of feedback during an individual's learning process. Furthermore, different types of feedback may be more effective during successive stages in the learning process. However, effective feedback may differ for different types of learners or at different stages in the learning

process. For example working memory efficiency and strategy-use have been implicated as predictors of performance in (computer-based) learning [Siegler and Svetina, 2002; Stevenson 2012; Tunteler et al. 2008]. In this study these factors were examined in conjunction with feedback-type as possible predictors of learning outcomes in a computerized training and assessment of analogical reasoning.

Initial ability or learning stage especially appears to play an important role in the effect of different forms of feedback on learning [Hattie and Timperley 2007]. For example, in a previous study on children's change in analogical reasoning training utilizing repeated simple feedback was contrasted with graduated prompting techniques, a form of elaborated feedback where increasingly specific strategic hints guide the child to the correct solution [Campione and Brown 1987; Resing and Elliott, 2011]. The researchers found that although graduated prompts led to greater performance gains on the whole, this form of training was most effective for children who performed poorly on the pretest [Stevenson et al. 2013a]. These results could not be explained by ceiling effects or regression to the mean. Furthermore, this result coincided with other cognitive training studies in various domains where interventions were generally more effective in initially lower performing or at-risk populations. Does this mean that providing elaborate versus simple feedback is not necessarily beneficial for more advanced learners?

To further explore the role of initial ability on feedback effects we examined the role of children's initial solution strategies (analogical versus non-analogical, see Figure 1) in the effect of three types of feedback: (1) step-wise elaborated feedback, (2) repeated simple feedback or (3) no feedback. The hypothesis was that children with initially weaker analogical reasoning strategies, characterized by "duplicate" (copying object next to empty box) solutions or "other / creating a zoo" solutions would benefit most from more elaborate forms of feedback whereas children who were already capable of applying analogical reasoning strategies (providing (partially) correct solutions) would not show differential benefit in the different types of feedback training. The role of working memory, which has often been shown to be related to analogy solving skills, but not always able to account for children's change in analogical reasoning [Stevenson et al. 2013b], was also taken into account in these analyses.

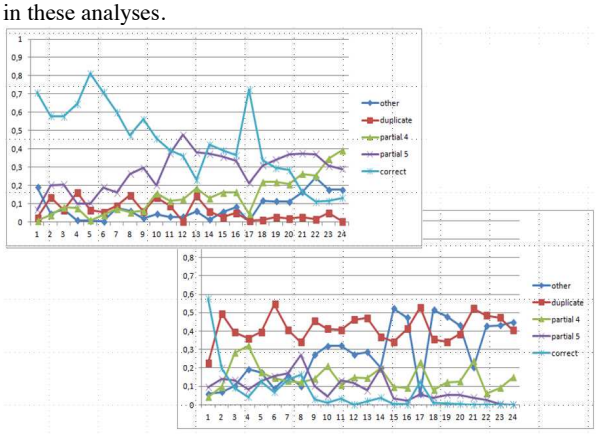


Figure 1. Depiction of strategy distribution within two pretest strategy groups: non-analogical reasoners (top left) and analogical reasoners (bottom right).

2. METHODS

2.1 Sample

1000 children from five age-groups (kindergarten, first through fourth grade) were recruited from public elementary schools of similar middle class SES in the south-west of the Netherlands. The sample consisted of 374 boys and 626 girls, with a mean age of 7 years, 3 months (range 4.9-11.3 years). The schools were selected based upon their willingness to participate and written informed consent for children's participation was obtained from the parents.

2.2 Design & Procedure

The data utilized in this study is a combination from five separate studies utilizing a pretest-intervention-posttest control-group design [Stevenson 2012]. In each study the children were randomly blocked to the step-wise elaborative feedback (graduated prompts), repeated simple feedback or a control condition without feedback based on their scores on a cognitive ability reasoning subtest (visual exclusion from the Revised Amsterdam Children's Intelligence Test [Bleichrodt et al. 1987] or the Standard Progressive Matrices [Raven et al. 2004]). The three intervention conditions presented in this study are: (1) stepwise elaborate feedback, (2) repeated simple feedback, or (3) no feedback. Four analogy testing and intervention sessions took place weekly and lasted 20-30 minutes each. Prior to the analogy testing sessions the children were also administered the Automated Working Memory Assessment to assess verbal (subtest listening recall) and visuo-spatial (spatial span) working memory [Alloway 2007]. All participants were tested individually in a quiet room at the child's school by educational psychology students trained in the procedure.

2.3 Analogical reasoning assessment

AnimaLogica was used to test and train children in analogical reasoning [Stevenson 2012]. The figural analogies (A:B::C:?) comprise of 2x2 matrices with familiar animals as objects (see Figure 2). The animals changed horizontally or vertically by color, orientation, size, position, quantity or animal type. The number of transformations – or object changes – provide an indication of item difficulty [Mulholland et al. 1980]. The children were asked to construct the solution to the analogy using drag & drop functions to place animal figures into the empty box in the lower left or right quadrant of the matrix. A maximum of two animals were present in each analogy. These were available in three colors (red, yellow, blue) and two sizes (large, small). The orientation (facing left or right) could be changed by clicking the animal figure. Quantity was specified by the number of animal figures placed in the empty box. Position was specified by location of the figure placed in the box.

The pretest and posttest items were isomorphs [Freund and Holling 2011] in which the items only differ in color and type of animal, but utilize the exact same transformations to ensure the same difficulty level. The number of items different per age group but included overlapping items ability could be estimated reliably using item response models. The internal consistency of each of the versions was considered very good with $\alpha \geq .90$.

Before each testing or training session two example items were provided with simple instructions on how to solve the analogies. If the child's solution was incorrect the correct solution was shown before proceeding to the next item. During the testing phases the remaining items were administered without feedback.

Table 1.

Overview of the prompts used in the elaborative feedback condition.

Prompt	Verbal Instruction
0	Here's a puzzle with animal pictures. The animals from this box have been taken away. Can you figure out which ones go in the empty box?
1	Do you remember what to do? Look carefully. Think hard. Now try to solve the puzzle.
2	This animal picture changes to this one. This one should change the same way.
3	So what changes here? Ok remember this one changes the same way.
4	See, this picture changes to this one because…
5	Which animal goes in the empty box? The elephant or the horse? What color should it be? Red, Yellow or Blue? …Size? Quantity? Orientation? Position?...

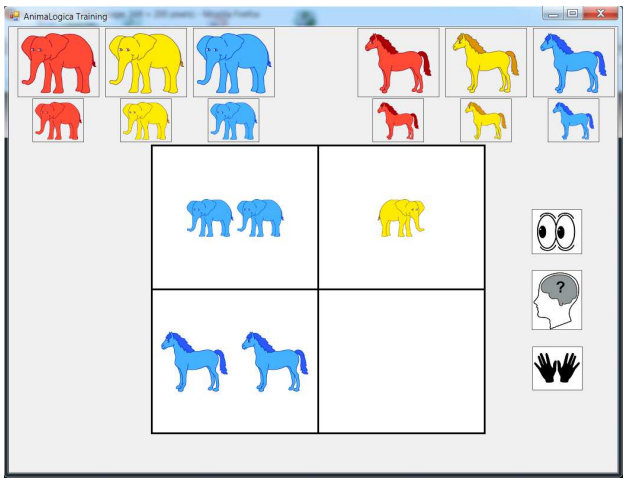


Figure 2. Depiction of visual effects emphasizes cues from prompt 1 to "Look carefully", "Think hard" and then "Try to solve the puzzle" (these are not all shown at once).

2.3.1. Feedback Interventions.

The *stepwise elaborate feedback condition* received training according to the graduated prompts method [Campione and Brown 1987; Resing and Elliott 2011] which consisted of stepwise instructions beginning with general, metacognitive prompts, such as focusing attention, followed by cognitive hints, emphasizing the transformations and solution procedure, and ending with step-by-step scaffolds to solve the problem (see Table 1). The prompts were mostly auditory in nature and accompanied by visual effects support the explanations (see Figures 2 & 3). A maximum of five prompts were administered. Once the child answered an item correctly the child was asked to explain his/her answer; no further prompts were provided and the next item was administered.

The *simple feedback condition* received auditory feedback on whether or not the outcome was correct and this was repeated until the item was solved correctly or five attempts were made to solve the item. After the fifth incorrect attempt the correct solution was shown before proceeding to the next item. If a correct solution was found before five attempts then the next item was administered.

In the *control condition* the children received the exact same items as in the other two conditions but did not receive help or feedback in solving them. Therefore, the children only practiced solving the items but were not trained in analogical reasoning.

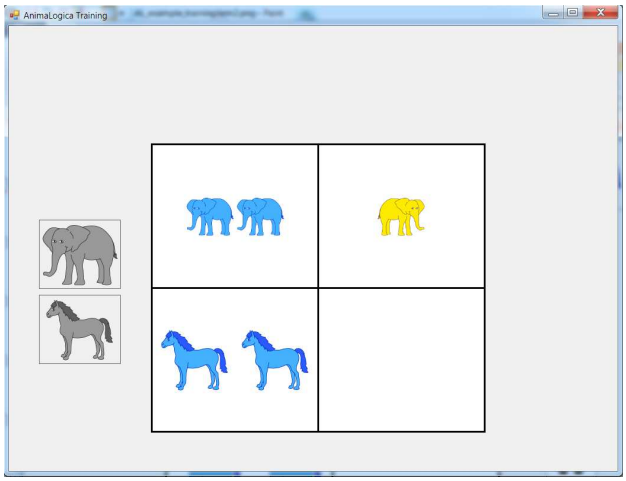


Figure 3a. Visual effects emphasizing prompt 5 where scaffolds are used to solve the puzzle: "Which animal belongs in the empty box?".

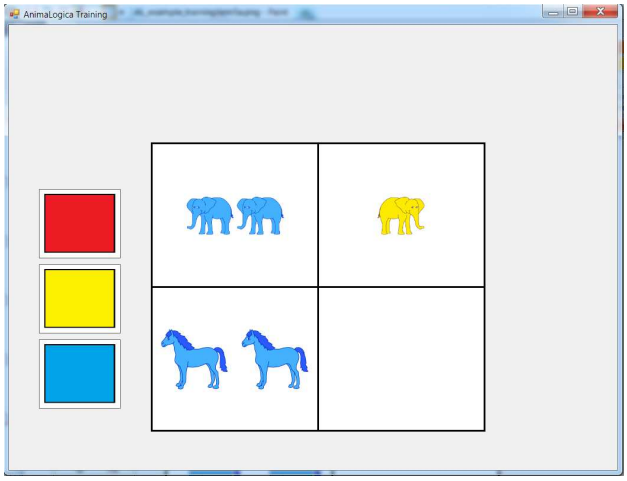


Figure 3b. Prompt 5 scaffold: "What color should it be?".

2.4 Statistical Models

Disentangling the complex changes in ability over time on an individual basis requires complex statistical models. For example, using raw gain scores (posttest minus pretest score) to measure change can lead measurement errors due to the unreliability of the gain score, the regression effect of repeated administration and that the scale units for change do not share constant meaning for test takers with different pretest scores and [de Bock 1976; Lord 1963]. These problems are potentially solved by placing ability scores for pretest and posttest on a

joint interval measurement scale using logistic models such as those employed in item response theory (IRT) [Embretson and Reise 2000]. In the Rasch model, one of the most simple IRT models, the chance that an item is solved correctly depends on the difference between the latent ability of the learner and the difficulty of the presented item or problem. The Rasch-based gain score provides a good basis for the latent scaling of learning and change because the gain score has the same meaning in terms of log odds (i.e. the logarithm of probability of correct vs. incorrect) across the entire measurement scale [Embretson and Reise 2000]. Therefore, this study applied IRT models to analyze individual differences in feedback effects on learning and change [Stevenson et al. 2013a].

2.4.1 Explanatory IRT analyses

Each of the hypotheses about the children's performance and change was investigated using model comparison. First a reference model was created and then predictors were added successively to so that the fit of the new model could be compared to the previous (nested) model using a likelihood ratio (LR) test, which assesses change in goodness of fit. The models were estimated using the lme4 package for R [Bates and Maecheler 2010] as described by [De Boeck et al. 2011].

2.4.2 Null model

The initial reference model (M0) was a simple IRT model with random intercepts for both persons and items (pretest and posttest) where the probability of a correct response of person *p* on item *i* is expressed as shown in equation 1.

$$P(y_{pi} = 1 \mid \theta_p, \beta_i) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)}$$

$$\text{where } \theta_p \sim N(0, \sigma_\theta^2) \text{ and } \beta_i \sim N(0, \sigma_\beta^2) \quad (1)$$

2.4.3 Modelling learning and change

This study employs repeated testing. In order to account for this effect a session parameter has to be added to the null model to represent average change from pretest to posttest. However, this model assumes the effect of retesting to be equal for all children. In order to allow for individual differences in improvement from pretest to posttest a random parameter that allows for the session effect to vary over persons was added. In this model, Embretson's Multidimensional Rasch Model for Learning and Change (MRMLC, see M2 in Table 1), the chance that an item is solved correctly (P_{ip}) also depends on the difference between the examinee's latent ability (θ_p) and the item difficulty (β_i) [Embretson 1991]. Yet, the ability is built up through the testing occasions *m* up to *k* in a summation term, which indicates which abilities (θ_{pm}) must be included for person *p* on occasion *k*.

$$P(y_{ipk} = 1 \mid \theta_{pk}, \beta_i) = \frac{\exp(\sum_m^k \theta_{pm} - \beta_i)}{1 + \exp(\sum_m^k \theta_{pm} - \beta_i)}$$

$$\text{where } \theta_{pm} \sim N(0, \sigma_\theta^2) \text{ and } \beta_i \sim N(0, \sigma_\beta^2) \quad (2)$$

The initial ability factor, θ_{p1}, refers to the first measurement occasion (i.e. pretest) and the so-called modifiabilities (θ_{pm} with $m>1$) represents the change from one occasion to the next. In the present model examining pretest to posttest change $k=2$ and the modifiability θ_{p2} refers to performance change from pretest to posttest.

2.4.4 Modelling sources of individual differences in learning and change

The formula in equation 2 can be extended by including other item or person predictor variables and evaluating their effects on the latent scale [De Boeck and Wilson 2004]. Person predictors are denoted as Z_{pj} $(j=1,\ldots,J)$ and have regression parameters ζ_j. The item predictor (e.g. number of transformations) can be denoted as X_i (k=1) and has the regression parameter δ. These predictors are successively entered into the null model (see equation 1) as follows, with indices *i* for items, *p* for persons, *j* for the person covariate used as a predictor variable and *k* for the item covariate used a predictor variable.

$$P(y_{pi} = 1 \mid Z_{p1} \ldots Z_{pJ}, \beta_i) = \frac{\exp(\sum_{j=1}^{J} \zeta_j Z_{pj} + \varepsilon_p + \delta X_{ik} + \varepsilon_i)}{1 + \exp(\sum_{j=1}^{J} \zeta_j Z_{pj} + \varepsilon_p + \delta X_{ik} + \varepsilon_i)}$$

$$\text{where } \varepsilon_p \sim N(0, \sigma_{\varepsilon p}^2) \text{ and } \varepsilon_i \sim N(0, \sigma_{\varepsilon i}^2) \quad (3)$$

This equation represents models M3-6 in the results presented in Table 2.

Table 2.

Overview of the estimated IRT models.

Model	Nested Model	Effects			AIC	BIC	-LL	LR test[a]	
		Fixed	Random over Persons	Random over Items				*df*	*Λ*
M0			Intercept	Intercept	37575	37600	18784		
M1	M0	+ Session	"	"	35741	35775	17866	1	1835.90***
M2	M1		+Session	"	34871	34922	17429	2	874.18***
M3	M2	+ Session* Condition	"	"	34063	34132	17024	2	811.52***
M4	M3	* Strategy group	"	"	33773	33944	16866	12	314.50***
M5	M4	* WM	"	"	18014	18236	8979	8	15775***

[a] The LR-test comprises a comparison between the model and the nested model. *** p < .001

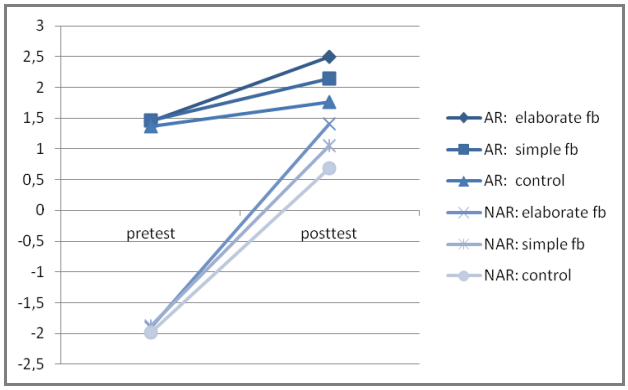


Figure 4. Plot of M5 with logit (y-axis) by Session (x-axis) for Analogical Reasoners (AR) versus Non-analogical reasoners (NAR) for each feedback condition (elaborate, repeated simple and control).

3. RESULTS

Table 2 displays the outcomes of the model building steps. As can be seen in the right-most column the addition of each new

predictor in the explanatory IRT model significantly improved model fit. From M0 to M1 we could statistically infer that there was a main effect for training. The inclusion of individual regression lines for performance change from the pretest to posttest was deemed warranted given the improved model fit from M1 to M2. The significant model comparison result from M2 to M3 shows us that the different types of feedback had different "change" slopes. The difference in performance change from pretest to posttest between the two strategy-groups is shown in model M4 (see Figure 4). Finally, from M4 to M5 we could statistically infer working memory was differentially related to performance change per condition and strategy group. Analysis of the simple contrasts indicated that working memory moderated feedback effects in the analogical reasoners (AR strategy group), but was unrelated to performance change in the non-analogical reasoners (NAR strategy group) (simple feedback: $B = -1.38$, $p < .01$ and elaborated feedback: $B = -1.37$, $p < .01$, reference category = no feedback / control condition).

Significant fixed main effects were found for Session, Strategy group, verbal and visuo-spatial Working memory. Significant fixed interaction effects were found for Session x Condition, Session x Strategy group, Session x Working memory, Strategy group x Working memory and Session x Strategy group x Working memory. Random intercepts were present for persons ($SD_{ability} = .62$, $SD_{modifiability} = .70$, $r = -.24$) and items ($SD = .74$).

Table 3.

Estimates of fixed effects in M5.

	B	SE	*p*
Intercept	- 0.32	.42	.44
Session (reference = pretest)	2.17	.16	<.001
Simple Feedback Condition (reference = control)	0.10	.10	.32
Elaborate Feedback Condition (reference = control)	0.08	.10	.41
Strategy-group (reference = non-analogical reasoners)	3.26	.11	<.001
Verbal working memory	0.23	.09	.01
Visuo-spatial working memory	0.26	.04	<.001
Session * Simple Feedback Condition	0.28	.13	.04
Session * Elaborate Feedback Condition	0.65	.13	<.001
Session * Strategy-group	-1.65	.12	<.001
Session * Verbal Working memory	0.47	.11	<.001
Strategy-group * Verbal Working memory	0.08	.10	.43
Session * Strategy-group * Verbal Working memory	-0.61	.13	<.001

4. CONCLUSION

This paper presented our recent research in the area of statistical models of formative feedback effects in performance and change in children's analogical reasoning. The results showed that individual differences stemming from initial strategy-use and working memory efficiency were present and influenced the effect feedback. Elaborate feedback was more effective than simple feedback. Working memory was a predictor of pretest performance. Working memory also moderated feedback effects but only in children in the advanced strategy-use group. Working memory most likely forms a bottleneck in children's analogical reasoning on difficult analogy tasks [Richland et al. 2006]; however children with less advanced strategies most likely were unable to solve the more difficult analogy items which would require accurate solving steps and the accompanying greater taxation of working memory to do so. Finally, initial strategy-use interacted with feedback-type in that children using less advanced strategies at pretest benefited more from each form of feedback during training compared to the children displaying more advanced strategies at pretest. On the whole, the main conclusion is that elaborated feedback, presently implemented using graduated prompting techniques, appears to be the advisable form of feedback in advancing children's change in analogical reasoning.

Given the great potential of computer-based interactive learning environments to provide feedback tailored to an individual's instructional needs an important task is creating algorithms to optimize feedback provision and thus learning. On the one hand (meta-analyses of) randomized pretest-training-posttest control experiments that contrast the effectiveness of different types of feedback and explore sources of individual differences herein as discussed in the present paper provide essential information concerning which factors could be used to optimize feedback. However an investigation of the effects of specific elaborated feedback prompts on a trial-by-trial basis [Golden et al. 2012] and the interactions with learner characteristics or task performance (e.g., strategy-use) using item response theory models is a promising next step towards the provision of optimal feedback in interactive learning environments. Thus the next step in this research project is to expand upon the present findings concerning the effectiveness of the stepwise elaborated feedback and disentangle the immediate effects of the separate prompts during the training process. It will be interesting to see whether different types of prompts better aid more or less advanced learners with more or less efficient working memory to solve the items presented during training.

5. ACKNOWLEDGMENTS

Thank you to Marian Hickendorff for her assistance with the initial analyses of the strategy groups presented in this paper. We also thank Femke Stad and Carlijn Bergwerff for their assistance with data collection and coding.

6. REFERENCES

[1] SHUTE, V. J. 2008. Focus on Formative Feedback, *Review of Educational Research*, 78 (1), 153-189.

[2] KLUGER, A. N. AND DENISI, A.1996. The Effects of Feedback Interventions on Performance: A Historical Review , a Meta-Analysis , and a Preliminary Feedback Intervention Theory. *Psychological Bulletin*, 2(2), 254-284.

[3] NARCISS, S. AND HUTH, K. 2006. Fostering achievement and motivation with bug-related tutoring feedback in a computer-based training for written subtraction. *Learning and Instruction*, 16(4), 310-322.

[4] HATTIE, J. AND GAN, M. 2011. Instruction Based on Feedback. In *Handbook of Research on Learning*, R. E. MAYER AND P. A. ALEXANDER, eds. New York, New York, USA: Routledge, 249-271.

[5] VAN DER KLEIJ, F. M. FESKENS, R. C. W. AND EGGEN, T. J. H. M. submitted 2013. The effectiveness of methods for providing written feedback through a computer-based assessment for learning: A systematic review. 1-25.

[6] STEVENSON, C.E. 2012. *Puzzling with Potential Dynamic testing of analogical reasoning in children*.

Doctoral dissertation. Amsterdam: Leiden University, 1-191.

[7] SIEGLER, R. S. AND SVETINA, M. 2002. A microgenetic/cross-sectional study of matrix completion: comparing short-term and long-term change. *Child development*, 73(3) 793-809.

[8] TUNTELER, E., PRONK, C. M. E. AND RESING, W. C. M. 2008. Inter- and intra-individual variability in the process of change in the use of analogical strategies to solve geometric tasks in children: A microgenetic analysis. *Learning and Individual Differences*, 18(1), 44-60.

[9] HATTIE, J. AND TIMPERLEY, H. 2007. The Power of Feedback, *Review of Educational Research*, 77 (1), 81-112.

[10] CAMPIONE, J. C. AND BROWN, A. L. Linking dynamic assessment with school achievement. In *Dynamic assessment: an interactional approach to evaluating learning potential*, C. S. LIDZ, Ed. New York, New York, USA: Guilford Press, 82-109.

[11] RESING, W. C. M. AND ELLIOTT, J. G. 2011. Dynamic testing with tangible electronics: measuring children's change in strategy use with a series completion task. *The British journal of educational psychology*, 81(4), 579-605.

[12] STEVENSON, C. E., HICKENDORFF, M. RESING, W. C. M., HEISER, W. J. AND DE BOECK, P. A. L. 2013a. Intelligence Explanatory item response modeling of children's change on a dynamic test of analogical reasoning, *Intelligence*, 41(3), 157-168.

[13] STEVENSON, C. E., HEISER, W. J. and RESING, W. C. M. 2013b. Working memory as a moderator of training and transfer of analogical reasoning in children, *Contemporary Educational Psychology*, 38(3) 159-169.

[14] MULHOLLAND, T. M., PELLEGRINO, J. W. AND GLASER, R. 1980. Components of geometric analogy solution. *Cognitive psychology*, 12(2) 252-284.

[15] FREUND, P. A. AND HOLLING, H. 2011. How to get really smart: Modeling retest and training effects in ability testing using computer-generated figural matrix items, *Intelligence*, 39(4), 233-243.

[16] F. DE BOCK, 1976. Basic issues in the measurement of change. In *Advances in psychological and educational measurement*., D. N. M. DE GRUIJTER AND VAN DER L. J. T. KAMP, Eds. New York, New York, USA: Wiley.

[17] LORD, F. M. 1963. Elementary models for measuring change. In *Problems in measuring change*., C. W. HARRIS, Ed. Madison, Wisconsin, USA: University of Wisconsin Press, 21-38.

[18] EMBRETSON, S. E. AND REISE, S. 2000. *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

[19] BATES, D. AND MAECHELER, M. 2004, lme4: Linear Mixed-Effects models using S4 Classes. http://r-forge.r-project.org/projects/lme4/.

[20] DE BOECK, P. A. L., BAKKER, M., ZWITSER, R., NIVARD, M., HOFMAN, A. TUERLINCKX, F. AND PARTCHEV, I. 2011. The estimation of item reponse models with the lmer Function from the lme4 Package in R, *Journal of Statistical Software*, 39 (12), 1-27.

[21] EMBRETSON, S. E. 1991. A multidimensional latent trait model for measuring learning and change, *Psychometrika*, 56(3), 495-515.

[22] DE BOECK, P. A. L. AND WILSON, M. 2004. *Explanatory item response models: A generalized linear and nonlinear approach*. New York, New York, USA: Springer.

[23] RICHLAND, L. E. MORRISON, R. G. AND HOLYOAK, K. J. 2006. Children's development of analogical reasoning: insights from scene analogy problems. *Journal of experimental child psychology*, 94(3), 249-73.

[24] GOLDIN, I. M. KOEDINGER, K. R. AND ALEVEN, V. A. W. M. M. 2012. Learner Differences in Hint Processing, In *Proceedings of the 5th International Conference on Educational Data Mining*, Chiana, Greece, 2012.

[25] BLEICHRODT, N., DRENTH, P.J.D., ZAAL, J.N. & RESING, W.C.M. 1987. *Handleiding bij de Revisie Amsterdamse Kinder Intelligentie Test [Manual of the Revised Amsterdam Child Intelligence Test]*. Lisse: Swets & Zeitlinger.

[26] RAVEN, J. RAVEN, J.C. & COURT, J.H. 2004. Manual for Raven's Progressive Matrices and Vocabulary Scales. San Antonio, Texas: Harcourt Assessment.

[27] ALLOWAY, T.P. 2007. *Automated Working Memory Assessment (AWMA)*. London: Harcourt Assessment.

An Intelligent Tutoring System for Japanese Language Particles with User Assessment and Feedback

Zachary T. Chung
Ateneo de Manila University
Department of Information Systems and Computer Science (DISCS)
Katipunan Ave., Loyola Heights, Quezon City, Philippines
+63-2-426-6001 local 5660
zachary.chung@obf.ateneo.edu

Hiroko Nagai, Ph.D.
Ateneo de Manila University
Japanese Studies Program (JSP)
Katipunan Ave., Loyola Heights, Quezon City, Philippines
+63-2-426-6001 local 5248
hyabut@ateneo.edu

Ma. Mercedes T. Rodrigo, Ph.D.
Ateneo de Manila University
Department of Information Systems and Computer Science (DISCS)
Katipunan Ave., Loyola Heights, Quezon City, Philippines
+63-2-426-6001 local 5660
mrodrigo@ateneo.edu

ABSTRACT

In recent years, an increasing number of Ateneo students have been taking an interest in the Japanese language. For Ateneo students beginning their study of the language however, Japanese particles are difficult concepts because they cannot be translated to equivalent words in English. For a beginner learner, it is inevitable to view a second language with the lens of a first language as shown by the concept of transfer in second language acquisition. As a result, learners tend to misconstrue Japanese particles by attempting to understand them with respect to non-existent equivalents in English.

In this research, we develop an intelligent tutoring system for Ateneo students taking introductory Japanese (FLC 1JSP) to aid them better understand Japanese particles. The system would assess the learner's understanding of Japanese particles by practice and depending on which particle where most mistakes are made, the system would give instructional feedback. Feedback to be implemented in the system use visual prototypes that represent the meaning of the particle. We hope to see if visual representations can also teach Japanese particles to students as an alternative to text-detailed explanations such as those commonly found in textbooks.

Categories and Subject Descriptors

K.3.1 [**Computer Uses in Education**]: Computer-assisted instruction (CAI), Distance learning

General Terms

Design, Experimentation, Human Factors, Theory

Keywords

Intelligent Tutoring Systems (ITS), Japanese particles, Case Particles, Japanese language, Visual prototypes

1. INTRODUCTION

1.1 Context of the Study

An increasing number of Ateneo students are minoring in Japanese Studies to learn more about the Japanese language and culture. Students beginning Japanese in their FLC 1JSP (Introduction to Japanese) course encounter difficulty with Japanese particles regarding proper usage and context: に (ni)、へ (e)、を (wo)、と(to)、で (de)、の (no) 、は (wa) 、が (ga)

1.2 Research Objectives

In this paper, we discuss the development of a web-based Intelligent Tutoring System (ITS) addressing the difficulty of Ateneo students with Japanese particles - a system that facilitates practice with feedback that clarifies particle usage and meaning. We attempt the following questions:

1. How do we create an intelligent tutoring system for Japanese to help students better understand the concept of Japanese particles?
2. Other than the topic and subject marking particles は(wa) and が(ga) respectively, which particles do students make the most mistakes with in FLC 1JSP?
3. What do these errors imply about the student's mental model of Japanese particles?

1.3 Scope and Limitations

Users of the system developed are primarily FLC 1JSP students of Ateneo de Manila University, hence system content is scoped to the said course. We aim to supplement the language knowledge of FLC 1JSP students; instruction in the system is geared towards clarifying understanding, as opposed to teaching anew.

Finally, we utilize visual feedback in the system based on prototypes by Sugimura (discussed in section 2.1) because we like to know if Japanese particles can also be taught by animations aside from explanations of their meaning. For particle and word combinations that do have not have any visual representations, we use textual feedback based on Socratic questioning as our alternative form of feedback. We hope to see if computer animations and our combination thereof can be an effective means to clarify these Japanese particles to students.

2. FRAMEWORK

2.1 Visual Prototypes for Japanese Particles

Japanese particles can be taught using images representative of their meaning. Sugimura demonstrates that each Japanese particle can be represented by a prototype image and he states that learners would have less cognitive load learning Japanese particles in this manner than rote memorization of a definition [11]. In this research, we develop visual feedback, based on five prototype images of the following particles from FLC1 JSP: **ni, e, to, no, de**.

1. The particle **ni**

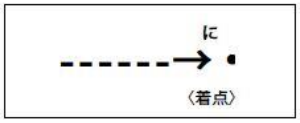


Figure 2.1: Prototypical meaning of ni [11]

Ni shows the directionality of an agent's action and its binding effect to a target [11]; **ni** can also indicate the place or time of existence of a subject [11]. These two usages are generalized into the image of a point, indicating a destination or a point in time shown above. Compared to **e**, **ni** emphasizes the destination as opposed to the process, depicted by the dotted arrow in figure 2.1.

2. The particle **de**

The particle **de** indicates *space* where an action takes place in the nominative or accusative case [11]. The prototype of this particle is shown in figure 2.2 below:

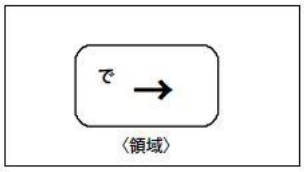


Figure 2.2: Prototypical meaning of de [11]

The arrow in figure 2.2 above represents some force acting within an enclosed space. Though **de** is likewise represented with an arrow like **ni**, **de** emphasizes an action performed within the bounds of a certain space [11].

3. The particle **e**

In essence, **e** is similar to **ni** for indicating the direction of an action. Compared to **ni**, **e** puts emphasis in the process or means of an agent to get to a destination [11; Dr. Hiroko Nagai personal communication, May 5, 2012]. The particle **e** is represented according to Sugimura in figure 2.3 below [11]:

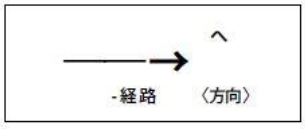


Figure 2.3: Prototypical meaning of e [11]

4. The particle **to**

According to Morita, the particle **to** has a unificative meaning associated to its usage [11], where two agents work together to perform an action. In a prototype image, Sugimura depicts the meaning of the particle **to** as follows [11] (Refer to Figure 2.4):

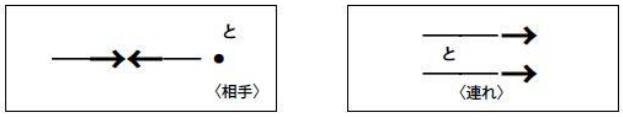


Figure 2.4: Prototypical meaning of to [11]: An action performed together in companionship.

5. The particle **no**

No denotes relations between nouns but these have various forms hence, we only consider **no** for the following usages in our research as scoped in FLC1 JSP:

1. A is the possessor of B (like the B of A or A's B) such as: watashi **no** kaban (My bag)
2. A is the location where B belongs to (B in/at A) such as: ateneo **no** gakusei (A student in Ateneo) and;
3. A created B hence B is possessed by A such as: gakusei **no** sakubun (A student's essay)

In all these three cases above, the particle **no** connects nouns together, such that the preceding noun phrase forms a phrase to modify a following noun phrase [6]. According to Oya, Japanese language adviser of the Japan Foundation, Manila, the particle **no** can be depicted in a prototypical image of a circle (noun 2) inside a larger circle (noun 1) and so on as follows for these three usages:

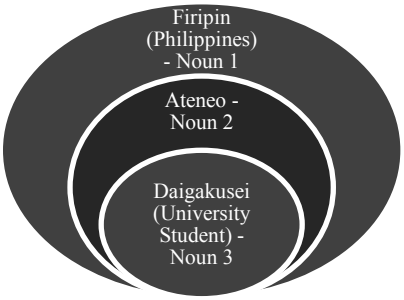


Figure 2.5: Firipin no ateneo no daigakusei: Combining nouns with no

In figure 2.5 above, the largest circle sets a scope to the circle(s) enclosed within. In this representation, Ateneo is in the Philippines and the student is affiliated with the Ateneo, thus a set of concentric circles. The enclosed nouns are connected by **no,** forming one noun, meaning "A University Student of Ateneo in the Philippines".

2.2 Visuals as Feedback in Multimedia Learning

Students learn best by seeing the value and importance of information presented so it is important to sustain interest using a feedback medium that coincides with the learning style of a student, which is "the manner in which individuals perceive and process information in learning situations" [4].

According to the Cognitive Theory of Multimedia Learning by Mayer, Multimedia instructional messages designed according to how the human mind works are more likely to lead to meaningful learning than those that are not [7]. The theory states that humans seek to make sense of multimedia presentations in relation to their collected experiences. Hence, visual feedback would be effective given that it resembles common human experience while depicting the meaning of Japanese particles. Table 3.1

summarizes the theory regarding how learners relate visuals to experience.

Table 3.1 Image-related Processes in the Cognitive Theory of Multimedia Learning: Building Connections between Pictorial Models with Prior Knowledge

Process	Description
Selecting images	Learner pays attention to relevant pictures in a multimedia message to create images in working memory.
Organizing images	Learner builds connections among selected images to create a coherent pictorial model in working memory.
Integrating	Learner builds connections between pictorial models and with prior knowledge.

As guidelines for our design of visual feedback, the following are prescribed by the theory [1, 2]:

1. **Focus on Task-Relevant Aspects of Information:** Research show that guiding learners' attention is only useful if it leads the learner to a deeper understanding of the task-relevant parts of the information presented.
2. **Limit Unnecessary Information:** Each piece of information, useful or not has to be processed by the learner so it is additive to cognitive load. According to the Apprehension Principle, information that is not required for the task or problem solving, such as seductive details or eye-catching illustrations, produce extraneous cognitive load that ties attention to less relevant concepts and therefore reduces knowledge acquisition [1].
3. **Attention-guiding Principle:** Supporting the process of selecting relevant information will be useful because it shifts the learners' attention to those parts of information that are needed to understand the key concept of presented materials. Also, since animation is fleeting by nature, often involving simultaneous display changes, it is important to guide learners in understanding the animation so that they do not miss the change. Highlights, visual cues and color coding seem to be appropriate visual instructional aids because novice learners are not able to distinguish between relevant and irrelevant features.
4. **Personalize Instruction:** Learner's attention can be activated in a more effective way if instructions are personalized rather than anonymous, for example by addressing the learner in the first person.

2.3 Error Isolation and Feedback

Mistakes are part of the learning process. According to Gass and Selinker, second language errors do not reflect faulty imitation by a learner; they are attempts to figure out a system by imposing regularity on the language being learned. In fact, mistakes are structured; there is an underlying generalization and this shows a certain level of development [3, 9].

Mistakes are akin to slips of the tongue but errors are systematic and recurring [3]. Errors mean that the learner does not recognize that it is wrong, and by consistent reproduction, he has incorporated it into his system of the target language [3]. In our system, we isolate errors by a pre-test and when an error has been committed at least twice (same particle and context), then feedback is given, targeting the faulty knowledge only as much as possible.

Feedback in our system is designed to let the learner realize his own mistake. We do this by presenting the animation of a learner's erroneous particle side-by-side with the animation of the correct particle. Alternatively, we pose questions or hints to challenge the learner to reconsider his answer instead. In this manner, we allow the learner an opportunity to explore and adjust the application of the form or rule he used to derive his wrong answer to what is correct – *restructuring* in interlanguage processes [9]. This is more effective because it does not interrupt the learner because of fear of being directly corrected [5].

3. METHODOLOGY

3.1 Development Methodology

The Intelligent Tutoring System (ITS) developed in this research is web-based for simpler deployment and testing; Adobe Flash was used to drive animations.

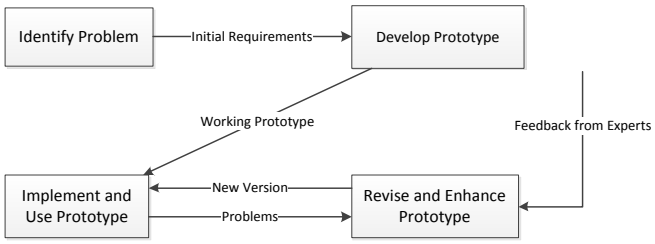


Figure 3.1: The Prototyping Methodology [8]

Based on consultations with FLC 1JSP instructors, students have difficulty mastering case particles because they confuse the different notions these particles provide in sentences. We identified particle pairs students frequently have misconceptions with such as **ni** and **de**, **to** and **no** or **ni** and **to**, etc., then we developed prototype animations that highlight their semantic differences. Then, we showed these animations to instructors for feedback and we improved them to ensure that visual feedback developed in any form teach the correct notion of Japanese particles. Consultations were performed during development mainly with Dr. Hiroko Nagai, Director of the Ateneo Japanese Studies Program, as well as with Mr. Susumu Oya, Japanese Language Adviser of the Japan Foundation, Manila, observing the processes of the prototyping methodology in software development as shown in figure 3.1 above.

3.2 Student Modeling

Student models provide descriptions of learning at a level of granularity that facilitates the encoding of principles and rules in a teaching system [12]. Learner models approximate student behavior by tracking misconceptions in comparison with substandard reasoning patterns. This is performed with the goal of supporting weak students' knowledge and to develop the students' strengths [13]. In our system, we used an overlay model to model the student-user of our system. The model is able to show "the difference between novice and expert reasoning, by indicating how students rate on mastery of each topic, missing knowledge and which curriculum elements need more work" [13]. Since an overlay model is a model of a proper domain subset (i.e. Japanese particles in grammar), we used this model to evaluate students and give feedback accordingly.

The disadvantage of overlay modeling is that students may have knowledge that is not part of an expert's knowledge, thus it is not represented in the student model [13]. However, we mitigate this by creating a multiple-choice based system, where possible answers are contained only within the domain knowledge we teach. Since Japanese particles also have distinct grammatical usages at the level of FLC1 JSP, creating this model is simple because the domain knowledge itself is a matter of conforming to concise grammar rules.

To create the overlay model of the student, we broke down the concept of Japanese particles from FLC1 JSP into its base knowledge components[1]. Among Japanese particles, this is the production rule learned and referenced by a learner to know how to use a Japanese particle. For example, a student can have the following knowledge component: "to indicate the existence of a living or non-living thing, the particle **ni** is used". In total, we have nine (9) knowledge components in our ITS, following a permutation of nine possible contextual usages of all the Japanese particles in our system designed for FLC1 JSP. Note that the particle **e** and the particle **ni** for indicating a place where something moves (direction) are both singly counted as one knowledge component, whereas the rest are considered as individual knowledge components. This is because FLC1 JSP does not yet teach students to differentiate the nuance of both these particles. Also, a more detailed description of how our overlay model operates is discussed below, where we also describe the general operation of the system.

3.2 General ITS System Operation Flow

Students create an account and the ITS presents a pre-test called "Learning Check 1" (See Figure 3.2). This activity shows a battery of eighteen (18) Japanese sentences using the Japanese particles taught in FLC1 JSP; the task for the student in this section is to complete the sentence by choosing the right particle to complete the statement.

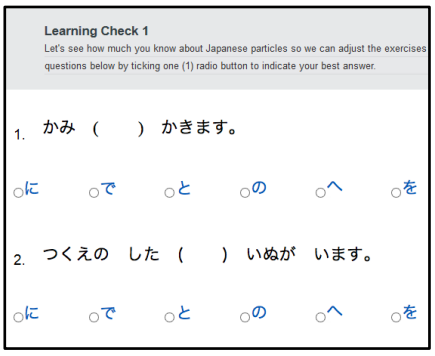


Figure 3.2: Learning Check 1 – Students complete the sentences by supplying the missing particles using the choices provided.

Learning Check 1 is used by the system to create an overlay model of the student. This is used to measure the extent of a student's knowledge of Japanese particles. The model works by assigning points per knowledge component[2] and if a student uses a particle given a context correctly, one (1) point is assigned to the corresponding knowledge component. The model works like a table, where we distribute points across rows and each row is a knowledge component. At the level of FLC1 JSP, since we have nine (9) contextual usages for the particles taught in the course and we have two questions for each usage, we have eighteen (18) questions for Learning Check 1 (See figure 3.3 below):

Pseudo-Overlay Model		
Particle	**Context**	**Pts.**
Ni	Indicate a point in time something takes place.	2
	Indicate a place where something or someone exists.	2
	Indicate target of an action by an agent (uni-directional target).	2
ni/e	Indicate a place towards which something moves.	2
De	Indicate where an event/action takes place.	2
O	Direct objects	2
No	Noun phrase modification to indicate property	2
To	Connect nouns together 'AND'	2
	Indicate target of an action by an agent (bi-directional target).	2
	Total	**18/18**

Figure 3.3 Overlay Model: Point distribution across knowledge components. Maximum attainable score is 18/18

Based on the model, the system displays content in the following section, "Learning Check 2", where actual tutoring takes place. Here, another battery of Japanese sentences is *selectively* presented about the Japanese particles the student appears to have a lack of knowledge with, had the student not met the established minimum number of points per row of the overlay model. While the student is answering, tutoring is now provided - feedback is presented on-the-fly upon mouse clicks in Adobe Flash (See Figure 3.4):

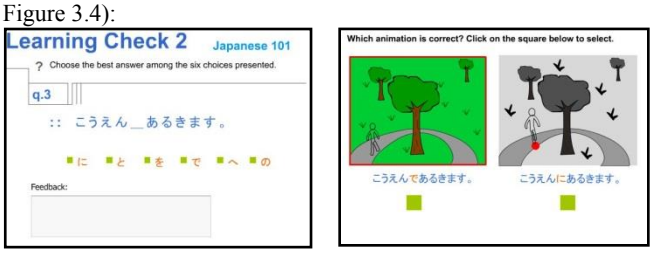


Figure 3.4: Learning Check 2 shows another sentence using 'de'; feedback as needed.

Following Learning Check 2, we present the student a post-test to measure improvements in knowledge. The post-test also serves as a follow-up learning opportunity for the student and the questions used in this section are similar to the questions in the pre-test in terms of count, particle usage and presentation but arranged in a different order. We simply changed the nouns or verbs in the

[1] A knowledge component is a process or a generalization that a learner uses alone, or in combination with other knowledge components to solve a problem [10].

[2] A knowledge component is a process or a generalization that a learner uses alone, or in combination with other knowledge components to solve a problem [7].

sentences and we also maintained two questions per context, hence also making eighteen (18) questions. This allows for comparison on an equal basis between both sections in terms of scoring. Also, to mitigate the possibility that the pre-test is more difficult than the post-test and vice-versa, we also swapped the questions we used in the pre-test with those in the post-test at random. Finally, after using the system, we show a report page to the student concluding the use of the system and how many points were earned based on the overlay model[3]. We also suggest grammar points to the student where more review is recommended based on the result of the post-test (See Figure 3.5 below).

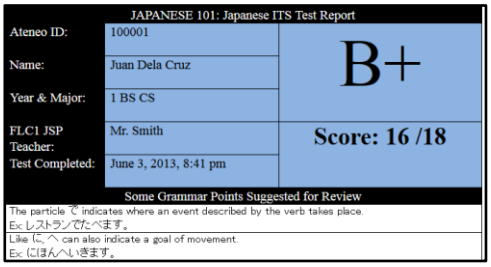


Figure 3.5: Report Page

3.3 Feedback Design

Feedback is given by animations based on the prototype of Japanese particles (See section 2.1). For Japanese particles and their combinations thereof with certain words, forming sentences yielding an image-based representation, we show the student animations with the correct particle and the incorrect particle subtituted in the sentences side-by-side. The goal of this mode of presenting feedback is to allow the student to think for himself the correct answer before the system explicitly shows the answer with explanation. However, for cases non-illustratable, we used textual feedback based on Socratic questioning with cues. The system was designed in mind only to show explicit correction as a last resort because our goal is to restructure grammar knowledge in this tutoring system without being obstrusive to student motivation.

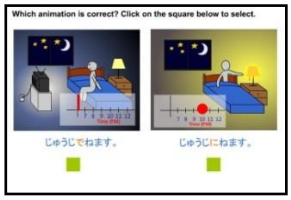


Figure 3.6: Animation Selection: With 'de' for the sentence "juuji _ nemasu (Sleep at 10pm).", the animation of the incorrect answer (left) versus the correct answer (right) is shown.

If the student chooses the correct animation, he is praised and he is shown an explanation why his answer is correct. Otherwise, if the student still chooses the wrong animation, the system shows an explanation of the error and it allows the student to try completing the sentence again (See figure 3.6 below).

[3] Each correct answer in Learning Check 1 is one (1) point. If a student commits an error, the missed points, synonymous to the number of errors made in Learning Check 1, can still be earned back provided that the student answers the corresponding follow-up questions in Learning Check 2.

Figure 3.7 System Responses: Choosing the right animation leads to praise (left); choosing the wrong animation, leads to an explanation of the answer (right).

In cases when animations are not applicable, we give textual feedback in the form of clues based socratic questioning as shown in figure 3.7 below:

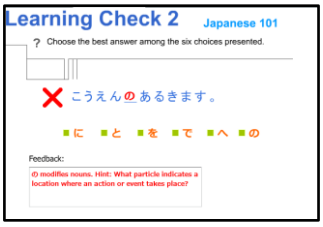


Figure 3.8: Textual feedback for syntatically impossible cases.

4. Results

4.1 Field Testing

As a system designed to target students beginning their study of Japanese in Ateneo, field testing was conducted with the aforementioned students during their FLC1 JSP classes. Students were brought to a computer lab to access the tutoring system online and a total of forty-five (45) students participated in testing across classes handled by three different instructors.

For our results in this research, we focus on presenting analysis based on the results of our pre-test versus post-test scores to see if the students improved using our ITS. Also, we evaluate the experience of the students who used our tutoring system via survey to give us an idea how they find our ITS.

4.2 Testing Methodology

Participants were divided into two (2) groups: twenty-one (21) and twenty-four (24) participants respectively. One group used the ITS such that at the onset of a mistake, corresponding feedback is already shown in Learning Check 2. Another group used the ITS such that the pair of sentences per particle and its context in Learning Check 1 must be incorrect for feedback to be given in Learning Check 2. We formed the two test groups to see how much consideration is adequate before feedback is delivered, although the latter case is ideal based on the notion of error consistency from second language acquisition. A single mistake may not necessarily translate to malformed knowledge about a concept (i.e. a mouse misclick) hence, we believe that consistency is key to isolating true faulty knowledge [3]. During testing, no student was allowed to use any references regarding Japanese particles over the internet.

Figure 4.1: Computer Laboratory Setup

4.4 Pre-test and Post-test Comparison

Table 4.1: Group 1 – One mistake, then Feedback

ID Number	Pre-test (18)	Post-test (18)	Δ
120864	8	9	1
110882	10	12	2
110966	8	4	-4
111329	6	5	-1
91388	9	13	4
122145	7	11	4
112807	10	11	1
123232	12	16	4
123653	8	10	2
123743	9	11	2
123796	9	11	2
123800	4	7	3
114162	11	12	1
94060	5	11	6
120721	10	11	1
123283	9	9	0

Table 4.2: Group 2 – Two mistakes, then Feedback

ID Number	Pre-test (18)	Post-test (18)	Δ
111662	10	11	1
114537	11	9	-2
114553	3	10	7
121314	9	14	5
121359	10	11	1
124592	10	8	-2
114512	5	9	4
110866	8	9	1
111399	11	9	-2
91957	9	8	-1
112107	3	5	2
112227	8	6	2
112017	3	5	2

In testing, we collated scores from different sections. The score in Learning Check 1 is the pre-test column. A separate post-test was carried out after Learning Check 2 to measure the change in knowledge of a student after going through the ITS.

4.5 Group 1 Analysis

For participants with a score of 13 and above in pre-testing for group 1, we did not count their results in our analysis because among all participants in this group, the highest change in score was six (6) points. This means that the highest possible improvement in points can only be measured with scores of twelve (12) and below. Students who obtained a score higher than twelve (12) can only get less than six (6) points to make it the perfect score of eighteen (18) which becomes a cap, hence there is a possibility of unequal comparison in terms of the maximum achievable improvement across students in the test group. To allow for equal and consistent comparison, these participants were excluded in the results [Dr. Joseph Beck, personal communication January 7, 2013].

All participants of group 1 found feedback in the system helpful with an average of 1.235 and 1.471 for their evaluation of the animation and textual feedback respectively on a scale of -2 to 2 (-2 as the lowest and 2 as the highest). Standard deviation values are 0.970 and 0.624 respectively for these averages. These mean that both forms of feedback used in the system are generally regarded as helpful by the participants in the group. Ease of use was evaluated by the students with an average of 1.176 and desire for a similar system for use in FLC 1JSP class was evaluated with an average of 1.294 on the same scale. Standard deviation values are 0.951 and 0.686 respectively for these averages, which point to a good consensus that the system is fairly simple to use and the students would like to have a similar system again in class. Content-wise, all the participants evaluated the system difficulty with 0.765 (from -2, easy until 2, hard) and the standard deviation is 0.437, implying that the system difficulty is manageable in terms of content. Word familiarity was evaluated with an average of 0.294 (-2 as least familiar and 2 as most familiar) with a standard deviation of 0.588. While the averages tell us that students are generally knowledgeable with the words in the system, it is neither high to indicate an excellent understanding of words nor the students are unfamiliar with the words in the system. Based on raw answers collected through the system, knowledge of words pose as a factor behind student errors because to use the correct particle, understanding the notion of words lead the decision to use the correct particle to relate them in sentences.

4.6 Group 2 Analysis

As with group 1, for students who received a score of twelve (12) and above in pretesting, we did not consider their results in our analysis to yield an equal and consistent comparison.

It appears that group 2 participants had a lower average for word familiarity at 0.000, yet the same participants found the system in terms of difficulty easier with an average of 0.615, compared to group 1 on the same scale of -2 to 2. Standard deviation values are both 0.100 and 0.650 respectively for these averages. These mean that while the participants are generally familiar with the words in the system, it also varies greatly per individual. On the other hand, system difficulty is moderate for the participants of this group. Notably, lower averages were attained with 0.667 and 1.083 regarding feedback helpfulness in animation and text respectively. The standard deviations for these values are 0.778 and 0.669 respectively. Ease of use and desire for use of the system in FLC1 JSP gained lower averages at 0.846 and 1.077 with standard deviations values of 0.689 and 0.641 respectively. For these lower scores, it is possible that because participants received feedback less in this group, they found the system less helpful hence more difficult.

5. Conclusion

Table 5.1: Average Delta in Scores (Pre-test vs. Post-test)

1 Mistake (Group 1)	1.75pts.
2 Mistakes (Group 2)	1.38pts.

Findings show that the ITS is effective for both test groups as shown by the positive increase in average delta scores for both test groups. However, more aggressive feedbacking seem to lead to a better perception of the ITS and higher improvement in scores among participants are evident in group 1 than in group 2. In computer-based teaching, it appears that immediate feedback is

better whenever an error is committed at the onset, contrary to what we posited based on concepts in second language acquisition, where it is best to wait for consistent error production first before feedback. In classroom-based teaching, direct correction is not advised, however in computer-based teaching where correction is already indirect by nature through a screen and not by person, immediate correction is more effective and best at the onset of an error.

As initial work in the field, much improvement can still be done to further this ongoing research. In consultation with Dr. Joseph Beck, a visiting professor from Worcester Polytechnic Institute, he suggests to add follow-up questions with our animations, confirming if the user did understand what is taught by the system right after any feedback. Also, from theory to our direct application of image-based teaching of Japanese particles by Sugimura, more investigation regarding effective visual feedback design could be carried out because how we translated the theory into animation based on theoretical meaning may not deliver the intended idea of what we mean to show the student. By doing so, it is possible to uncover the elements in animated feedback students find particularly helpful regarding these particles. From this endeavor, we know that an effective intelligent tutoring system centered on animations for Japanese particles works when it guides the self-discovery learning of students. Success is notable when the students themselves can reproduce the correct answer on their own on a similar question immediately after feedback.

Finally, to have a more in depth understanding of the causality of learner errors and to further confirm our analysis regarding trends among these Japanese particles, we plan to conduct follow-up interviews with select participants to factor in how a user understands certain aspects of the system in relation to a participant's understanding of Japanese.

6. ACKNOWLEDGMENTS

Our thanks to the Ateneo Laboratory for the Learning Sciences; the Department of Science and Technology Philippine Council for Industry, Energy, and Emerging Technology Research and Development (PCIEERD) for the grant entitled, "Development of Affect-Sensitive Interfaces"; the Engineering Research and Development for Technology (ERDT) program for the grant entitled, "Development of an Educational Data Mining Workbench."; Joseph Beck, Ph.D. from Worcester Polytechnic Institute and Susumu Oya from the Japan Foundation.

7. REFERENCES

[1] BETRANCOURT, M. 2005. The animation and interactivity principles in multimedia learning. In The Cambridge Handbook of Multimedia Learning, R.E. Mayer, Ed. Cambridge University Press, NY, 287-295.

[2] BRUNKEN, R. AND PLASS, J. AND LEUTNER, D. 2002. How instruction guides attention in multimedia learning. In Proceedings of the 5th International Workshop of SIG 6 Instructional Design of the European Association for Research on Learning and Instruction (EARLI), Eurfurt, June 2002, H. NEIGEMANN, D. LEUTNER AND R. BRUNKEN, Eds. Waxmann Munster, Munchen, Berlin, 122-123.

[3] GASS, S.M. AND SELINKER, L. 2001. Second language acquisition: An introductory course, 2nd ed. Lawrence Erlbaum Associates, Mahwah, NJ.

[4] GILAKJANI, A.P. AND AHMADI, S.M. 2011. The Effect of Visual, Auditory and Kinesthetic Learning Styles on Language Teaching. In International Conference on Social Science and Humanity, Singapore, February, 2011. IACSIT Press, Singapore. 469-472.

[5] KODAMA, Y. AND KIDA, M. 2010. Teaching grammar. In Nihongo Kyoujuho Series, Japan Foundation. Sanbi Printing, Bunkyo-ku, Tokyo. 25-32.

[6] MAKINO, S AND TSUTSUI, M. 1989. A dictionary of basic Japanese grammar. The Japan Times, Tokyo, Japan.

[7] MAYER, R.E. 2005. Cognitive theory of multimedia learning. In The Cambridge Handbook of Multimedia Learning, R.E. Mayer, Ed. Cambridge University Press, NY, 32-43

[8] NAUMANN, J.D. AND JENKINS, A.M. n.d. Prototyping: The new paradigm for systems development. MIS Quarterly, 6 (3). 29-44.)

[9] ORTEGA, L. 2009. Understanding second language acquisition. In Understanding Language Series, B. Comrie and G. Corbett, Eds. Hodder Education, London, United Kingdom. 116-118.

[10] PITTSBURGH SCIENCE OF LEARNING CENTER: LEARNLAB. 2011. http://www.learnlab.org/research/wiki/index.php/Knowledge_component

[11] SUGIMURA, T. 2002. Teaching each Japanese particles through images (イメージで教える日本語の格助詞, trans.). Language Culture Research Series. Nagoya University International Language and Culture Research Center, Japan. 39-55.

[12] WOOLF, B.P. 1992. Towards a computational model of tutoring. Educational Technology Research and Development, 40. (4) 49-64.

[13] WOOLF, B.P. 2009. Building intelligent interactive tutors: Student centered strategies for revolutionizing e-learning. Morgan Kaufmann Publishers, Burlington, MA

Towards Formative Feedback on Student Arguments

Nancy L. Green
University of North Carolina
Greensboro
Greensboro, NC 27402 USA
1-336-256-1133
nlgreen@uncg.edu

ABSTRACT

This paper presents our ideas on generating formative feedback in the Genetics Argumentation Inquiry Learning (GAIL) system. GAIL will provide undergraduate biology students with tools for constructing Toulmin-style arguments on questions in genetics. Feedback will be based in part on the output of GAIL's argument analyzer, which will compare learner arguments to automatically constructed expert arguments. In addition to identifying problems in the learner's arguments, the analyzer will recognize the argumentation scheme used to construct acceptable arguments. From that, GAIL can instantiate critical questions, a unique form of feedback in intelligent learning environments.

Keywords

Educational Argumentation Systems, Undergraduate Genetics Education.

1. INTRODUCTION

We are developing the Genetics Argumentation Inquiry Learning (GAIL) system for improving undergraduate biology students' argumentation skills in the domain of genetics. As in many educational argumentation systems, GAIL will provide the learner with tools for representing arguments in diagrams due to the cognitive benefit of diagrams [1-3]. In addition, educational systems can exploit the learner's argument diagram as a source of information for providing educational feedback. A prototype graphical user interface (GUI) for GAIL is shown in Figure 1. The top left-hand side of the screen presents a problem, e.g., to make an argument for the claim that J.B., an imaginary patient, has the genetic condition called cystic fibrosis. Below that are possible hypotheses, data about the patient and his biological family members, and biomedical principles that may be relevant to the current problem. The learner can drag these elements into the argument diagramming workspace in the center of the screen to construct an argument in a Toulmin-influenced [4] box-and-arrow notation; a vertical arrow from the *data* points upward to the *claim/conclusion* and the *warrant* is attached at a right-angle to the arrow.

In this paper we describe our planned approach to providing formative feedback based upon automatic analysis of learners' argument diagrams. Expert models for argument analysis will be automatically constructed by GAIL using an argument generator module similar to the argument generator developed for the GenIE Assistant [5]. The expert model will contain all acceptable arguments that can be generated automatically for a given claim from an underlying knowledge base (KB) representing the problem domain. GAIL's argument analyzer will compare the user's argument to the generated expert arguments to identify acceptable learner arguments and weaknesses in the learner's argument. Weaknesses in student arguments are identified using non-domain-specific, non-content-specific rules that recognize common error types, e.g., those observed in a pilot study reported in section 3. In addition, if an argument is acceptable, the analyzer will recognize and output the argumentation scheme underlying the student's argument and its associated critical questions. The output of GAIL's argument analyzer will be utilized by GAIL's feedback generator to provide formative feedback.

In some previous educational argumentation systems, the student's argument diagram is compared to a manually-constructed expert model to provide problem-specific support. However, expert models are expensive to construct and may not cover all possible solutions or errors [6]. In GAIL's approach the expert model is constructed automatically. Other systems use simulation of reasoning to evaluate formal validity but do not provide problem-specific support [6]. GAIL's approach is similar in that it reasons like an expert to generate an argument. Unlike those systems, however, GAIL's approach will provide problem-specific support.

This paper presents how the expert model is generated (section 2), a pilot study of GAIL's GUI prototype that motivated the classification of weaknesses in learners' arguments (section 3), implementation of a prototype argument analyzer (section 4), some issues to be addressed in the planned feedback generator (section 5), and conclusions (section 6).

2. EXPERT MODEL

Generation of expert arguments in GAIL will be done following the approach to argument generation used in the GenIE Assistant, a proof-of-concept system for generating first-drafts of genetic counseling patient letters [5]. Written by genetic counselors to their clients, this type of letter contains biomedical arguments to justify diagnostic testing, the diagnosis of genetic conditions, and the probable genotypes of family members. GenIE's internal components include

- *domain models*, causal models of genetic conditions used by genetic counselors in communication with their clients [7],
- an *argumentation engine* that uses computational definitions of *argumentation schemes* [8] to guide search in the domain model for data and warrant needed to support a particular claim, and
- a *letter drafter* that organizes and expresses the arguments as English text using natural language generation techniques.

GAIL's expert arguments will be produced using a similar approach to the GenIE Assistant's domain models and argumentation engine. However, the natural language generation

module, the letter drafter, will not be needed to generate expert arguments.

The domain models in the GenIE Assistant are represented computationally as qualitative probabilistic networks (QPN) [9]. A QPN consists in part of a directed acyclic graph whose nodes are random variables. In addition, a QPN specifies qualitative constraints on variables in terms of influence (S^+, S^-), additive synergy (Y^+, Y^-), and product synergy (X^0, X^-) relations. For (Boolean) random variables A, B and C, $S^+(A,B)$ [or $S^-(A,B)$] can be paraphrased as *If A is true then it is more [less] likely that B is true*; $Y^+(\{A,C\},B)$ *[or* $Y(\{A,C\},B)$ as *If A and C are true then A enables [prevents] C from leading to B being true;* $X^0(\{A,C\},B)$*[or* $X^-(\{A,C\},B)$*]* as *if both [either] A and C are true then it is likely that B is true*.

To illustrate S^+, if a patient has two mutated BRCA1 alleles then it is more likely she will develop breast cancer; $Y+$, someone who has inherited a genetic mutation for familial hypercholesterolemia is at a higher risk of heart disease if she is obese; X^-, breast cancer can be caused by mutation of BRCA1 or some other gene; and X^0, together the mother and the father can pass an autosomal recessive mutation to their offspring. A QPN representing knowledge about a genetic condition can be reused for different patient cases. Representative domain models for testing the GenIE Assistant were built quickly using information from genetics reference books. The size of a QPN to be used in GAIL would be of the same scale as those used to generate letters in the GenIE Assistant (less than 50 nodes). For more information on domain modeling see [5].

Computational definitions of argumentation schemes are used by the GenIE Assistant's argumentation engine to construct a genetic counselor's arguments for the diagnosis and genotypes of family members [5]. The argumentation schemes are formalized in a structure including *claim*, *data*, and *warrant*. Since the argumentation engine and schemes do not encode domain-specific or patient case-specific content, they can be used to generate arguments in any domain whose domain knowledge can be represented in a similar format. The propositions used as claim or data describe states of variables in a QPN. The warrant expresses formal constraints on the nodes of the QPN in terms of influence and synergy relations mentioned above. The distinction between the two types of premises reflects their difference in function and source of information. Claims and data are facts or hypotheses about a particular case, whereas warrants describe (biomedical or other) generalizations.

In addition to those components, argumentation schemes in the GenIE Assistant include a field called the *applicability constraint*, a constraint that must be true to generate an argument from that scheme. Note that conclusions of the argumentation schemes are not necessarily deductively valid, and the *applicability constraint* is a type of critical question [8]. As discussed in section 5, the *critical questions* of GAIL's argumentation schemes provide a systematic means of challenging the conclusion of an argument.

To illustrate, consider an abductive reasoning scheme used in the GenIE Assistant:

Claim: $A \geq a$
Data: $B \geq b$
Warrant: $S^*(<A,a>, <B,b>)$
App. constraint: $\neg$ *exists* C $X(\{C,A\},<B,b>)$: $C \geq c$

In the above, uppercase-initial terms -- *A, B, C* -- are random variables in the QPN, S^* is a chain of one or more positive influence relations S^+. Lowercase-initial terms – *a, b, c* – are values of the random variables, and in this scheme are threshold values. To paraphrase this scheme, (warrant) there is a (chain of) possible positive causal influence(s) from *A* to *B*; (data) *B* is at least *b*; therefore (claim) *A* is at least *a*; (applicability constraint) provided that there is no *C* such that *C* and *A* are mutually exclusive positive influences on *B* and *C* is at least *c*. For example, (warrant) having a genotype with two mutated alleles of CFTR can lead to (abnormal CFTR protein which can lead to abnormal pancreas enzyme level which can lead to) growth failure; (data) this patient has growth failure; therefore (claim) this patient has cystic fibrosis; (applicability constraint) as long as there is no other condition believed to explain growth failure.

An argument for a given claim is automatically constructed by searching the domain model and data about the patient's case for information fitting GenIE's argumentation schemes instantiated with the claim. In addition to the above abductive argumentation scheme, other schemes support abductive reasoning about alternative causes or jointly necessary causes, reasoning from cause to effect, reasoning from negative evidence, and reasoning by elimination of alternatives. The argumentation schemes reflect those used in a corpus of genetic counselor-authored letters. Note that the GenIE Assistant's argumentation engine can construct complex arguments involving multiple pieces of evidence and chains of arguments. The same approach will be used in GAIL to generate expert arguments for a given claim. In a performance evaluation of the GenIE Assistant, two letters, each containing multiple arguments, were generated in 22 seconds on a desktop computer [5]. Note that the time should be less than that in GAIL, since the arguments will not be realized in English. Also, they can be generated off-line if necessary.

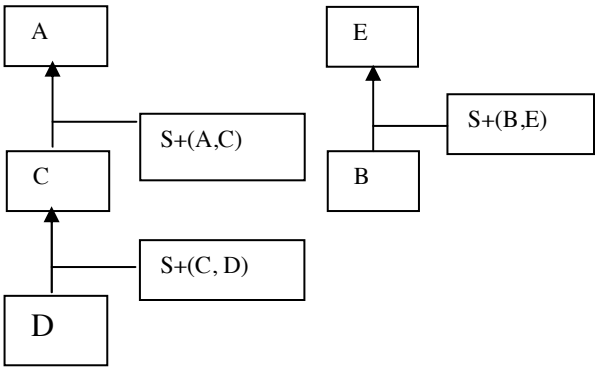


Fig. 2. Example of simple argument structures.

Some example arguments that can be generated are illustrated in Figures 2 and 3 in the box and arrow style of notation used in the GAIL interface. (To save space, the diagrams contain variables rather than the text that would be used in the GUI.) The diagram on the left of Figure 2 is a chain of two abductive arguments. The claim (A) that patient P has cystic fibrosis (two mutated CFTR alleles) is supported by the hypothesis (C) that P has abnormal CFTR protein and is warranted by the positive influence relation between CFTR alleles and CFTR protein. Hypothesis C is supported by the data (D) that P has frequent respiratory infections and the positive influence relation between CFTR protein and respiratory infections. The diagram on the right of

Figure 2 is a causal/predictive argument for the claim (E) that individual M (the patient's mother) is a carrier of a CFTR mutation. E is supported by the family history data that M has a certain ethnicity and is warranted by the higher probability of being a carrier if an individual has that ethnic background.

Figure 3 shows part of an argument for the claim (A=1) that P's mother has exactly one mutated CFTR allele. The left-hand subargument is for the hypothesis that she has one or two mutated CFTR alleles. That subargument is supported by the hypothesis (D=2) that P has cystic fibrosis (two mutated CFTR alleles), and is warranted by the synergy relation, X^0(<A=1,B=1>, D=2), i.e., that a child who has two mutated alleles inherited one from the mother and one from the father. Note that the claim D=2 would be supported by another subargument (not shown in Figure 3). The right-hand subargument is for the hypothesis that the mother does not have two mutated CFTR alleles. This is supported by the data (¬C) that she does not have cystic fibrosis symptoms, and warranted by the positive influence relation between CFTR alleles and symptoms of cystic fibrosis.

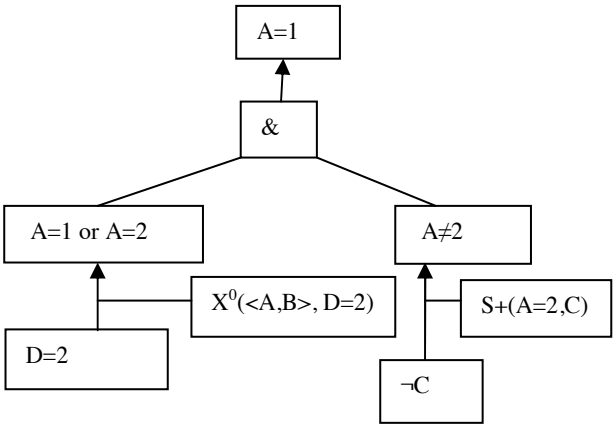


Fig. 3. Example of part of more complex argument.

3. PILOT STUDY

A formative evaluation of GAIL's prototype user interface was done in fall 2011 through spring 2012 with a total of 10 paid undergraduate volunteers, the first seven of which were recruited from biology classes and the last three computer science students. Each participant was first asked to read a seven-page patient education document, which we had found on the internet and printed for this study, on the inheritance and diagnosis of cystic fibrosis. After a participant read the document, it was put away and the research assistant narrated a silent video tutorial describing the components of an acceptable argument, and showing the features of the GAIL GUI and the process of constructing several different arguments using GAIL. Afterwards, the research assistant pointed out a chat box in the GAIL GUI for communicating with the assistant if necessary. The assistant then left the room, but could view the participant's computer screen on another computer monitor.

Listed in the upper left-hand corner of the GAIL GUI, the problems for which the first seven participants were asked to construct arguments are as follows.

Problem 1: Give two arguments for the diagnosis that J.B. has cystic fibrosis.

Problem 2: Give one argument for the diagnosis that J.B.'s brother has cystic fibrosis.

Problem 3: Give one argument against the diagnosis that J.B.'s brother has cystic fibrosis.

Problem 4: Give one argument for hypothesis that J.B.'s mother and father are both "carriers" of the CFTR gene mutation that causes cystic fibrosis

Note that the hypotheses, observations, generalizations (warrants), and problems shown on GAIL were written by the author of this paper based on information from a college genetics textbook. (J.B. refers to a fictitious patient.)

None of the first seven students created acceptable arguments. At that point in the study, it was decided to modify the materials and procedure. First, the problems were reduced in number (eliminating Problem 2, requiring an argument with conjunction). Second, when the participant submitted a response, the research assistant reviewed it using a checklist of error types created by the author after reviewing the arguments created by the first group of participants. If the participant's response contained any of those types of errors then the research assistant gave the participant feedback (as discussed below) through the chat box and asked the student to revise his argument. After three tries, the student was told to proceed to the next problem in the set. Third, to expedite the revised study, the remaining three students were recruited from computer science.

The distribution of error types is shown in Table 1. A Type 1 error was an argument whose claim did not match the claim for which the student was asked to give an argument. Type 2 was an argument where the data was not evidence for the claim. Type 3 was an argument where the warrant did not relate the data to the claim. Type 4 was an argument where the opposite type of link was required. Type 5 was a chained argument in which a subargument was missing or incorrect. For example, consider the chained argument on the left of Figure 2. If the learner failed to give a subargument in support of C, or if the learner skipped the intermediate conclusion C and showed D as directly supporting A, the error would be classified as Type 5. Type 6 errors involved incorrect use of conjunctions. Type 7 was omission of the warrant.

Table 1. Average number of errors per error type per person in each group

Error Type	Group 1	Group 2
1:Incorrect claim	1.9	0.8
2:Incorrect data	2.6	0.3
3:Incorrect warrant	2	1
4:Incorrect pro/con	0.9	0.3
5:Incorrect/missing chained claim	1.4	0
6:Incorrect/missing conjunction	0.9	NA
7: Missing warrant	0.1	0.4

In Table 1, Group 1 comprises the first seven students, who were given no feedback. Group 2 comprises the last three students, who were given feedback and three tries on each problem. The number of errors on each try for each student in Group 2 was totaled and the average was computed by dividing by nine (i.e., three students with three tries each). From the first group, it can be seen that the

most frequent errors (in descending frequency) were incorrect data, incorrect warrant, and incorrect claim. Although the quantity of errors in the first and second groups cannot be compared, it should be noted that the top three error types in Group 1 remained the top three in Group 2.

Group 2 received feedback from the research assistant based on the following guidelines:

1. Does the hypothesis match the problem? If not, tell the student that the hypothesis must match the problem.
2. Is everything OK except that the student has used Pro instead of Con or vice versa? If so, explain the difference.
3. Is the data relevant to the hypothesis (could you make a good argument using that data)? If not, suggest he/she try to use some other data.
4. Is the data relevant but the generalization (warrant) does not link the data to the hypothesis? If yes, suggest he/she try a generalization that links the two.
5. Is the generalization (warrant) relevant (could you make a good argument with it) but the data does not fit the warrant? If yes, suggest that he/she try different data that fits the warrant.
6. Did the student include some data in a conjunction that is unnecessary? If so, suggest that he/she remove the conjuncts that do not fit the warrant.
7. Did the student appear to skip a step in a chained argument that has a sub-argument for the data of the top argument? If yes, help the student break it into the main argument and the sub-argument.

Table 2 shows the types of errors made by the three students in Group 2 after receiving feedback on their first and second answers on each problem. Problem 1 was solved correctly by two students on the first try, and by the third student on the second try. Problems 2 and 3 were solved correctly by only one student (on the third try). Problem 3 was solved correctly by two students on the second try. These results suggest that on the more difficult problems (Problems 2 and 3), the feedback may have helped to reduce the number of errors.

Table 2. Types of errors in group 2 (after feedback).

Student	Try	Problem 1	Problem 2	Problem 3
1	1st		1, 3 ,4	2, 3
	2nd		1, 3	7
	3rd		3, 4	2, 7
2	1st	1	1, 3	1, 7
	2nd		1, 3	
	3rd		1	
3	1st		3, 4	2, 3, 7
	2nd		3	
	3rd			

At the end of the session, students were asked to complete a user experience survey. The survey results, shown in Table 3, indicate that the students had a favorable response to using the software despite making errors.

Table 3. Average scores on user experience survey (N=10). Possible responses: 3(True), 2(Somewhat true), 1(False).

Question	Score
My background … helped me answer the problems in this study.	2.3
I found the subject of genetic conditions and inheritance interesting.	3
I found the tools for diagramming arguments easy to use.	2.8
I found the tutorial on how to use the argument diagramming tools helpful.	3
I prefer using the argument diagramming tools to writing arguments.	2.7
I would like to use a program like this in my courses on genetics	2.9

4. ARGUMENT ANALYZER

The expert model will contain all acceptable arguments that can be automatically generated for a given claim from an underlying knowledge base (KB) representing the problem domain. The generated arguments are simple or complex argument structures containing KB elements. Text elements provided to the learner through GAIL's GUI are linked internally to KB elements. The inputs to GAIL's argument analyzer will be the learner's argument and the expert model, both in the same format. Implemented in Prolog, the prototype argument analyzer determines if a student's argument diagram represents an acceptable argument and if not acceptable, identifies its weaknesses.

The algorithm to determine acceptability merely checks whether the user's argument matches one of the acceptable arguments. If the user's argument does not match an acceptable argument, its weaknesses are identified using pattern-matching rules motivated mainly by the types of errors seen in the study described in the previous section. The rules are non-domain-specific and non-problem-specific. For example, if the user's data and claim match the expert's, but the warrant does not, the analyzer identifies the problem as an unacceptable warrant (Type 3). The prototype argument analyzer implementation outputs an error message for each error detected. However, in the future implementation of GAIL, the argument analyzer's output would be used by the Feedback Generator, which will be responsible for selecting which error(s) to highlight and providing appropriate feedback.

If the learner's argument is acceptable, i.e., it matches an expert argument, then knowledge of the argumentation scheme used to generate the expert argument provides an additional resource for generation of feedback as described in the next section.

5. FEEDBACK GENERATOR

The feedback generator has not been implemented yet. Currently, we are gathering information to guide its design. As discussed in the previous section, the feedback generator will have access to the output of the argument analyzer. If the learner's argument contains errors such as those types listed in Table 1, some design questions are: which of the errors to address (and in what order), when to provide feedback, what feedback content to provide, and in what syntactic form. Before designing a feedback generator that

answers these questions, we are running a think-aloud study to get a better understanding of why students make these errors. For example, a type 4 error might be due to a misunderstanding of the argument representation used in GAIL's GUI. If that is indeed the case, then it would seem that addressing such an error should be given higher priority by the feedback generator. On the other hand, we hypothesize that a type 1, 2 or 3 error may be due to a deeper problem, either in the learner's understanding of what constitutes an acceptable argument, or in understanding the genetics information provided by GAIL as possible building blocks for the learner's argument diagram.

A key point to note is that our approach supports content-based feedback. Many of the types of errors listed in Table 1 are content-based errors that can be detected by the argument analyzer based on the expert model. In addition to using it to identify content-based errors, GAIL will be able to use the expert model to provide content-based feedback. This is illustrated in the following imaginary scenario. Figure 4 depicts abstractly a student argument diagram in which the data, B, is not related by the warrant, S+(A,C), to the conclusion A. Our approach supports providing feedback to the effect that this argument is not acceptable because the warrant does not relate the data to the conclusion; and supports giving the advice to look for other data that is consistent with the given warrant or to look for another warrant that links the given data to the conclusion. Suppose that the expert model contains an argument similar to that in Fig. 4, but using C as data. If the student is unable to make use of the more general advice to replace the data or warrant in the diagram, a hint could then be generated asking whether C is in the observations or hypotheses on the GUI screen.

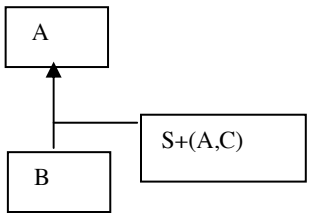


Fig. 4. Abstract example of unacceptable argument.

Figure 5 shows that with the help of this feedback, the imaginary student has replaced the data in the argument diagram with C. However, suppose that C was listed on the GUI screen as a hypothesis rather than an observation. In that case, a sub-argument for C would be required. The argument analyzer could recognize that the sub-argument for C in the expert model is missing in the student's diagram. Then the feedback generator could inform the student that C must be supported by a sub-argument since it is only a hypothesis.

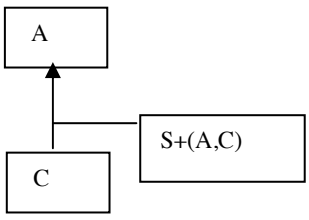


Fig. 5. Abstract example of partly fixed, unacceptable argument.

Figure 6 shows that with the help of this feedback the student adds a sub-argument for C to the diagram, matching an acceptable expert argument.

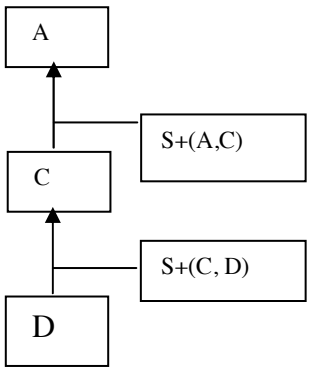


Fig. 6. Abstract example of acceptable argument.

In this domain, however, the conclusions of acceptable arguments are not necessarily deductively valid. As discussed in Section 2, each abstract argumentation scheme is associated with certain critical questions, which provide a way of challenging an argument constructed from that scheme. Critical questions support a different type of feedback, which could inspire a learner to consider multiple arguments pro and con the same claim. To illustrate, one of the critical questions of the abductive argumentation scheme is whether there is another plausible explanation of a certain observation. Having recognized the learner's argument as an instance of this scheme, the feedback generator could instantiate this critical question. Suppose that the learner has constructed an acceptable abductive argument for a diagnosis of cystic fibrosis; instantiating this critical question could support generating feedback such as *Can you make an argument for an alternative diagnosis that explains the patient's frequent respiratory infections?* or, *What if he has some other condition that could explain those symptoms?*

Some other critical questions of GAIL's abductive argumentation schemes, where B is an observation and A is a putative cause of B, include (Green 2010):

- **(Missing Enabler)** is there a C such that C is required for A to cause B, and C is absent? (Example: *Has exposure to bacteria occurred, which is required for thickened mucous to lead to frequent respiratory infections?*)
- **(Mitigation)** is there a C whose presence may mitigate the effect of A on B? (Example: *Is the patient taking antibiotics, which will prevent respiratory infections?)*
- **(Inapplicable Warrant)** Despite the similarity of individual I to the population described by the warrant, is there is a difference that could make it inapplicable to I? (Example: *Although the mother is from a geographic region with a high rate of cystic fibrosis, is her ethnic background different from most of the population there?*)
- **(False Positive)** Is p(¬A | B) too high? (Example: *Is the false positive rate for the laboratory test used to diagnose this condition high?*)
- **(Low Certainty of Data)** Is p(B) too low? (Example: *Are we confident that there is accurate information about the health of the biological mother who gave the patient up for adoption when he was an infant?*)

Again note that feedback can be given without requiring problem-specific knowledge to be embedded in the feedback generator. Also note that semantic, not syntactic, forms of critical questions are associated with argumentation schemes. Thus, using natural language generation from semantic forms to generate syntactic variations, one could study the varying effectiveness of different ways of asking the same critical question.

6. CONCLUSIONS

This paper presents our ideas on generating formative feedback in the Genetics Argumentation Inquiry Learning (GAIL) system. GAIL will provide learners with tools for constructing Toulmin-style arguments in diagrams using blocks of text provided by the system. The text is linked internally to KB elements. An argument generator like one previously developed for another application will use the KB and abstract argumentation schemes to automatically generate expert arguments. GAIL's argument analyzer will determine if a learner's argument is acceptable by comparing it to the expert arguments. A prototype argument analyzer has been implemented using non-domain-specific, non-content-specific rules that recognize common error types. The error types are based on those observed in a pilot study. GAIL's formative feedback generator will use the argument analyzer's output. In addition to identifying problems in the learner's argument, if the argument is acceptable the analyzer will inform the feedback generator of critical questions of the argumentation scheme underlying the student's argument. The critical questions can be used to generate feedback stimulating the learner's critical thinking.

7. ACKNOWLEDGMENTS

Graduate students B. Wyatt and C. Martensen implemented the prototype of GAIL's GUI in summer 2011, and Martensen ran the user study in fall 2011 through spring 2012; both received support from a UNCG Faculty Research Grant. We would like to thank the reviewers as well for asking many interesting questions that we have tried to address in the camera-ready version of this paper or would like to address in future work.

8. REFERENCES

[1] Kirschner et al. 2003. Kirschner, P.A., Buckingham Shum, S.J., and Carr, C.S. (Eds.) 2003. *Visualizing Argumentation*. London: Springer.

[2] Scheuer et al. 2010. Scheuer, O., Loll, F., Pinkwart, N., and McLaren, B.M. 2010. Computer-Supported Argumentation: A Review of the State of the Art. *Computer-Supported Collaborative Learning* 5(1): 43-102.

[3] Pinkwart and McLaren 2012. Pinkwart, N., McLaren, B.M. (Eds.) 2012. *Educational Technologies for Teaching Argumentation Skills*. Sharjah: Bentham Science Publishers

[4] Toulmin, S. 1998. Toulmin, S.E. 1998. *The uses of argument*, Cambridge: Cambridge University Press.

[5] Green, N., R. Dwight, K. Navoraphan, and B. Stadler. 2011. Natural language generation of transparent arguments for lay audiences. *Argument and Computation*, 2(1): 23-50.

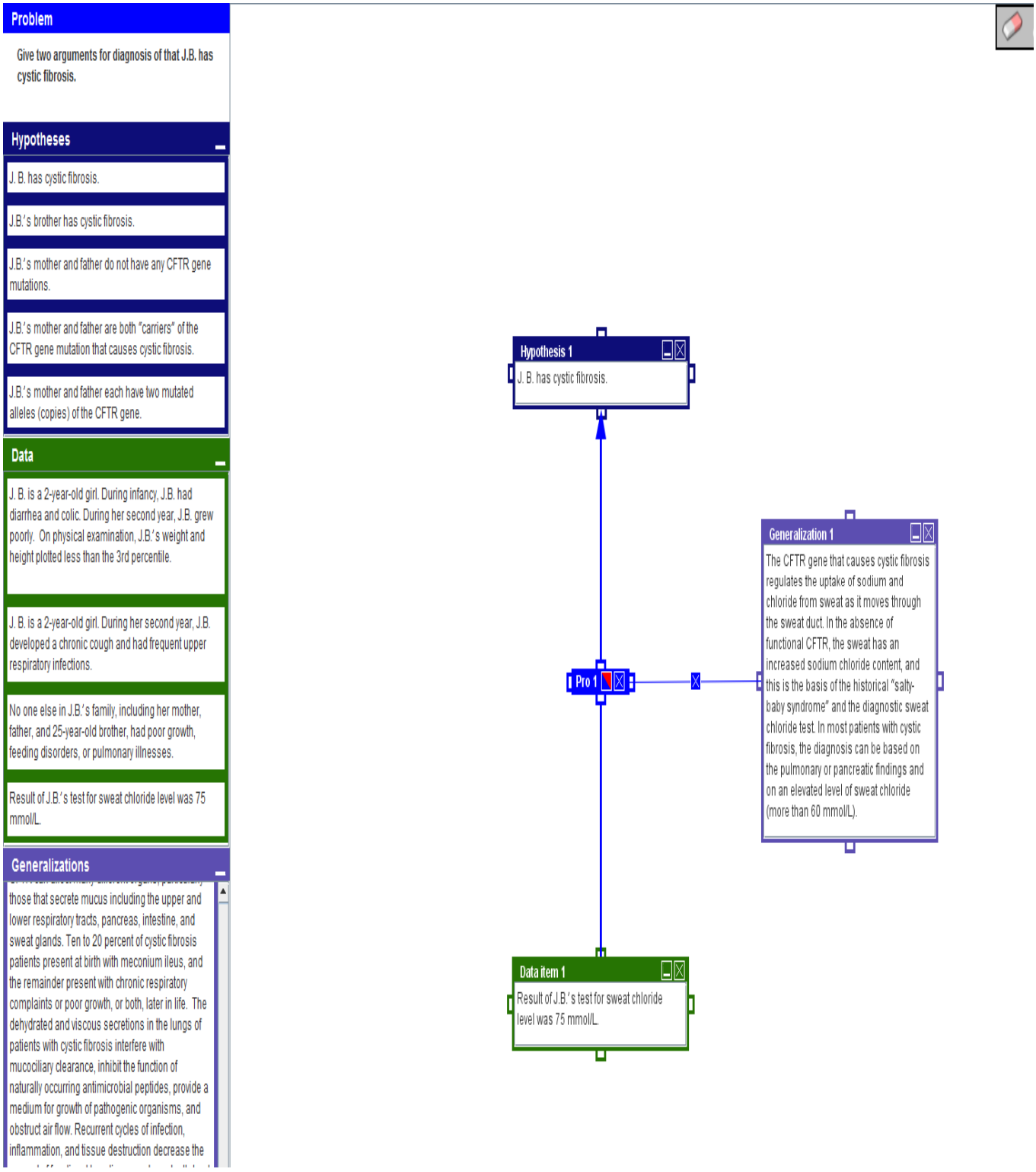
[6] Scheuer et al. 2012. Scheuer, O., McLaren, B.M., Loll, F., Pinkwart, N. 2012. Automated Analysis and Feedback Techniques to Support and Teach Argumentation: A Survey. In Pinkwart and McLaren (Eds.) 2012. *Educational Technologies for Teaching Argumentation Skills*.

[7] Green, N. 2005 A Bayesian network coding scheme for annotating biomedical information presented to genetic counseling clients. *Journal of Biomedical Informatics* 38, 130-144.

[8] Walton et al. 2008. Walton, D., C. Reed, and F. Macagno. 2008. *Argumentation Schemes*, Cambridge: Cambridge University Press.

[9] Druzdzel and Henrion 1993. Druzdzel, M. J., and Henrion, M. 1993. Efficient Reasoning in Qualitative Probabilistic Networks. In *Proceedings of the 11th National Conference on AI*, 548-553. Washington, DC.

[10] Green, N. 2010. Towards intelligent learning environments for scientific argumentation. In *Workshop on Ill-defined Problems and Ill-defined Domains, Intelligent Tutoring Systems 2010* (Pittsburgh, PA).



.

Fig. 1. Screen shot of GAIL prototype user interface in formative evaluation of fall 2011 – spring 2012.

Formative feedback in *Digital Lofts*: Learning environments for real world innovation

Matthew W. Easterday
easterday@northwestern.edu

Daniel Rees-Lewis
daniel2011@u.northwestern.edu

Elizabeth Gerber
egerber@northwestern.edu

Northwestern University, Evanston, IL, 60208

ABSTRACT

C*ivic innovators* design real-world solutions to societal problems. Teaching civic innovation presents serious challenges in *classroom orchestration* because facilitators must manage a complex learning environment (which may include community partners, open-ended problems and long time scales) and cannot rely on traditional classroom orchestration techniques (such as fixed schedules, pre-selected topics and simplified problems). Here we consider how *digital lofts--online learning environments for civic innovation* might overcome orchestration challenges through the use of badges, cases, crowd-feedback, semi-automatically created instruction, self-assessment triggered group instruction, social media, and credentialing. Together these features create three types of feedback loops: a crowd critique loop in which learners receive formative feedback on their innovation work from a broader community, a case development loop in which examples of student work are semi-automatically created to provide instruction, and a learner-driven instructional loop, in which self-assessments determine which group instruction is provided. Researching and developing digital lofts will help us to understand how to support real-world innovation across design disciplines such as engineering, policy, writing and even science; and result in technologies for disseminating and scaling civic innovation education more broadly.

Keywords

Digital lofts, feedback, civic innovation, online learning environments

1. INTRODUCTION

Many of the challenges facing our society such as global warming, poverty, and illiteracy are *political* problems that cannot be solved through engineering alone. For example, to create environmentally sustainable cities we would have to train engineers to redesign the land, water, energy and information systems of the city. And while we do train engineers to design membrane filtration-, renewable energy-, and mass transit-systems, we do not teach them about changing economic policy to promote conservation, energy initiatives to discourage fossil fuel use, or zoning rules to encourage mass transit. We do teach engineers about complex mechanical systems and how to communicate effectively as a team, but we don't teach them that sustainable infrastructure might also require changes in policy. Even when we do teach them about policy, we don't teach them how to change it, and even if they did know how to change it, they can't change it alone, leaving us with engineers who are at the mercy of policy problems, not ones that can solve them. In short, good technology and bad policy means no impact (Easterday, 2012).

To overcome societal challenges, we must train ***civic innovators*** who can identify, design and engineer solutions to societal problems. Civic innovators must be able to develop, modify, and implement ideas while navigating ambiguous problem contexts, overcoming setbacks, and persisting through uncertainty in their community. To become civic innovators, learners must gain experience identifying and tackling complex, ill-structured design challenges that are not easily solved within a fixed time frame. Civic innovation education is thus a kind of *service learning* that *"*...integrates meaningful community service with instruction and reflection to enrich the learning experience, teach civic responsibility, and strengthen communities..." (ETR Associates, 2012). However, unlike other forms of service learning, civic innovation focuses on *design*--whereas service learning might ask students to pick up trash in a riverbed to motivate learning about ecology, civic innovation might ask students to pick up trash in a riverbed to motivate learning about ecology in order to identify, design, and engineer solutions to reduce environmental pollution.

But embedding learning in real-world activities makes civic innovation difficult to teach: individual mentoring can be effective but expensive; extra-curricular environments provide flexibility but insufficient guidance; and classroom instruction is too rigid and time-bound for solving complex societal problems. Embedding learning in real-world activities creates a serious challenge of ***classroom orchestration.*** Classroom orchestration (Dillenbourg & Jermann, 2010) involves satisfying the constraints of curriculum, assessment, time, energy, space, etc. required to promote learning in a given context. Embedding learning in the real-world increases the orchestration challenge because orchestration techniques that work in the classroom (such as using simple problems, making students complete assignments at the same pace) can't be used when learners are working on real-world problems. Adding community clients and professional design mentors only makes orchestration more challenging.

New cyberlearning technologies, such as web 2.0, social media, reputation systems, and crowdsourcing offer new ways to orchestrate learning environments for civic innovation. Just as we create instructional *lab*s to teach science, the purpose of this project is to develop instructional *lofts* to teach innovation. Our research question is: ***how might we create Digital Lofts: on-line, learning platforms for teaching civic innovation that overcome the orchestration challenge?***

Knowing how to design digital lofts that overcomes the orchestration challenge will allow us to amplify teaching resources to make civic innovation education feasible. Design principles for Digital Lofts would allow us to overcome orchestration challenge not just for civic innovation education, but for project-based learning environments as well, allowing us to design learning environments that are more sustainable, more easily scaled to new contexts, and more like real life.

2. BACKGROUND

Advantages of civic innovation learning communities

What do civic innovation learning environments look like? Civic innovation learning communities: (a) have pro-public missions, (b) teach learners how to design solutions to real problems, (c) are led by learners and supported by faculty and professional experts, and (d) extend nationally through a network of chapters. For example, in *GlobeMed,* students work on international health challenges. In *Engineers for a Sustainable World,* students work on projects that promote environmental, economic, and social sustainability. It is important to stress the pro-public mission of these learning communities. Learners are tackling problems that require them to address societal challenges and to understand policy issues. For example, by tackling the problem of energy sustainability, students are forced to consider the environmental, economic and legal policies that constrain the effectiveness of technological interventions. For this project, we consider *Design for America*, which provides an ideal model of a learning community for civic innovation.

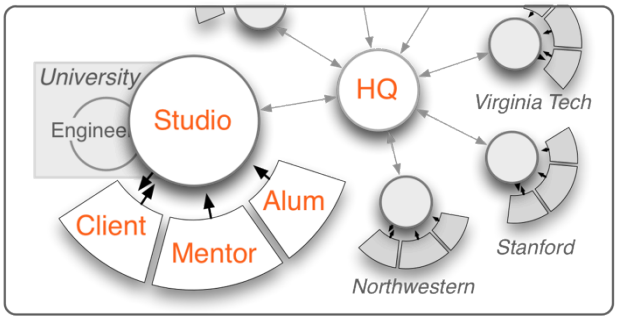


Figure 1*. Design for America's community of practice. The 14 studios are hosted on University campuses and interact with, but do not replace the existing curricula. Studios incorporate local clients, mentors and alumni and communicate directly with DFA Headquarters.*

Design for America (DFA) is a learner-directed, extracurricular service-learning environment that is succeeding at developing civic innovators. Universities host on-campus DFA studios in which student teams work on self-selected civic innovation projects throughout the academic year, applying the skills and expertise they've gained through academic coursework (Figure 1 & 2). Student teams identify challenges in healthcare, environment, and education in their local community such as reducing hospital-acquired infections and reducing water waste in cafeterias. They work with organizational partners to: understand stakeholder needs, ideate, prototype, test, and implement solutions. During the annual 4-day *Leadership Studio*, experienced student leaders train new student leaders in studio management and leadership.

Design for America was conceived by co-author Gerber during the 2008 presidential election to engage university students in solving civic issues using human-centered design. As an assistant professor of design, Gerber joined student co-founders Mert Iseri, Yuri Malina, and Hannah Chung, to start the first studio at Northwestern University. Currently, there are 14 studios hosted by universities throughout the country (including Stanford, Virginia Tech, and Northwestern) involving 1800 students (58% women), aged 18-30 from over 60 majors, working on over 50 projects; 15 faculty mentors; and 80 professional mentors. And the number of studios is expected to grow to 30 by 2015. In just four years, DFA has produced two start-ups that have raised over $1.5 million in funding. DFA has been featured in Fast Company, Oprah, and the Chicago Tribune.

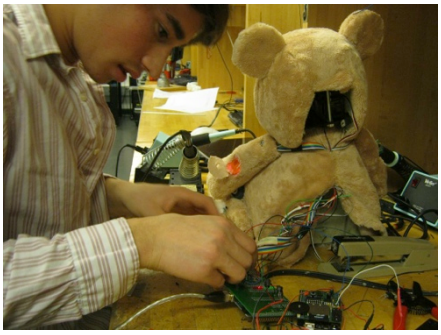


Figure 2*: Design for America students learn civic innovation through projects that require designing, building, and implementing solutions.*

Findings from surveys, daily diaries, interviews, and observations suggest that DFA students develop confidence in their ability to act as civic innovators through successful task completion, social persuasion, and vicarious learning in communities of practice with clients, peers, industry professionals, and faculty. Furthermore, students attribute achievement of learning outcomes outlined by the Accreditation Board for Engineering and Technology including identifying, formulating, and solving problems; functioning on a multidisciplinary team; communicating effectively; and knowledge of contemporary issues to their participation in Design for America. (Gerber, Marie Olson, & Komarek, 2012); (ABET Engineering Accreditation Commission, 2011).

Design for America's civic innovation model follows many recommendations of the learning sciences for improving motivation and transfer such as using real world problems that require design of meaningful products with social relevance. DFA encourages students to work on authentic problems (Shaffer & Resnick, 1999) to motivate learning and transfer. Students identify and select projects and self-direct the innovation and discovery process including observation, idea generation, prototyping, and testing (Kolodner, Crismond, Gray, Holbrook, & Puntambekar, 1998); (Puntambekar & Kolodner, 2005). By trying to apply their knowledge to a problem, students come to understand what they know and when they need more information (Edelson, 2001). Like service learning (Furco, 1996), DFA increases civic awareness, interest in the real needs of people, and contemporary issues by focusing on innovating solutions to local community challenges (Gerber et al., 2012).

Unlike traditional classrooms, Design for America's community of practice (Figure 1) expands beyond the physical boundaries of the student community to include experienced, local professionals, local clients and community members, as well as beyond the temporal boundaries of student life as learners continue to participate in projects as alumni. Students' involvement in a community of practice (Lave & Wenger, 1991) includes engaging with peer mentors, professionals and faculty in a non-evaluative environment over an extended timeframe. Communities of practice foster innovation self-efficacy (i.e., learners' belief in their ability to innovate, (Gerber et al., 2012) and such beliefs influence goal setting, effort, persistence, learning and attribution of failure (Bandura, 1997); (Deci & Ryan, 1987); (Ryan & Deci, 2000). Students select real world

challenges (Shaffer & Resnick, 1999) that are personally meaningful, build and test solutions to problems, and share their work with the community through review sessions (Papert & Harel, 1991); (Papert, 1980); (Resnick, 2009); (Kolodner, Owensby, & Guzdial, 2004). Because DFA projects are extracurricular, they conclude when ideas are implemented, rather than when the academic term ends.

Orchestration challenges in civic innovation learning communities

While learning environments for civic innovation have many potential advantages, they also face many challenges. Civic innovation teachers face serious orchestration challenges because they have to teach many different project teams, with different levels of expertise, working on different problems for different community clients. The orchestration challenge makes civic innovation difficult to teach well.

Like many extra-curricular organizations, DFA students often suffer from a lack of guidance. Our needs analysis of Design for America found that, unsurprisingly, learners would benefit from more scaffolding and feedback on the innovation process including: (a) planning and conducting research on their project challenge; (b) using initial research to inform proposed solutions; (c) selecting and conducting appropriate design activities for their project challenge; and (d) discounting initial solutions if these solutions prove not to be viable. While DFA has been very successful at attracting learners, these learners report that frustrations from lack of progress makes them question their commitment to the work they are undertaking. And while leaders (student facilitators) experienced in project work and trained at the DFA leadership studio require less support, they find helping other students very challenging. In interviews, these student leaders asked for more granular 'how to' guides from DFA headquarters.

DFA students also often struggle to access available resources that could help them in their projects. While students are aware that they can reach out to experts within the DFA network generally, they struggled to identify specific individuals or instructional resources that can help them. Learners often fail to ask for support from more experienced members of the community because they don't know whom or for what to ask. Similarly, learners find it challenging to locate helpful instruction. They report floundering for long periods of time trying to find resources and as well as not knowing where to start looking.

In fact, these issues are challenges in project-based learning and criticisms of minimally guided instruction in general. Without sufficient guidance, learners become lost, confused and frustrated, which can lead to misconceptions (Kirschner, Sweller, & Clark, 2006); (Hardiman, Pollatsek, & Well, 1986); (Brown & Campione, 1996). Furthermore, students often need to develop additional help-seeking skills in order to learn effectively (Gall, 1981; Pintrich, 2004); (Ryan, Pintrich, & Midgley, 2001). Learning science provides myriad ways to offer guidance such as providing explanations, worked examples, process worksheets, prompts, (and many more) (Scardamalia, Bereiter, & Steinbach, 1984); (Reiser, 2004); (Edelson, Gordin, & Pea, 1999); (Puntambekar & Kolodner, 2005); (Kolodner et al., 2004).

Note that we do not wish to re-litigate the discovery vs. direct instructional debate here--achieving the proper balance between providing and withholding assistance (a.k.a the assistance dilemma) remains a fundamental and enduring question in the learning sciences (Koedinger & Aleven, 2007). Our point is merely that civic innovation facilitators cannot effectively deliver *any* instructional model (constructionist, direct, or otherwise) because they cannot effectively orchestrate learning at DFA studios. In other words, we cannot answer the fundamental questions about civic innovation without addressing orchestration.

The need for new orchestration technologies

In a typical classroom, orchestration is relatively easy. But the traditional classroom approaches to orchestration don't work for civic innovation. For example, to make classroom teaching easier, we often give students identical, simplified problems (in the words of one DFA student: "well-defined problems on a platter.") We use schedules that keep learners moving at the same pace so we can teach the same skills and knowledge to the whole class. This is an easy way to orchestrate groups of learners when we have a limited set of teaching resources.

Unfortunately, when we use simplified, artificial problems, we don't give students a chance to practice the skills for coping with design complexity we want them to learn. We also destroy the motivational benefits that come from working on real world problems. For example, if we want students to practice "scoping," (i.e., identifying important but tractable problems to solve) then we need to give them ill-defined problems that can be scoped in different ways and that may not fit neatly into the academic calendar. If we want them to practice communicating with clients, then we must accept unclear and changing project goals. If we want to take advantage of students' intrinsic motivation to address real world problems on topics they feel are important, then we must accept a certain level idiosyncrasy of projects. But once we start letting different groups work on different, more complex problems, at different speeds, working with clients in the community, and so on, it becomes almost impossible for a single teacher to orchestrate learning in a productive way.

Could technology help teacher orchestrate civic innovation learning environments? Existing online learning management platforms do not address the orchestration problem. Many of the most popular general-purpose online platforms assume a classroom model and are designed for distributing online books or lectures, such as academic platforms like the Open Learning Initiative (Lovett, Meyer, & Thille, 2008), MIT Opencourseware (Massachusetts Institute of Technology, 2012), and Coursera (Severance, 2012), which do not help us orchestrate design projects. Other technologies provide no pedagogical help but rather tools for managing files and conversations, such as Blackboard (Blackboard Inc., 2012), Canvas (Canvas, 2012), Lore (Lore, 2012), and Sakai (Sakai Foundation, 2012). Some technologies for orchestration focus on only small portions of the challenge such as managing a single activity (Dillenbourg & Jermann, 2010). And while there has been great progress in technologies for orchestrating scientific inquiry (Peters & Slotta, 2010), such as BioKIDS (Songer, 2006), BGuILE (Reiser et al., 2001); (Sandoval & Reiser, 2004), Inquiry Island (White et al., 2002), KIE (Bell, Davis, & Linn, 1995), and WISE (Slotta, 2004), these platforms are not appropriate for teaching civic innovation.

Solving the orchestration challenge is not simply another application of technology to teaching, it is absolutely essential for creating the civic innovation learning environments urgently needed to prepare learners for the societal challenges that await them.

3. TECHNOLOGICAL INNOVATION

Orchestration of civic innovation is difficult because there are too many moving pieces: different learners, with different abilities,

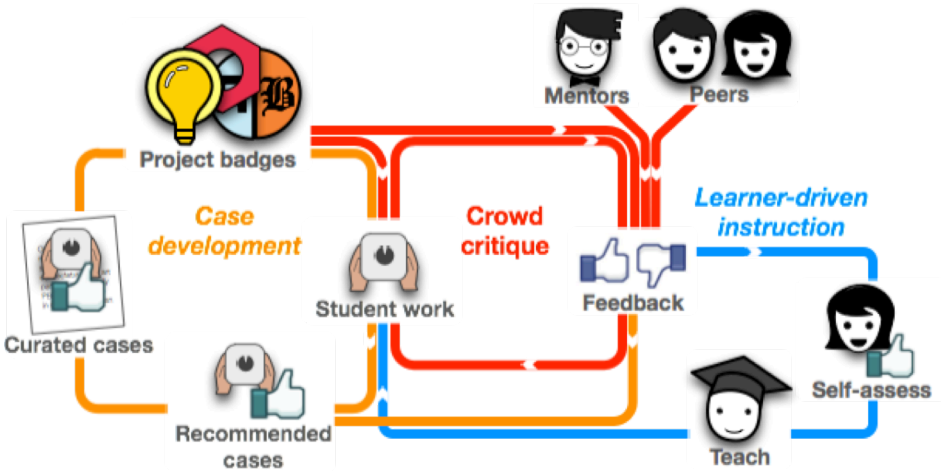


Figure 3. *Digital Lofts merge curriculum and data in three integrated feedback loops: the crowd-critique loop, the case development loop and the learner-driven instructional loop.*

working on different (complex) design problems, at different speeds, with different community clients. We could solve the orchestration challenge by giving each project team it's own professional design teacher but doing so is costly. However, with new technologies like web 2.0, crowdsourcing, and social media, we may be able to reduce the orchestration challenge for teachers and give them additional resources to overcome it. Specifically: we can use web 2.0 to scaffold the innovation process and provide flipped, just-in-time instruction relevant to students' current goals; we can use crowd-feedback to provide learners with more frequent, higher quality feedback on their progress; we can use recommender systems to semi-automatically create case libraries of successful designs; and we can automatically monitor group progress so teachers can give the right instruction to the right group at the right time.

Design hypothesis. Our initial design hypothesis argues that we can teach civic innovation by using what we call ***Digital Lofts*** to overcome the orchestration challenge. Digital Lofts are online learning platforms for support learning in real world contexts that:

1. use badges to scaffold the innovation process,
2. provide a student-generated and curated case-library linked to badges to teach design,
3. use crowd-feedback to increase the frequency and quality of feedback,
4. use recommender systems to semi-automatically create case-based instructional material,
5. use self-assessment to trigger maximally relevant group instruction,
6. use social media to facilitate participation and support, and
7. use recognition and credentialing to facilitate help-seeking and connections to resources.

These features allow us to create a curriculum that dynamically adapts to the needs of the learner, that is, to merge curriculum and data. By merging curriculum+data, we can reduce the challenge of orchestrating civic innovation to a manageable level.

To understand how Lofts help us orchestrate civic innovation, we can think of Lofts as supporting 3 interrelated feedback loops: (a) **a crowd-critique loop** in which students receive feedback on their work through project critiques, (b) a **case development loop** in which student work is used to semi-automatically create case studies of successful and unsuccessful designs which are then used to teach design principles, and (c) a **learner-driven instructional loop** in which students' self assessments trigger face-to-face group instruction taught by facilitators (Figure 3).

The crowd-critique loop

Designers and engineers often organize their work according to an innovation process. Figure 4 shows the high level steps or goals of a simplified innovation process consistent with the processes used by leading design and engineering firms like IDEO and Cooper (Dubberly, 2005) by the Stanford *d.school* (Beckman & Barry, 2007) or defined in engineering education standards (Massachusetts Department of Education, 2006). In Figure 4, the first stage of design is to "focus" by identifying a potential topic to address such as "water conservation at universities." The second stage is to "immerse" or study the user-needs, constraints and technologies involved in the issue. The third stage is to "define" a specific problem that can be solved, such as "reduce water use in the college cafeteria by 30%." The fourth stage is to "ideate" by generating a wide range of potential solutions. The fifth stage is to "build" the design using sketches, prototypes and high-fidelity implementations that realize the design idea. The sixth stage is to "test" the design. Even in simplified models like that in Figure 4, the design process is applied in an iterative and non-linear manner.

Figure 4. *Badges scaffold complex design processes for the novice into smaller, more manageable challenges and identify members who have passed the challenges as potential mentors.*

Design can be thought of as a process of learning (Beckman & Barry, 2007); (Owen, 1998). Designers construct new knowledge through observations that yield insights; insights support frameworks that inspire ideas that lead to innovative solutions (Beckman & Barry, 2007). Through this process, people construct knowledge (Dong, 2005), moving back and forth from the analytic phase of design, which focuses on finding and

discovery, to the synthetic phase, which focuses on invention and making (Owen, 1998). Beckman and Barry (2007) describe knowledge creation through the design process as movement between concrete experiences and abstract conceptualization, reflective observation, and active experimentation. Inductive and deductive practices support the construction of new knowledge that designers use to shape the environment in ways that did not previously exist.

So how can teachers guide design groups working on different, complex problems? One of the most important ways to promote learning is to provide learners with scaffolding and feedback on their work.

The Loft's crowd-critique loop scaffolds the design process and provides feedback using project critiques. The crowd-critique loop starts with **project badges** (like girl scout badges) that break the complex design process into a series of manageable mini-challenges (Figure 4). For example, for the *focus* badge, learners have to scope an important but tractable issue such as *hospital acquired infections*; for the *immerse* badge, learners have to conduct user-research on their target population to better understand their needs. In the second step of the crowd-critique loop, learners use the resources attached to each badge to help them solve the challenge--each badge is linked to *flipped* (*blended)* instructional material (Khan, 2012); (Lovett et al., 2008) that includes resources, principles, and examples that can help the learners solve the design challenge. For example, the "build" badge for a web design project might include a video lecture on writing html, an interactive javascript tutorial, on-line readings about web-design principles, or examples of the different stages of creating a well-designed website In the final step of the crowd-critique loop, (after students have worked on a badge and submitted their work to the Loft), the Loft solicits feedback on students' work from professional design mentors and peers who have previously completed the badge. The mentors and peers use the badge assessment rubrics to provide feedback to students.

The widespread use of badges in online games has led to a surge of interest in badges for learning (Duncan, 2011). However, civic innovation students are already intrinsically motivated to work on real world design problems, so it doesn't make sense to use badges as extrinsic rewards that might decrease motivation (Deci, Koestner, & Ryan, 1999) and encourage gaming the system (Kraut & Resnick, 2012). So instead, Lofts use badges to scaffold the design process and communicate learning goals, which should increase learning (Ambrose, Bridges, DiPietro, Lovett, & Norman, 2010).

Combining flipped instruction with face-to-face teaching can be more effective than face-to-face teaching or online-only teaching alone (Scheines, Leinhardt, Smith, & Cho, 2005); (Lovett et al., 2008). Our flipped instructional material will use a guided-experiential learning approach shown to improve learning outcomes relative to traditional project-based learning (Velmahos et al., 2004); (Clark, 2004/2008).

Providing high quality feedback to learners is one of the most effective ways to increase learning (Hattie, 2009); (Hattie & Timperley, 2007); (Ambrose et al., 2010). The Loft provides learners with two underutilized sources of feedback: professional mentors and peers. Giving peers well-designed assessment rubrics can make their feedback as effective as instructor feedback (Sadler & Good, 2006). The Loft thus uses ***crowd-feedback*** to increase the frequency and quality of feedback available to learners.

But what if students refuse to submit work or mentors and peers refuse to review it (Kraut & Resnick, 2012)? Our needs analysis found that DFA students are hungry for feedback on their project and very willing to submit work to get this feedback. Professional design mentors are also very willing to provide this feedback assuming that students 'drive' the process by providing them with well-prepared material from their design process (which the badges help students to do).

The case development loop

Developing useful learning resources can be a challenging task especially with design teams that may all be pursuing different directions at different times--how can cyberlearning technologies help produce effective and engaging learning resources?

Our needs analysis found that DFA students prefer to share design lessons through stories about how they created their designs and how well those designs worked. In the learning sciences, this falls under the heading of *case-based reasoning,* where each story describes an example or case of a design that worked (or didn't work) along with an explanation of the key features that led the result, in which context, and so on. Teaching effectively with cases has been well studied in several forms, including learning from cases (Kolodner*,* 1993; 1997), analogies (Gentner, Loewenstein, & Thompson, 2003), and worked-examples (Ward & Sweller, 1990; Salden, Aleven, Renkl, & Schwonke, 2009).

Unfortunately, DFA students' learning from cases suffers many limitations: (a) it is done informally, so knowledge of particular cases is not spread widely; (b) students do not effectively teach with cases, sometimes hiding illustrative mistakes, promote their projects rather than teaching, and failing to highlight the key design lesson or principle; and (c) students do not present contrasting cases that would allow learners to understand the deep features and the context of applicability of a case. Such knowledge sharing is typical of large distributed organizations (Argote, 1999).

Furthermore, it is difficult to create case-based teaching material both in terms of creating a useful library of cases and in creating ways for learners to find the appropriate case when needed (Kolodner, 1997).

Digital Lofts overcome this challenge through a case development loop. In the ***case development loop***, the Loft uses assessments of students' work to semi-automatically create *case libraries*--examples of student work that include reflections about what worked, what didn't, in what context. First, the crowd-feedback from the crowd-critique loop is used to recommend particularly successful and unsuccessful examples of each design step, producing sets of contrasting cases. Second, an instructional designer creates *curated cases* by selecting cases that best illustrate key design principles. The instructional designer then refines these cases. Finally, the contrasting cases are then presented as an instructional resources linked to each badge.

The crowd-feedback and badging systems of the Loft reduce the orchestration challenge of providing relevant and engaging instruction to a manageable level in several ways. First, the Loft continually collects student work from multiple campuses, so we get the initial material for the case library "for free" using crowdsourcing, or production of work by a distributed crowd of people (Von Ahn & Dabbish, 2004). Second, project critiques act as a *recommender system* (Kiesler, Kraut, Resnick, & Kittur, 2012) sorting student work into contrasting cases. Third, cases are already linked to particular phases of the design process

through the badges, so we automatically generate index that links the case to the relevant goal the student is working on. After the Digital Loft has done the heavy-lifting of generating, recommending, and indexing cases, the instructional designers can make the final case selection. Instructional designers can also edit the cases to improve their quality (Puntambekar & Kolodner, 2005; Kolodner et al., 2004), and present related so to encourage case comparison thus improving the chances of transfer (Thompson, Gentner, & Loewenstein, 2000; Gentner et al., 2003).

The learner-driven instructional loop

One of the difficulties of teaching groups of students of varying abilities engaged in projects at differing stages is how to provide face-to-face group instruction in a relevant and timely manner. When should a facilitator lead a "user research" workshop if each group is at a different stage of the design process? While the Loft tailors feedback and instruction to each project team, there is still a need for group instruction taught by a knowledgeable facilitator.

In the ***learner-driven instruction loop***, students' self assessments of their abilities and interest in learning different design skills are collected and monitored by the Loft. When enough students indicate a desire to learn a certain skill set, facilitators are notified that there is an opportunity to teach a workshop on an in-demand topic. The learner-driven instructional loop begins after students complete a badge. At this point, the Loft reminds learners to update their "individual development plans" (Beausaert, Segers, & Gijselaers, 2011). An individual development plan (IDP) is a list of skills along with the learner's self-assessment of his current ability level and desire to learn that skill. As students take on new badge challenges, the skills necessary for completing that badge are added to their IDP. Once a given number of students at a DFA studio or classroom express an interest in learning a particular skill, facilitators are notified that they should conduct a particular workshop (and provided with a facilitator's guide for that workshop). Because these workshops are triggered by students' current interests, the workshops maximally target students' interests and needs. While students may not be perfectly accurate in their self-assessments, feedback from mentors and peers provide a reality check on the students' self-assessments (i.e., negative feedback from mentors will prompt students to reassess their skills).

People who implement career goal plans report greater success and satisfaction in their career (Ng, Eby, Sorensen, & Feldman, 2005), so IDPs for civic innovation should increase the success and satisfaction of novice civic innovators on their journey to become more successful designers.

4. CONCLUSION

The study of Digital Lofts will lead empirically-grounded principles for designing online environments for civic innovation education, contributing to number of research areas including digital badges, crowdsourcing, learning-by-cases, design-based learning, and online learning communities. Because many domains can be framed as design disciplines including engineering (making technologies), policy (creating government programs), English language arts (creating texts and speeches), and even science (creating research studies), principles for online innovation education apply to myriad disciplines. And by coordinating groups of learners and mentors throughout the design process, Digital Lofts blur the boundaries between informal and formal learning environments: making extra curricular environments more effective and classroom environments more like real life. This project seeks to lay a theoretical foundation for understanding the broader ecosystem of online, social, design-based learning environments.

More broadly, our goal is to create a widely adopted online learning environment that will support civic innovation training. The Digital Loft platform will be disseminated broadly, targeting use in the teaching, training, and learning of civic innovation. This will fill an urgent need for learning environments that educate civic innovators who can solve our greatest societal challenges. Foreseeable impacts on higher education and society include: increasing the number of graduates motivated and capable of broader societal impact, improved education, curricular changes, and support for future interventions. Successful output of this project will help to foster and support a culture of innovation in our future workforce. By developing a scalable, cost-effective, online platform for design-based learning across many disciplines (design, engineering, speaking, etc.) Digital Lofts have the potential to fundamentally transform online learning.

5. ACKNOWLEDGMENTS

We are grateful to the participants of Design for America for their enthusiastic participation. This work was generously funded by the Digital Medial Learning Competition supported by the MacArthur and Mozilla Foundations.

REFERENCES

ABET Engineering Accreditation Commission. (2011). *Criteria for accrediting engineering programs*. Baltimore, MD: ABET. Retrieved from http://www.abet.org/uploadedFiles/Accreditation/Accreditation_Process/Accreditation_Documents/Current/eac-criteria-2012-2013.pdf

Ambrose, S. A., Bridges, M. W., DiPietro, M., Lovett, M. C., & Norman, M. K. (2010). *How learning works: 7 research-based principles for smart teaching.* San Francisco, CA: Jossey-Bass.

Argote, L. (1999). *Organizational learning: Creating, retaining and transferring knowledge.* New York: Springer.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* W. H. Freeman and Company.

Beausaert, S., Segers, M., & Gijselaers, W. (2011). The use of a personal development plan and the undertaking of learning activities, expertise-growth, flexibility and performance: The role of supporting assessment conditions. *Human Resource Development International*, *14*(5), 527-543.

Beckman, S. L., & Barry, M. (2007). Innovation as a learning process: Embedding design thinking. *California Management Review*, *50*(1), 25.

Bell, P., Davis, E. A., & Linn, M. C. (1995). The knowledge integration environment: Theory and design. In *The first international conference on computer support for collaborative learning* (pp. 14-21).

Blackboard Inc. (2012). *Blackboard* [Computer Software]. Retrieved from http://www.blackboard.com/

Brown, A. L., & Campione, J. C. (1996). Psychological theory and the design of innovative learning environments: On procedures, principles, and systems. In L. Schauble & R. Glaser (Eds.), *Innovations in learning: New environments for education* (pp. 289-322). Mahway, NJ: Erlbaum.

Canvas. (2012). *Instructure* [Computer Software]. Retrieved from http://www.instructure.com/

Clark, R. E. (2008). *Design document for A guided experiential learning course*. Submitted to satisfy contract DAAD 19-99-D-0046-0004 from TRADOC to the Institute for Creative Technologies and the Rossier School of Education, University of Southern California. (Original work published 2004) Retrieved from http://www.cogtech.usc.edu/publications/gel_design_document.pdf

Deci, E. L., & Ryan. (1987). The support of autonomy and the control of behavior. *Journal of Personality and Social Psychology*, 53(6), 1024-1037

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, *125*(6), 627-668.

Dillenbourg, P., & Jermann, P. (2010). Technology for classroom orchestration. In M. S. Khine & I. M. Salhey (Eds.), *New science of learning* (pp. 525-552). New York: Springer.

Dong, A. (2005). The latent semantic approach to studying design team communication. *Design Studies*, *26*(5), 445-461.

Dubberly, H. (2005). *How Do You Design?* Retrieved from http://www.dubberly.com/wp-content/uploads/2008/06/ddo_designprocess.pdf

Duncan, A. (2011, September 15). Digital badges for learning: Remarks by secretary duncan at 4th annual launch of the macarthur foundation digital media and lifelong learning competition. Retrieved from http://www.ed.gov/news/speeches/digital-badges-learning

Easterday, M. W. (2012). Matthew easterday: Cyber-Civics 101. Presentation at the NSF cyberlearning summit, jan 18, 2012, washington D.C. Retrieved from from http://www.youtube.com/watch?v=UBPeDVR2nOo&feature=youtu.be

Edelson, D. C. (2001). Learning-for-Use: A framework for the design of technology-supported inquiry activities . *Journal of Research in Science Teaching*, *38*(3), 355-385.

Edelson, D. C., Gordin, D. N., & Pea, R. D. (1999). Addressing the challenges of inquiry-based learning through technology and curriculum design. *Journal of the Learning Sciences*, *8*(3-4), 391-450.

ETR Associates. (2012). What is service-learning? [Web page] Retrieved from http://www.servicelearning.org/what-service-learning.

Furco, A. (1996). Service-learning: A balanced approach to experiential education. *Expanding Boundaries: Serving and Learning*, *1*, 1-6.

Gall, S. N. -L. (1981). Help-seeking: An understudied problem-solving skill in children. *Developmental Review*, *1*(3), 224 - 246. doi:10.1016/0273-2297(81)90019-8

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, *95*(2), 393.

Gerber, E. M., Marie Olson, J., & Komarek, R. L. D. (2012). Extracurricular design-based learning: Preparing students for careers in innovation. *International Journal of Engineering Education*, *28*(2), 317.

Hardiman, P. T., Pollatsek, A., & Well, A. D. (1986). Learning to understand the balance beam. *Cognition and Instruction*, *3*(1), 63-86.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.*

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112.

Khan, S. (2012). *The khan academy.* [Web page] Retrieved from http://www.khanacademy.org/

Kiesler, S., Kraut, R., Resnick, P., & Kittur, A. (2012). Regulating behavior in online communities. In P. Resnick & R. Kraut (Eds.), *Evidence-based social design: Mining the social sciences to build online communities* (pp. 125-178). Cambridge, MA: MIT Press.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75-86.

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, *19*(3), 239-264.

Kolodner, J. L. (ed.) (1993). *Case-based learning.* Boston: Springer.

Kolodner, J. L. (1997). Educational implications of analogy: A view from case-based reasoning. *American Psychologis*, *52*(1), 57.

Kolodner, J. L., Crismond, D., Gray, J., Holbrook, J., & Puntambekar, S. (1998). Learning by design from theory to practice. *Proceedings of the International Conference of the Learning Sciences*, 16-22.

Kolodner, J. L., Owensby, J. N., & Guzdial, M. (2004). Case-based learning aids. In *Handbook of research for educational communications and technology* (Vol. 2, pp. 829-861).

Kraut, R., & Resnick, P. (2012). Encouraging contribution to online communities. In P. Resnick & R. Kraut (Eds.), *Evidence-based social design: Mining the social sciences to build online communities* (pp. 21-76). Cambridge, MA: MIT Press.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge university press.

Lore. (2012). Lore. [Computer Software] Retrieved from http://lore.com/

Lovett, M., Meyer, O., & Thille, C. (2008). The open learning initiative: Measuring the effectiveness of the OLI statistics course in accelerating student learning. *Journal of Interactive Media in Education.* Retrieved from *http://jime.open.ac.uk/2008/14*

*M*assachusetts Department of Education. (2006). *Massachusetts science and technology/engineering curriculum framework*. Massachusetts. Retrieved from www.doe.mass.edu/frameworks/scitech/1006.pdf

Massachusetts Institute of Technology. (2012). *MIT opencourseware* [Computer Software]. Retrieved from http://ocw.mit.edu/index.htm

Ng, T. W. H., Eby, L. T., Sorensen, K. L., & Feldman, D. C. (2005). Predictors of objective and subjective career success: A meta-analysis. *Personnel Psychology*, *58*(2), 367-408.

Owen, C. L. (1998). Design research: Building the knowledge base. *Design Studies*, *19*(1), 9-20.

Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas*. Basic Books, Inc.

Papert, S., & Harel, I. (1991). Situating constructionism. In *Constructionism* (pp. 1-11).

Peters, V. L., & Slotta, J. D. (2010). Scaffolding knowledge communities in the classroom: New opportunities in the web 2.0 era. In *Designs for learning environments of the future* (pp. 205-232). New York: Springer.

Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review*, *16*(4), 385-407.

Puntambekar, S., & Kolodner, J. L. (2005). Toward implementing distributed scaffolding: Helping students learn science from design. *Journal of Research in Science Teaching*, *42*(2), 185-217.

Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *The Journal of the Learning Sciences*, *13*(3), 273-304.

Reiser, B. J., Tabak, I., Sandoval, W. A., Smith, B. K., Steinmuller, F., & Leone, A. J. (2001). BGuILE: Strategic and conceptual scaffolds for scientific inquiry in biology classrooms. In *Cognition and instruction* (Vol. 25, pp. 263-306).

Resnick, M. (2009). Scratch: Programming for all. *Communications of the ACM*, *52*(11), 60-67. doi:10.1145/1592761.159277

Ryan, A. M., Pintrich, P. R., & Midgley, C. (2001). Avoiding seeking help in the classroom: Who and why? *Educational Psychology Review*, *13*(2), 93-114.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54-67. doi:10.1006/ceps.1999.1020

Sadler, P. M., & Good, E. (2006). The impact of self-and peer-grading on student learning. *Educational Assessment*, *11*(1), 1-31.

Sakai Foundation. (2012). *Sakai* [Computer Software]. Retrieved from http://www.sakaiproject.org/

Salden, R. J. C. M., Aleven, V. A. W. M. M., Renkl, A., & Schwonke, R. (2009). Worked examples and tutored problem solving: Redundant or synergistic forms of support? *Topics in Cognitive Science*, *1*(1), 203-213. doi:10.1111/j.1756-8765.2008.01011.x

Sandoval, W. A., & Reiser, B. J. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education*, *88*(3), 345-372.

Scardamalia, M., Bereiter, C., & Steinbach, R. (1984). Teachability of reflective processes in written composition. *Cognitive Science*, *8*(2), 173-190.

Scheines, R., Leinhardt, G., Smith, J., & Cho, K. (2005). Replacing lecture with web-based course materials. *Journal of Educational Computing Research*, *32*(1), 1-26.

Severance, C. (2012). Teaching the world: Daphne koller and coursera. *Computer*, *45*(8), 8-9.

Shaffer, D. W., & Resnick, M. (1999). Thick" authenticity: New media and authentic learning. *Journal of Interactive Learning Research*, *10*(2), 195-215.

Slotta, J. D. (2004). The web-based inquiry science environment (WISE): Scaffolding knowledge integration in the science classroom. In M. C. Linn, E. A. Davis, & P. Bell (Eds.), *Internet environments for science education* (pp. 203-232). Lawrence Erlbaum Mahwah, NJ.

Songer, N. B. (2006). BioKIDS: An animated conversation on the development of complex reasoning in science. In R. Keith Sawyer (Ed.), *The cambridge handbook of the learning sciences* (pp. 355-370). New York: Cambridge University Press.

Thompson, L., Gentner, D., & Loewenstein, J. (2000). Avoiding missed opportunities in managerial life: Analogical training more powerful than individual case training. *Organizational Behavior and Human Decision Processes*, *82*(1), 60-75.

Velmahos, G. C., Toutouzas, K. G., Sillin, L. F., Chan, L., Clark, R. E., Theodorou, D., & Maupin, F. (2004). Cognitive task analysis for teaching technical skills in an inanimate surgical skills laboratory. *American Journal of Surgery*, *187*(1), 114-9.

Von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 319-326).

Ward, M., & Sweller, J. (1990). Structuring effective worked examples. *Cognition and Instruction*, *7*(1), 1-39.

White, B., Frederiksen, J., Frederiksen, T., Eslinger, E., Loper, S., & Collins, A. (2002). Inquiry island: Affordances of a multi-agent environment for scientific inquiry and reflective learning. In *Proceedings of the fifth international conference of the learning sciences (ICLS)* . Mahwah, NJ: Erlbaum.

What is my essay really saying? Using extractive summarization to motivate reflection and redrafting

Nicolas Van Labeke[1], Denise Whitelock[1], Debora Field[2], Stephen Pulman[2], John Richardson[1]

[1] Institute of Educational Technology
The Open University
Walton Hall, Milton Keynes, MK7 6AA, UK
Nicolas.Vanlabeke@open.ac.uk
Denise.Whitelock@open.ac.uk
John.T.E.Richardson@open.ac.uk

[2] Department of Computer Science
University of Oxford
Wolfson Building, Parks Road, Oxford OX1 3QD, UK
stephen.pulman@cs.ox.ac.uk
debora.field@cs.ox.ac.uk

ABSTRACT

This paper reports on progress on the design of OpenEssayist, a web application that aims at supporting students in writing essays. The system uses techniques from Natural Language Processing to automatically extract summaries from free-text essays, such as key words and key sentences, and carries out essay structure recognition. The current design approach described in this paper has led to a more "explore and discover" environment, where several external representations of these summarization elements would be presented to students, allowing them to freely explore the feedback, discover issues that might have been overlooked and reflect on their writing. Proposals for more interactive, reflective activities to structure such exploration are currently being tested.

Keywords

Essay writing; Extractive Summarization; Formative Feedback; External Representations; Reflective Activities.

1. INTRODUCTION

Written discourse is a major class of data that learners produce in online environments, arguably the primary class of data that can give us insights into deeper learning and higher order qualities such as critical thinking, argumentation and mastery of complex ideas. These skills are indeed difficult to master as illustrated in the revision of Bloom's Taxonomy of Educational Objectives (Pickard 2007) and are a distinct requirement for assessment in higher education. Assessment is an important component of learning and in fact (Rowntree 1987) argues that it is the main driver for learning and so the challenge is to provide an effective automated interactive feedback system that yields an acceptable level of support for university students writing essays.

Effective feedback requires that students are assisted to manage their current essay-writing tasks and to support the development of their essay-writing skills through effective self-regulation.

Our research involves using state-of-the-art techniques for analyzing essays and developing a set of feedback models which will initiate a set of reflective dialogic practices. The main pedagogical thrust of e-Assessment of free-text projects is how to provide meaningful "advice for action" (Whitelock 2010) in order to support students writing their summative assessments. It is the combination of incisive learning analytics and meaningful feedback to students which is central to the planning of our empirical studies. Specifically, we are investigating whether summarization techniques (Lloret & Palomar 2012) could be used to generate formative feedback on free-text essays submitted by students.

This paper is organized as follows. We briefly describe the context and research questions that are informing the design principles of our platform, OpenEssayist. We then describe the basic processes behind the summarization techniques implemented in the system and, finally, demonstrate the current stage of design of the prototype, in particular the use of external representations for the summarization elements. We conclude this paper by sketching our current and planned evaluations.

2. DEFINING A DESIGN SPACE FOR OPENESSAYIST

2.1 WRITING SUMMARIES VS. REFLECTING ON SUMMARIES FOR WRITING.

Writing summaries has been a long-standing educational activity and has received some serious attention in delivering computer-based support. For example, systems such as SummaryStreet (Wade-Stein & Kintsch 2004) or Pensum (Villiot-Leclercq *et al.* 2010) aim to help students *write* summaries as a learning, skills-based, task.

But using summaries as a source of reflection on your own writing seems to be a more open issue. Recent research on formative feedback suggests indeed that essay summarization, understood to comprise both a short summary of the essay and a simple list of its main topics, could be useful for students, e.g. "*to help determine whether the actual performance was the same as the intended performance*" (Nelson & Schunn 2009, p. 378).

With this in mind, one of our research questions is how to use advances in Natural Language Processing to design an automated summarization engine that would provide a good foundation for a dedicated model of formative feedback. Can we use summarization elements to help students identify or visualize patterns in their essays, as explored by (O'Rourke & Calvo 2009)? Or to trigger questions and reflective activities, as implemented in Glosser (Villalon *et al.* 2008)?

2.2 SUPPORTING ESSAY WRITING IN DISTANCE LEARNING

The context of application of our research agenda is supporting students at the Open University (OU) in writing assignment essays. Specifically, we have been working closely with a postgraduate module *Accessible online learning: Supporting disabled students* (referred to as H810). This postgraduate module runs twice a year for about 20 weeks and contributes to a Master of Arts (MA) in Online and Distance Education. All courses, materials and support are delivered online. Students on this module, as is the case for most of the students at the OU, are typically part-time, mature students, who have not been in formal education for a long period of time. It is therefore unsurprising that writing essays, a common assignment in most of the OU courses, proves to be a challenging task for students (and, anecdotal evidence suggests, a common reason for drop-out).

At the same time, OU students often have extensive work experience in a wide variety of areas, and that experience is explicitly capitalized on in the assignments. This means that essays can vary greatly in subject matter. To illustrate this point, two examples of assignment tasks are given in Table 1.

Table 1. Examples of assignment tasks.

TMA1 (1500 words)
Write a report explaining the main accessibility challenges for disabled learners that you work with or support in your own work context(s). Use examples from your own experience, supported by the research and practice literature. If you're not a practitioner, write from the perspective of a person in a relevant context. Critically evaluate the influence of the context (e.g. country, institution, perceived role of online learning within education) on the: (1) identified challenges; (2) influence of legislation; (3) roles and responsibilities of key individuals; (4) role of assistive technologies in addressing these challenges.
TMA2 (3000 words)
Critically evaluate your own learning resource in the following ways: (1) Briefly describe the resource and its accessibility features; (2) Evaluate the accessibility of your resource, identifying its strengths and weaknesses; (3) Reflect on the processes of creating and evaluating accessible resources.

The questions we are considering, given this context, is how we can support these students as they write essays and what the implications are for the design of a computer- and summarization-based approach.

In the initial phase of the project, we ran a couple of focus groups with OU students that helped to identify many aspects of the students' personal approach to essay writing (Alden *et al*. 2013).

Writing an essay is a task that can involve several stages: preparation of material, drafting of essay, reflecting on feedback, summative evaluation by tutors. But not all of them are suitable, or even desirable, for support in an automated assessment system.

Moreover, writing a 1500+ word essay is not a casual operation, nor is it handled in the same way by different students. For example, we discovered that some students are not using computers to draft their essays, because of unease, lack of permanent access to a desktop computer or simply because they still prefer to write their text with paper-and-pencil before typing for the final submission.

Relying on embedded text editors or on cloud-based solutions such as Google Docs – as done by (Southavilay *et al*. 2013) for collaborative writing – is therefore not a viable solution in our context. The system will have to accept texts written with whatever platform students are using to organize, draft and revise their essay. Ultimately, the system will have to be seen and used as a resource, the way forums, online textbooks and other digital tools are used by OU students.

One of the consequences of such selective support is that the flow of activities during the overall writing process is likely to be highly scattered in time: the core of the activity (i.e. writing) will take place *outside* the system's ecology and its use will be mostly as an ancillary to that main task. Careful attention will have to be paid to trade-offs between support and distraction, especially when it comes to interaction, formal reflective activities, accessibility and usability[1].

Finally, the diversity of content in student essays is one of the motivations for investigating summarization techniques as a backbone for formative feedback. Unlike other NLP techniques such as Latent Semantic Analysis (LSA), used in many educational systems, we will not be relying on a corpus of essays to compare and grade new essays accordingly. Summarization using the text alone with no domain-specific knowledge will enable OpenEssayist to handle assignments which have open topics, as well as enabling it to be applied without extensive further development to new subject areas.

2.3 A WEB APPLICATION FOR SUMMARIZATION-BASED FORMATIVE FEEDBACK.

OpenEssayist is developed as a web application and is composed primarily of two components (Figure 1, see appendix). The first component, EssayAnalyser, is the summarization engine, implemented in Python with NLTK[2] (Bird *et al*. 2009) and other toolkits. It is being designed as a stand-alone RESTful web service, delivering the basic summarization techniques that will be consumed by the main system. The second component is OpenEssayist itself, implemented on a PHP framework. The core system consists of the operational back-end (user identification, database management, service brokers, feedback orchestrator) and the cross-platform, responsive HTML5 front-end.

The intended flow of activities within the system can be summarized as follows. Students are registered users and have assignments, defined by academic staff, allocated to them. Once they have prepared a draft offline and seek to obtain feedback, they log on to the OpenEssayist system and submit their essay for analysis, either by copy-and-paste or by uploading their document. OpenEssayist submits the raw text to the EssayAnalyser service and, upon completion, retrieves and stores the summarization data. From that point on, the students, at their own pace, can then explore the data using various external

[1] Worth noting is that students who mention that they don't use computers for drafting their essays also report that they are using smart phones. A focus on responsive user interface suitable for mobile (and tablet) and on asynchronous data access will be an issue for serious consideration in this project.

[2] Natural Language Processing Toolkit, see http://nltk.org/

representations made available to them, can follow the prompts and trigger questions that the Feedback Orchestrator might generate from the analysis and can then start planning their next draft accordingly.

Again, this rewriting phase will take place offline, the system merely offering repeated access to the summarization data and feedback, as a resource, until the students are prepared to submit and explore the summarization feedback on their second draft and on the changes across drafts. This cycle of submission, analysis and revision continues until the students consider their essay ready for summative assessment.

3. EXTRACTIVE SUMMARIZATION

We decided to start experimenting with two simpler summarization strategies that could be implemented fairly quickly: key phrase extraction and extractive summarization, following the TextRank approach proposed and evaluated in (Mihalcea & Tarau 2004). Key phrase extraction aims at identifying which individual words or short phrases are the most suggestive of the content of a discourse, while extractive summarization is essentially the identification of whole key sentences. Our hypothesis is that the quality and position of key phrases and key sentences within an essay (i.e., relative to the position of its structural components) might give an idea of how complete and well-structured the essay is, and therefore provide a basis for building suitable models of feedback.

The implementation of these summarization techniques is based on three main automatic processes: 1) recognition of essay structure; 2) unsupervised extraction of key words and phrases; 3) unsupervised extraction of key sentences.

Before extracting key terms and sentences from the text, the text is automatically pre-processed using some of the NLTK modules (tokenizer, lemmatizer, part-of-speech tagger, list of stop words).

3.1 STRUCTURE IDENTIFICATION

The automatic identification of essay structure is carried out using handcrafted rules developed through experimentation with a corpus of 135 essays that have been previously submitted for the same H810 module. The system tries to automatically recognize which structural role is played by each paragraph in the essay (summary, introduction, conclusion, discussion, references, etc.). This identification is achieved regardless of the presence of content-specific headings and without getting clues from formatting mark-up. With the essays in the corpus varying greatly in structure and formatting, it was decided that structure recognition would be best achieved without referring to a high-level formatting mark-up.

3.2 KEY WORD EXTRACTION

EssayAnalyser uses graph-based ranking methods to perform unsupervised extractive summarization of key words. The 'key-ness' value of a word can be understood as its 'significance within the context of the overall text'.

To compute this key-ness value, each unique word in the essay is represented by a node in a graph, and co-occurrence relations (specifically, within-sentence word adjacency) are represented by edges in the graph. A centrality algorithm – we have experimented with betweenness centrality (Freeman 1977) and PageRank (Brin & Page 1998) – is used to calculate the significance of each word. Roughly speaking, a word with a high centrality score is a word that sits adjacent to many other unique words which sit adjacent to many other unique words which…, and so on. The words with high centrality scores are the key words[3].

Since a centrality score is attributed to *every* unique word in the essay, a decision needs to be made as to what proportion of the essay's unique words qualify as key words. The distribution of key word scores follows the same shape for all essays, an acute "elbow" and then a very long tail, observed for word adjacency graphs by (Ferrer i Cancho & Solé 2001). We therefore currently take the key-ness threshold to be the place where the elbow bend appears to be sharpest.

Once key words have been identified, the system matches sequences of these against the surface text to identify within-sentence key phrases (bigrams, trigrams and quadgrams).

3.3 KEY SENTENCE EXTRACTION

A similar graph-based ranking approach is used to compute key-ness scores to rank the essay's sentences. Instead of word adjacency (as in the key word graph), co-occurrence of words across pairs of sentences is the relation used to construct the graph. More specifically, we currently use cosine similarity to derive a similarity score for each pair of sentences. Whole sentences become nodes in the graph, while the similarity scores become weights on the edges connecting pairs of sentences. The TextRank key sentence algorithm is then applied to the graph to compute the centrality scores.

3.4 ESSAY ANALYSIS OUTPUT

The text submitted for analysis is stripped of its surface formatting and returned as a *new* annotated structured text, reflecting the various elements identified by EssayAnalyser: sentences and paragraphs, labeled with their structural roles (body, introduction, headings, conclusions, captions, etc.) and confidence levels.

Key words and key phrases are returned as an ordered list of terms, associated with various metrics such as centrality, frequency of inflected forms, etc. Key sentences are identified within the annotated text by their ranked centrality scores.

In addition to the core summaries of the essay, various metrics and specialized data structures are made available, for use by the system for diagnosis purpose (or by researchers for analysis): word and sentence graphs, word count, paragraph and sentence density and length, number of words in common with the module textbook, average frequency of the top handful of most frequent words, etc.

Our task is now to look for ways of presenting and exploiting these results and, ultimately, to devise effective models of feedback using them.

4. OPENESSAYIST: EXTERNAL REPRESENTATIONS AND REFLECTIVE ACTIVITIES

The design of the first version of the system has focused on defining the essay summarization engine and integrating it into a working web application that supports draft submission, analysis and reporting, using multiple external representations.

[3] In the actual process, we are in fact ranking *lemmas* (the canonical form of a set of words) rather than their inflected forms in the surface text. For brevity's sake, we will keep the terms 'words' and 'key words' in this document.

At the front-end level, the instructional interactions have been deliberately limited to fairly unconstrained forms, leading the system towards a more "explore and discover" environment. Our aim was to establish a space where emerging properties of the interventions under investigation (i.e. using summarization techniques for generating formative feedback) could be discovered, explored and integrated into the design cycles in a systematic way, contributing to both the end-product of the design cycle (the system itself) and to its theoretical foundations.

Several external representations have been designed and deployed in the system, reporting the different elements described above in different ways, trying to highlight such properties in the current essay (or, in changes over successive drafts).

The main view of the system is a mash-up of the re-structured raw text, highlighting many of the features extracted by EssayAnalyser in context, using a combination of HTML markers and JavaScript-enabled interactive displays (Figure 2). Sentences, paragraphs and headings (as identified by EssayAnalyser) are displayed as blocks of text, with visual markers on the left-hand side indicating their diagnosed structural role (e.g. introduction, headings, conclusion, etc.). Key words and key phrases are also highlighted with specific visual markers, as with the top-ranked key sentences.

A control-box allows the student to change the visibility of selected elements of the essay: show/hide specific structural components (e.g. only show the introduction), key words (or user-defined categories, see below), top-ranked sentences, etc. (Figure 3).

The intended purpose of this dynamic essay representation is to attract the attention of the student away from the surface text to issues at a more structural level that might become apparent once an alternative viewpoint is considered.

For example, if confidence levels were low in the structural recognition of an introduction, the visual indicator would reflects that degree of (un)certainty about their exact role of this paragraph, requiring the student to reflect on his intention (or on the fact that an introduction might be missing in the essay or seems to be too long or too short).

Similarly, the highlighting of key words and key phrases, in context within the essay, is intended to trigger reflection on their occurrence within the text. Its purpose is different from a dedicated external representation of the key words as such (Figure 4), where the focus is more on individual terms, and on their relative importance in the essay (as indicated by their centrality score or frequency in the surface text). In the mash-up view, the key word centrality score is played down (we do not represent any attribute other than its identification as a key word) while we try to focus on whether key word *dispersion* across the essay might help identify the flow of ideas and arguments.

To complement the main mash-up view and to alleviate potential overload, we are also designing and deploying ad-hoc external representations on specific aspects of the summarization.

For example, we are exploring whether more compact representations of the dispersion of key words across the essay (Figure 5) might provide a more suitable ground for insight into its meaning. In this graph, each key word (or category of key words, if they have been defined) is plotted on a scale showing the flow of the essay (the figure uses words on the x-axis but sentences and paragraphs can also be used as units). By adding on the scale markers for the introduction, the conclusion (or any other structural elements), the student has immediate access to the overall flow of key words across the text and within specific parts of it: patterns of occurrence or omission might provide opportunity to detect an overlooked mistake (e.g. what can be said about the fact that "learning resource", ranked as a top key word by the system, only occurs in the first few paragraphs of the essay?).

On a more experimental approach, we are also exploring the possibility of visually exploiting the networks that constitute the core internal representation of the key word and key sentence extraction, using various visualization tools (e.g. force-directed graph, adjacency matrix). A case for their informational and – more importantly – formative values remains to be made.

However, we are also arguing that, to help students explore the significance of summarization elements in their essay, visualization on its own will not be enough. Support for reflective *action* is needed to resolve a key question students are likely to ask: "what are the key words (and key sentences) and how do they help me?"

Let's consider the key words. In the current version of the system, key words are presented in a very simple fashion (Figure 4): ranked by their centrality score and by their dimension (i.e. bigrams, trigrams and so on). This is a reflection of the domain-independent, data-driven design approach followed so far; key words are derived on the basis of co-occurrence, i.e. identity relation, not on the basis of semantic relations such as synonymy or hyponymy.

We can therefore have situations, as in Figure 4, where key words such as "learning experience" and "study experience" both occur as distinct bigrams, whereas, for the student who used them, they might mean very similar things. More fine-grained approaches could be implemented in EssayAnalyser to address such situation at detection level, but, ultimately, the *intention* of the student is the only safe ground for deciding on the usage of both terms. Hence the need to support some user interaction with the system, especially if it can act as a reflective scaffold.

A first example of support for reflective action is made available to the students immediately after a draft has been analyzed by the system: to let them organize key words according to their own schema, using as many categories as they wish or need (see Figure 6). This serves two purposes: it helps the students to reflect on the content of the essay and helps the system to adapt the content of every external representation accordingly, by clustering key words together (as seen in Figure 5).

Another key-word-related activity relies on the fact that a decision is made by the system on what constitutes a key word, a decision that might be at odds with the intention of the student. So we are offering the possibility for students to define – or select – their own key words. With the extraction process deriving a centrality score and frequency count for every unique word in the text, the student's decision to flag a word as a key word can be matched with that information, encouraging her to reflect on why it might be that the words she thinks should be key words are not being recognized by the system as such.

5. CONCLUSION

The first phase of the design of OpenEssayist, as reported in this paper, has focused on devising a range of external representations on the various elements that the summarization engine is extracting, notably key words, key sentences and the structural role of paragraphs in the essay.

We have implemented a working prototype that delivers a fairly unconstrained, unstructured exploration of these elements, The

drive of our design approach has been to consider how these elements, either separately or combined, would create a space where students (and researchers) could discover emerging properties of the essay, triggering deeper reflection on their own writing.

Our objective is now to consider how we structure these reflective episodes for support within the system, and how we design dedicated reflective activities that will prove to deliver formative feedback for students.

Our work is continuously focusing on three parallel but inter-connected lines of experimentation and evaluation:

1) improve the different aspects of the summarization engine;
2) experiment with it on various corpora of essays to identify trends and markers that could be used as progress and/or performance indicators (Field *et al*. 2013);
3) refine the educational aspect of the system, identify possible usage scenarios (Alden *et al*. 2013), test pedagogical hypotheses and models of feedback.

At the time of writing, several usability/desirability inspection sessions are underway, using both semi-structured walkthrough protocols in a usability lab and self-guided remote sessions with students from the last presentation of the H810 module. Part of the aim of these empirical studies is to identify tutorial strategies to be used to scaffold the student's exploitation of the system.

Finally, we are planning two empirical educational evaluations of OpenEssayist in an authentic e-learning context, to take place in September 2013 and February 2014. All students enrolled on two different Master's degree modules will be offered access to the system for two of the module's assignments and encouraged to submit multiple drafts of their essays. In-system data collection, post-module surveys, and interviews with selected participants and their tutors will give us valuable information on their learning experience with the system.

ACKNOWLEDGEMENTS

This work is supported by the Engineering and Physical Sciences Research Council (EPSRC, grant numbers EP/J005959/1 & EP/J005231/1).

REFERENCES

Alden, B., Van Labeke, N., Field, D., Pulman, S., Richardson, J. T. E., and Whitelock, D. (2013). Using student experience to inform the design of an automated feedback system for essay answers. In *Proceedings of the 2013 International Computer Assisted Assessment Conference* (CAA'13, Southampton, UK). pp. to appear.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. Cambridge, MA: O'Reilly Media, Inc.

Brin, S., and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30(1), pp. 107–117.

Ferrer i Cancho, R., and Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 268(1482), pp. 2261–2265.

Field, D., Richardson, J. T. E., Pulman, S., Van Labeke, N., and Whitelock, D. (2013). Reflections on characteristics of university students' essays through experimentation with domain-independent natural language processing techniques. In *Proceedings of the 2013 International Computer Assisted Assessment Conference* (CAA'13, Southampton, UK). pp. to appear.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry* 40(1), pp. 35–41.

Lloret, E., and Palomar, M. (2012). Text summarisation in progress: a literature review. *Artificial Intelligence Review* 37(1), pp. 1–41.

Mihalcea, R., and Tarau, P. (2004). TextRank: Bringing order into texts. In *Proceedings of Empirical Methods in Natural Language Processing* (EMNLP'04, Barcelona, Spain). , pp. 404–411.

Nelson, M. M., and Schunn, C. D. (2009). The nature of feedback: how different types of peer feedback affect writing performance. *Instructional Science* 37(4), pp. 375–401.

O'Rourke, S. T., and Calvo, R. A. (2009). Analysing Semantic Flow in Academic Writing. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (AIED'09, Brighton, UK). IOS Press, pp. 173–180.

Pickard, M. J. (2007). The new Bloom's taxonomy: An overview for family and consumer sciences. *Journal of Family and Consumer Sciences Education* 25(1), pp. 45–55.

Rowntree, D. (1987). *Assessing Students: How Shall We Know Them?* London: Kogan Page.

Southavilay, V., Yacef, K., Reimann, P., and Calvo, R. A. (2013). Analysis of Collaborative Writing Processes Using Revision Maps and Probabilistic Topic Models. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (LAK'13, Leuven, Belgium). ACM, pp. 38–47.

Villalon, J., Kearney, P., Calvo, R. A., and Reimann, P. (2008). Glosser: Enhanced Feedback for Student Writing Tasks. In *Proceedings of the 8th IEEE International Conference on Advanced Learning Technologies* (ICALT'08, Santander, Spain). IEEE Press, pp. 454–458.

Villiot-Leclercq, E., Mandin, S., Dessus, P., and Zampa, V. (2010). Helping Students Understand Courses through Written Syntheses: An LSA-Based Online Advisor. In *Proceedings of the 10th International Conference on Advanced Learning Technologies (ICALT)* (ICALT'10, Sousse, Tunisia). IEEE Press, pp. 341–343.

Wade-Stein, D., and Kintsch, E. (2004). Summary Street: Interactive Computer Support for Writing. *Cognition and Instruction* 22(3), pp. 333–362.

Whitelock, D. (2010). Activating Assessment for Learning: Are We on the Way with Web 2.0? In *Web 2.0-Based E-Learning: Applying Social Informatics for Tertiary Teaching*, eds. Mark J.W. Lee and Catherine McLoughlin. Hershey, PA: IGI Global pp. 319–342.

APPENDIX

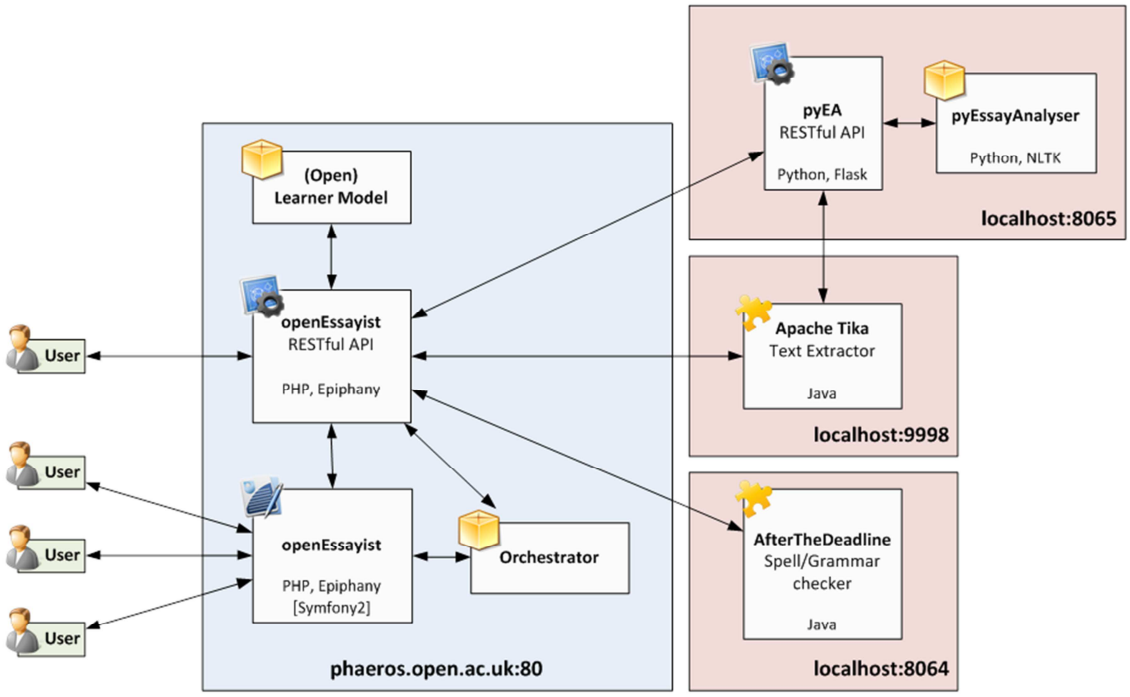


Figure 1. Architecture of OpenEssayist

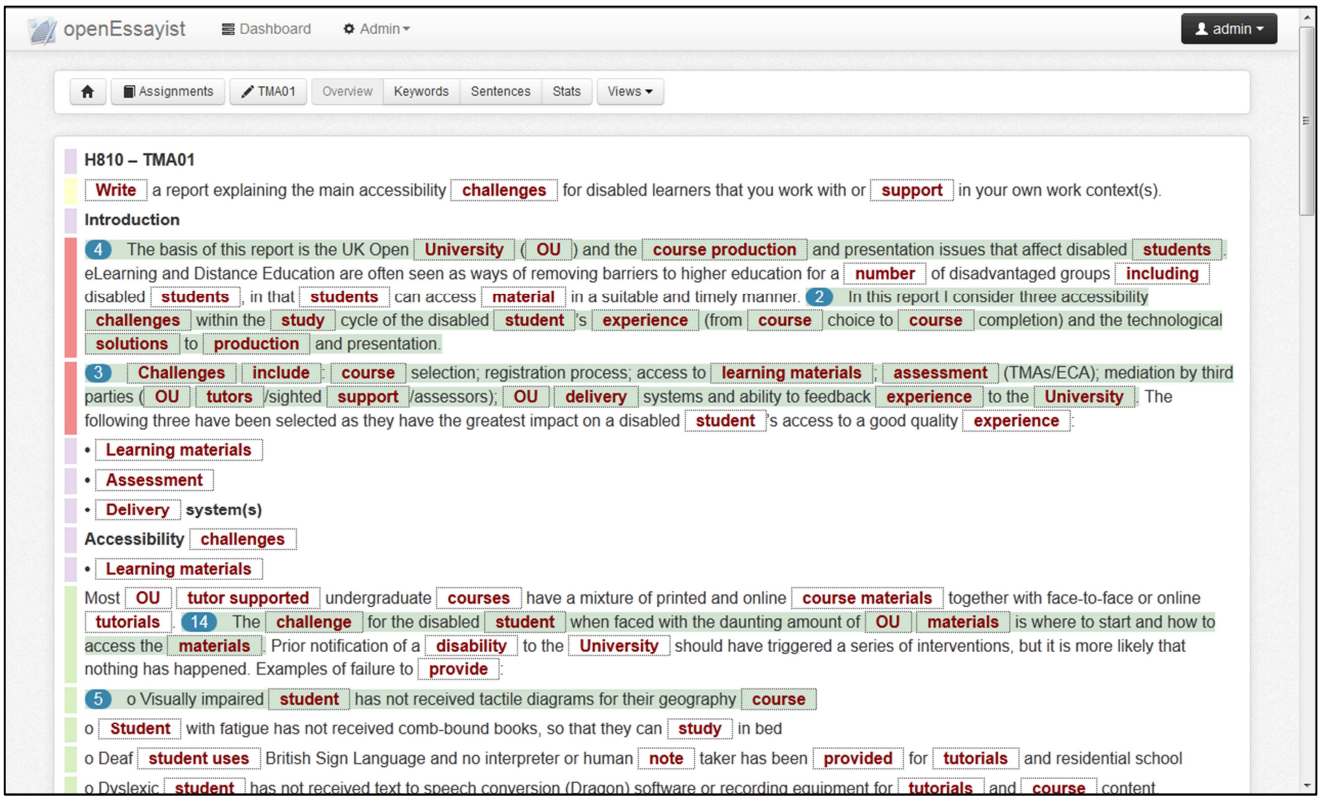


Figure 2. Key words, phrases and sentences visualized in the essay context. Sentences in light-grey (green) background are key sentences as extracted by the EssayAnalyser (the number indicates its key-ness ranking). Key words and key phrases are indicated in bold (red) and boxed.

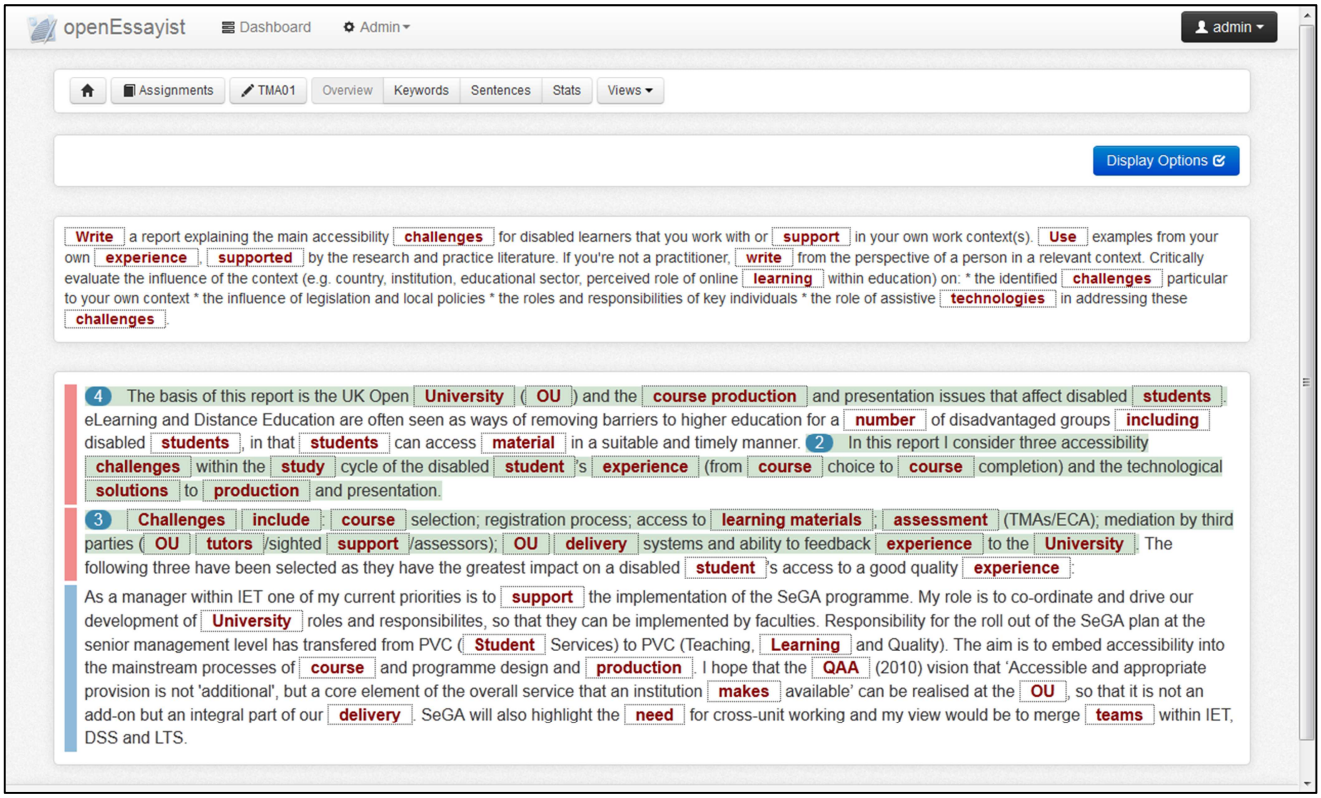


Figure 3. The structural elements of the essay can be used jointly with the key word extraction to highlight relevant information within specific parts of the essay, here in both introduction and conclusion (and the assignment question).

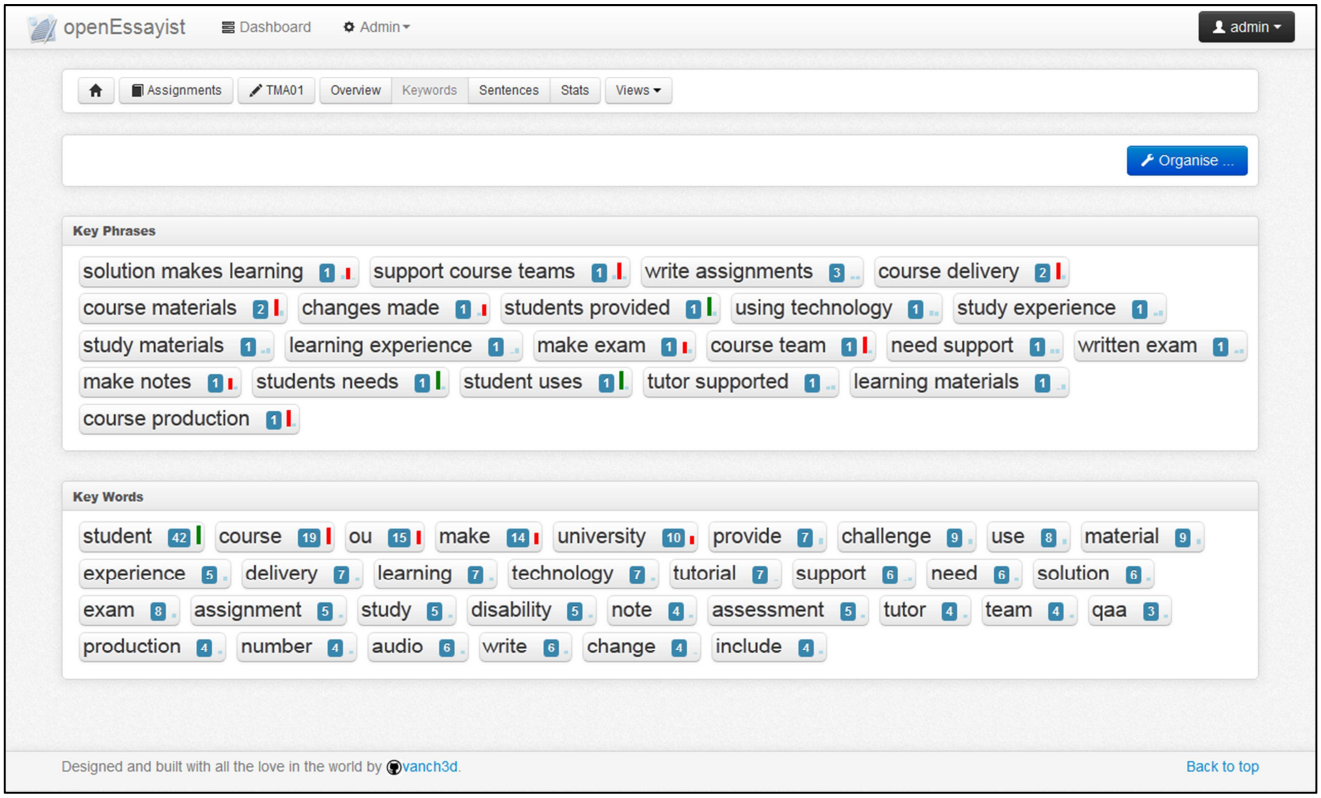


Figure 4. Key words and phrases as separate lists.

Figure 5. Dispersion of key words across the essay.

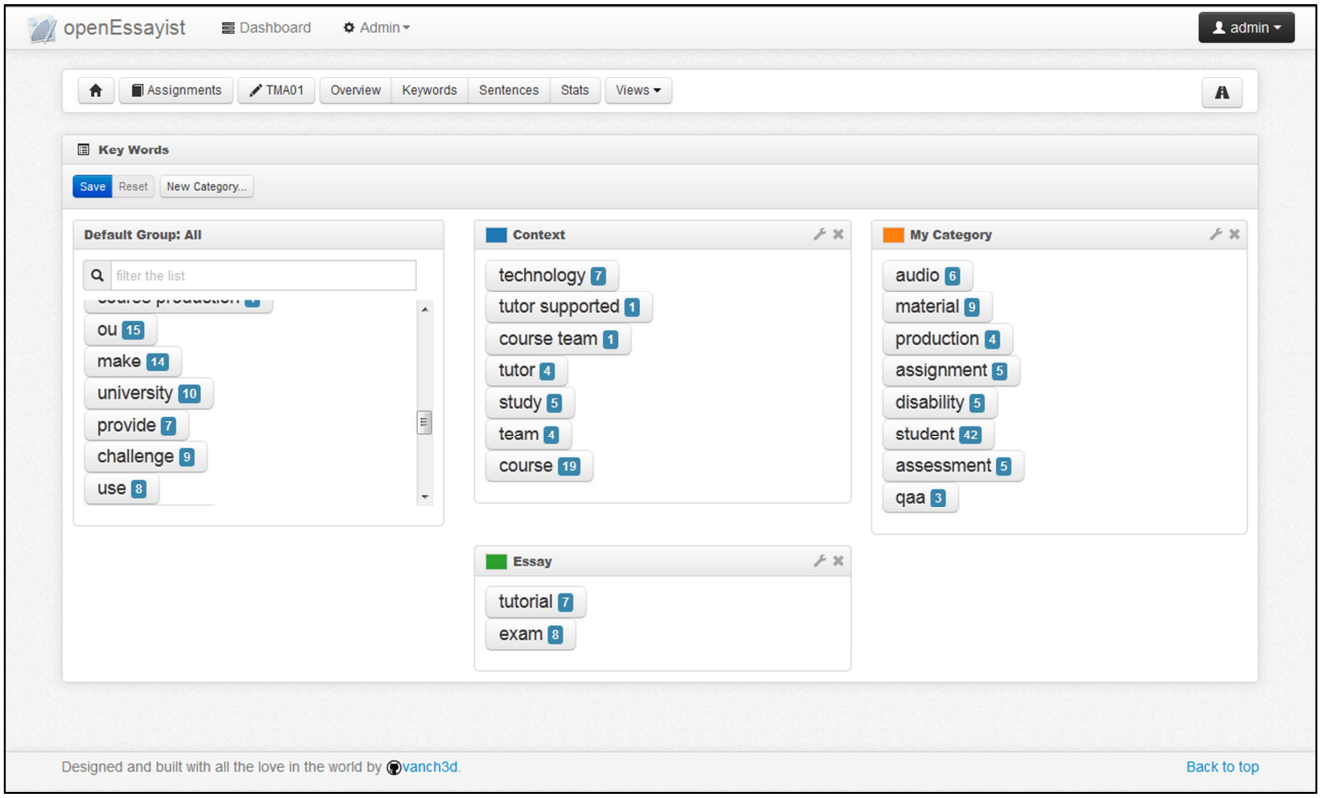


Figure 6. Key words extracted by the systems are re-organized by the students, using their own categories

A User Study on the Automated Assessment of Reviews

Lakshmi Ramachandran and Edward F. Gehringer
North Carolina State University
{lramach,efg}@ncsu.edu

ABSTRACT

Reviews are text-based feedback provided by a reviewer to the author of a submission. Reviews play a crucial role in providing feedback to people who make assessment decisions (e.g. deciding a student's grade, purchase decision of a product). It is therefore important to ensure that reviews are of a good quality. In our work we focus on the study of academic reviews. A review is considered to be of a good quality if it can help the author identify mistakes in their work, and help them learn possible ways of fixing them. *Metareviewing* is the process of evaluating reviews. An automated metareviewing process could provide quick and reliable feedback to reviewers on their assessment of authors' submissions. Timely feedback on reviews could help reviewers correct their assessments and provide more useful and effective feedback to authors. In this paper we investigate the usefulness of metrics such as *review relevance, content type, tone, quantity* and *plagiarism* in determining the quality of reviews. We conducted a study on 24 participants, who used the automated assessment feature on Expertiza, a collaborative peer-reviewing system. The aim of the study is to identify reviewers' perception of the usefulness of the automated assessment feature and its different metrics. Results suggest that participants find relevance to be the most important and quantity to be the least important in determining a review's quality. Participants also found the system's feedback from metrics such as content type and plagiarism to be most useful and informative.

Keywords

review quality assessment, metareview metrics, user experience survey

1. INTRODUCTION

In recent years there has been a considerable amount of research directed towards developing educational systems that foster collaborative learning. Collaborative learning systems provide an environment for students to interact with other students, exchange ideas, provide feedback and use the feedback to improve their own work. Systems such as SWoRD [1] and Expertiza [3] are web-based collaborative peer-review systems, which promote team work by allowing students to build shared knowledge with an exchange of ideas. These systems also provide an environment for students to give feedback to peers on their work.

The process of providing feedback to peers on their work may help students learn more about the subject, and develop their critical thinking. Rada et al. found that students who evaluated their peers' work were more likely to improve the quality of their own work than those students who did not provide peer reviews [4]. The peer review process may also help students learn to be more responsible.

The classroom peer review process is very much similar to the process of reviewing scientific articles for journals. Scientific reviewers tend to have prior reviewing experience and a considerable knowledge in the area of the author's submission (the text under review). Students on the other hand are less likely to have had any prior reviewing experience. They have to be guided to provide good quality reviews that may be useful to their peers.

Metareviewing can be defined as the process of reviewing reviews, i.e., the process of identifying the quality of reviews. Metareviewing is a manual process [2, 5, 6] and just as with any process that is manual; metareviewing too is (a) slow, (b) prone to errors and (c) likely to be inconsistent - the set of problems, which makes automated metareviewing necessary. An automated metareview process ensures consistent, bias-free reviews to all reviewers. This also ensures provision of immediate feedback to reviewers, which is likely to motivate them to improve their work and provide more useful feedback to the authors.

In this work we propose the use of a system that automatically evaluates student review responses. We use a specific set of metrics such as *review's relevance* to the work under review (or the submission), the *type of content* a review contains, *tone* of the review, *quantity* of feedback provided and presence of *plagiarism*, to carry out metareviewing. We have integrated the automated metareview feature (with the listed set of metrics) into Expertiza [3]. Expertiza is a collaborative web-based learning application. A screenshot of the metareview output from the system is shown in Figure1. We have conducted an exploratory analysis to study the importance of the review quality metrics and usefulness of the system's outputs, as judged by users of the system.

2. RELATED WORK

One of the earlier approaches to manually assessing the quality of peer reviews involved the creation and use of a Review Quality Instrument (RQI) [9]. Van Rooyen et al. use the RQI to check whether a reviewer discusses the following - (1) importance of the

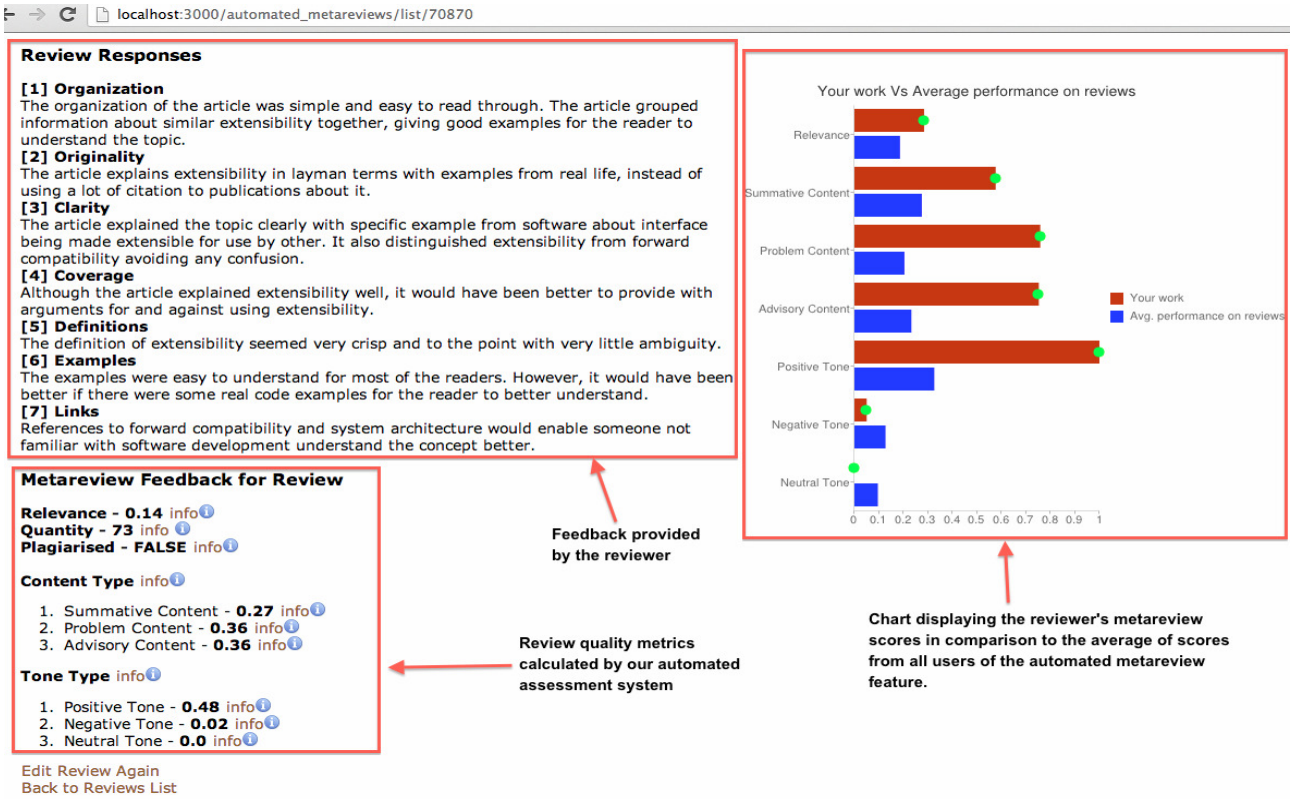


Figure 1: Output from the automated metareview feature on Expertiza [3]. We provide a comparison of the participant reviewer's scores with other reviewers' metareview scores (in a chart) to help reviewers gauge how well they are doing on a certain metric.

research question, (2) originality, (3) strengths and weaknesses, (4) presentation and interpretation of results. In addition, the RQI also checks if a review was constructive, and if the reviewer had substantiated his/her claims. We incorporate some of these metrics in our approach, e.g. detecting constructiveness in reviews (based on its content), checking whether reviewers substantiated their claims (by identifying relevance to the author's submission), to automatically assess review quality.

Nelson and Schunn studied feedback features that help authors understand and use reviews [10]. They found that features such as problem localization and solution suggestion helped authors understand feedback. These are some of the types of content we look for during review content identification.

Kuhne et al. use authors' ratings of reviews to identify the quality of peer reviews [5]. They found that authors are contented with reviewers who appear to have made an effort to understand their work. This finding is useful to our automatic review quality assessment system, which assesses reviews based on the usefulness of its content. Our system also detects the relevance of reviews, which may be indicative of the effort made by a reviewer to understand the author's work and provide specific feedback.

Xiong et al. look for problems identified by reviewers in the author's work in peer reviews from the SWoRD system [11]. Xiong et al. use a bag-of-words, exact match approach to detect problem localization features. They use a shallow semantic match approach, which uses counts of nouns, verbs etc. in the text as features. Their approach does not incorporate relevance identification nor does it identify content type. Cho uses machine classification techniques such as naïve Bayes, support vector machines (SVM) and decision trees to classify review comments [12]. Cho manually breaks down every peer comment into idea units, which are then coded as a praise, criticism, problem detection, solution suggestion, summary or off-task comment.

Some other approaches used to study the usefulness of reviews are those by Turney [15], Dalvi [16] and Titov [17]. Peter D. Turney uses semantic orientation (positive or negative) to determine whether a review can be classified as recommended or not recommended. Turney's approach to differentiate positive from negative reviews involves identifying similarity between phrases containing adverbs and adjectives to terms "excellent" and "poor". Turney uses semantic orientation to recommend products or movies. We also use semantic orientation (referred to as tone) to identify the degree of sensitivity with which reviewers convey their criticisms.

Lim et al. identify reviewers who target e-commerce products and applications, and generate spam reviews [18]. The problem of spamming may be analogous to the problem of copy-pasting text in order to game the automated assessment system into giving reviewers high scores on their reviews. Therefore, we use a metric to detect plagiarized reviews.

There exist research works that discuss metrics that are important in review quality identification, and some that apply shallow approaches to determine quality. However, there is no work that takes factors such as relevance, content type, tone, quantity and plagiarism into consideration while determining review quality. Our sys-

Table 1: Some examples of reviews.

S No.	Review
1	"The example needs work."
2	"Yes, the organization is poor."

tem is an amalgamation of existing research in the said areas. In the next section, we provide an overview of the different review quality metrics.

3. AUTOMATED REVIEW ASSESSMENT

In order to assess quality, reviews have to be represented using metrics that capture their most important features. In general a good review contains: (1) coherent and well-formed sentences, which can be easily comprehended by the author, and (2) sufficient amount of feedback. In this section we discuss the metrics we use to assess reviews.

3.1 Review relevance

Reviewers may provide vague, unjustified comments. Comments in Table 1 are generic, and do not refer to a specific object in the text under review. For instance, what type of "work" does the "example" need or, what is poor about the "organization"? These reviews are ambiguous, and need to be supported with more information. Also, how do we know if the review has been written for the right submission, for instance any article may contain an example. Review relevance helps identify if a review is talking about the right submission.

We identify relevance in terms of the semantic and syntactic similarities between two texts. We use a word order graph, whose vertices, edges and double edges help determine structure-based match across texts. We use WordNet to determine semantic relatedness. Our approach has been described in Ramachandran and Gehringer [19].

3.2 Review content

A review is expected to provide an assessment of the kind of work that was done - praising the submission's positive points, identifying problems, if any, and offering suggestions on ways of improving the submission. Review examples in Table 1 do not provide any details. Reviews must identify problems in the author's work, and provide suggestions for improvement in order to be useful to authors, thus helping them understand where their work is lacking or how it can be improved. Content of a review identifies the type of feedback provided by the reviewer. We look for the following types of content in a review:

- **Summative** - Summative reviews contain either a positive or a neutral assessment of the author's submission. *Example:* "I guess a good study has been done on the tools as the content looks very good in terms of understanding and also originality. Posting reads well and appears to be largely original with appropriate citation of other sources."
- **Problem detection** - Reviews in this category are critical of the author's submission and point out problems in the submission. *Example:* "There are few references used and there are sections of text quoted that appear to come from a multitude of web sites." However, problem detection reviews only find problematic instances in the author's work, and do not offer any suggestions to improve the work.
- **Advisory** - Reviews that offer the author suggestions on ways of improving their work fall into this category. *Example:* "Although the article makes use of inline citations which is a plus, there are only a few references. Additional references could help support the content and potentially provide the examples needed." Advisory reviews display an understanding of the author's work, since the reviewer has taken the effort to provide the author with constructive feedback.

Different types of review content have different degrees of usefulness. For instance summative reviews provide only summaries of the author's work and are less useful to the author, whereas reviews that identify problems in the author's work or provide suggestions can be used by authors to improve their work, and are hence considered more important. We use a cohesion-based pattern identification technique to capture the meaning of a class of reviews.

3.3 Review tone

Tone refers to the semantic orientation of a text. Semantic orientation depends on the reviewer's choice of words and the presentation of a review. Tone of a review is important because while providing negative criticism reviewers might unknowingly use words that may offend the authors. Therefore we use tone information to help guide reviewers. A review can have one of three types of tones - positive, negative or neutral. We look for positively or negatively oriented words to identify the tone of a review [15]. We use positive and negative indicators from an opinion lexicon provided by Liu et al [20] to determine the semantic orientation of a review. Semantic orientation or tone of the text can be classified as follows:

- **Positive** - A review is said to have a positive tone if it predominantly contains positive feedback, i.e., it uses words or phrases that have a positive semantic orientation. *Example:* "The page is very well-organized, and the information is complete and accurate." Adjectives such as "well-organized", "complete" and "accurate" are good indicators of a positive semantic orientation.
- **Negative** - This category contains reviews that predominantly contain words or phrases that have a negative semantic orientation. Reviews that provide negative criticism to the author's work fall under this category, since while providing negative remarks reviewers tend to use language that is likely to offend the authors. Such reviews could be morphed or written in a way that is less offensive to the author of a submission. *Example:* "The examples are not easy to understand and have been copied from other sources. Although the topic is Design Patterns in Ruby, no examples in Ruby have been provided for Singleton and Adapter Pattern."

  The given example contains negatively oriented words or phrases such as "not easy to understand" ,"copied", "no examples". Review segment "..have been copied from other sources..." implies that the author has plagiarized, and could be construed by the author as a rude accusation. One of the ways in which this review could be re-phrased to convey the message, so as to get the author to acknowledge the mistake and make amends, is as follows - "The topic on Design Patterns in Ruby could be better understood with more examples, especially for the Singleton and Adapter patterns. Please try to provide original examples from your experience or from what was discussed in class."

- **Neutral** - Reviews that do not contain either positively or negatively oriented words or phrases, or contain a mixture of both are classified into this category. *Example:* "The organization looks good overall. But lots of IDEs are mentioned in the first part and only a few of them are compared with each other. I did not understand the reason for that." This review contains both positively and negatively oriented segments, i.e., "The organization looks good overall" is positively oriented, while "I did not understand the reason for that." is negatively oriented. The positive and negatively oriented words when taken together give this review a neutral orientation.

In case of both content and tone, a single review may belong to multiple categories. For instance consider the review, "Examples provided are good; a few other block structured languages could have been talked about with some examples as that would have been pretty useful to give a broader pool of languages that are block structured." While classifying for content, we see that the first part of the review, "Examples provided are good" praises the submission, while the remaining part of the review provides advice to the author. Our content identification technique identifies the amount of each type of content or tone (on a scale of 0 - 1) a review contains. Similarly in the case of tone, we identify the degree of positive, negative or neutral orientation of each review.

3.4 Review quantity

Text quantity is important in determining review quality since a good review provides the author with sufficient feedback. We plan on using this metric to indicate to the reviewer the amount of feedback they have provided in comparison to the average review quantity (from other reviewers of the system), thus motivating reviewers to provide more feedback to the authors. We identify quantity by taking a count of all the unique tokens in a piece of review. For instance, consider the following review, "The article clearly describes its intentions. I felt that section 3 could have been elaborated a little more." The number of unique tokens in this review is 15 (excluding articles and pronouns).

3.5 Plagiarism

Reviewers tend to refer to content in the author's submission in their reviews. Content taken from the author's submission or from some external source (Internet) should be placed within quotes in the review. If reviewers copy text from the author's submission and fail to place it within quotes (knowingly or unknowingly) it is considered as *plagiarism*.

Each of the review quality metrics listed is determined independently, and are integrated into a complete review quality assessment system. Reviewers are given feedback on each of these listed metrics, so that they get a complete picture of the completeness and quality of their review.

4. USER EXPERIENCE STUDY

We decided to study the experience of using an automated metareview system, since different types of reviewers - students, teaching assistants and faculty may use this feature. We study the extent to which users of an automated quality assessment system would perceive it to be important, and the output of the system to be useful. The study is important because it helps us understand whether reviewers learn and benefit from such an automated metareview system. This study also helps us learn what aspects of the feature can be improved, by identifying what the surveyed reviewers liked or disliked about the feature. A positive experience from using this feature may mean that reviewers would be more inclined to use it to improve their reviews.

According to Kuniavsky [21], user experience is "the totality of end-users' perceptions as they interact with a product or service. These perceptions include effectiveness (how good is the result?), efficiency (how fast or cheap is it?), emotional satisfaction (how good does it feel?), and the quality of the relationship with the entity that created the product or service (what expectations does it create for subsequent interactions?)." There exist several other definitions for the term *user experience* (abbreviated as UX) [22]. UXMatters[1] defines user experience as that which "Encompasses all aspects of a digital product that users experience directly - and perceive, learn, and use - including its form, behavior, and content." They also state that "Learnability, usability, usefulness, and aesthetic appeal are key factors in users' experience of a product." Therefore, apart from a study of factors such as user's perceptions, feelings or responses to a system, a user experience survey should also involve a study of the learning gained from a system and the usefulness of a system.

The aim of this study is to identify the degree of importance participants attach to each of the metareview metrics–review relevance, content, tone, quantity and plagiarism. This study will help us identify how effective the system is at helping reviewers learn about characteristics of their reviews.

5. EXPERIMENTS

To study the usefulness of our review quality assessment system we investigate the following broad research questions:

RQ1: *Do automated metareviews provide useful feedback?*
RQ2: *Which of the review quality metrics are more or less important than the others?*
RQ3: *Which of the review quality metrics' output did the reviewers find more or less useful when compared to the others?*

5.1 Participants

In order to identify how useful users of the automated metareview feature find it to be, we recruited 24 participants to (1) use the feature on Expertiza and (2) provide us with information on their experience by filling out a survey. Participants were recruited with an email message, which explained to them the purpose of the study. The set of participants included 15 doctoral students, 3 masters' students and 1 undergraduate student, all of whom were from the computer science department at North Carolina State University, and 5 research scientists from academia and industry.

5.2 Data collection

Our data collection process involved two steps. In the first step, participants were asked to use the automated metareview feature on Expertiza. They use the system to write a review for an article. For our study, we chose a wiki article on *Software Extensibility*[2]. We chose this article since we were recruiting subjects from the field of computer science, and Software Extensibility is a topic most computer science students or researchers are familiar with. A detailed

[1]UXMatters - User experience definition - http://www.uxmatters.com/glossary/

[2]Software Extensibility - `https://en.wikipedia.org/wiki/Extensibility`

Table 2: Detailed set of instructions to help complete the survey

1.	Use username/password to log into Expertiza.
2.	Click on assignment "User Study"
3.	Click on "Others' Work" (Since you will be reviewing someone else's work.)
4.	Click on "Begin" to start the review.
5.	Click the url under the "Hyperlinks" section. Read the article on Software Extensibility. Please keep in mind that you are reviewing this article.
6.	Answer questions on the review rubric describing the quality of the article you read. After answering all the review questions, click on the "Save Review" button.
7.	Wait for a few minutes for the system to generate the automated feedback.
8.	Fill out the **user-experience questionnaire**.

set of instructions was provided to each of the participants to help them complete the study (Table 2).

A review rubric is provided to the participants to help them write the review. The rubric contains questions on the organization, originality, clarity and coverage of the article under review. The rubric also evokes information on quality of the definitions, examples and links found in the article.

When participants submit their reviews, they are presented with automated feedback from our system. This feedback gives them information on different aspects of their review such as (1) content type, (2) relevance of the review to the article, (3) tone, (4) quantity of text and (5) presence of plagiarism. A screenshot of the output is available in Figure1. The participant reviewer reads and understands the metareview feedback.

In the second step of data collection, the participant reviewer is asked to fill out a user experience questionnaire (Step 8 in Table 2). The user experience questionnaire is a big part of this study, and has been explained in detail in Section 6.

6. USER EXPERIENCE QUESTIONNAIRE

The user experience questionnaire consists of four sections - *participant background, importance of reviews, importance of metrics, usefulness of system's output*. The questions we use in our user experience survey are discussed in the following sections. Answers to each of these questions are given on a scale of 1 (lowest) to 5 (highest).

6.1 Participant background

In the *background* section, participants were questioned about their experience in writing reviews, and in their experience with using peer-review systems such as Expertiza. The exact questions were:

Q1: *Do you have prior reviewing experience?*
Q2: *Do you have prior experience using the Expertiza system?*
Q3: *Have you used a peer-review system before?*
Q4: *Are you a(n): Undergraduate, Masters or PhD student, or Other?*

6.2 Importance of reviews and metareviews

In the *importance* section, we questioned participants on the importance of reviews and metareviews to a system.

Q5: *How important do you think reviews are in a decision-making process?*
Q6: *How important do you think metareviews (review of a review) are in a decision-making process?*

Answers are given on a 5-point scale - *unimportant, somewhat important, neutral, important* and *extremely important*. This section also includes an open question to gather textual feedback from participants. All these questions are optional, i.e., the participant may choose not to respond to any of them.

We also gauge whether participants would be motivated to use reviews to improve the quality of their submission (as an author), and metareviews to improve the quality of their reviews (as a reviewer). We therefore included the following questions in the questionnaire:
Q7: *Would better reviews inspire you to use the feedback in your revisions?*
Q8: *Would automated metareviews motivate you to update your reviews?*
Q9: *Do the automated metareviews provide useful feedback?*

6.3 Importance of metareview metrics

In the *importance of metrics* section we identify how important participants think the different metareview metrics are in gauging the quality of a review.

Q10: *How important do you think each of the review quality metrics is in learning about the quality of your review? 1. Review relevance, 2. Review content 3. Tone 4. Quantity 5. Plagiarism*

The answers are given on a 5-point scale. This question helps us identify the metrics to which users of the system attach most importance, or to which ones they attach the least importance. This section also allows participants to provide any additional comments, to learn about the participants' opinions of the different metrics, or any other related information.

6.4 Usefulness of system's metareview output

This section helps us study the usefulness of the system's outputs. These questions gauge whether reviewers learned something about their review's quality from the automated feedback.

Q11: *How useful do you think the output from each of the review quality metrics is (from what you saw on Expertiza)? 1. Relevance, 2. Review content 3. Tone 4. Quantity 5. Plagiarism*

Answers are given on a 5-point scale and range from *not useful, somewhat useful, neutral, useful* or extremely useful. The ratings indicate usefulness of the chosen design for the system' output. These questions help us learn whether participants are able to successfully comprehend the meaning of the system's output. This information coupled with the information from the previous question on *importance of metrics* would help us identify the set of metrics that need improving. This section also includes an open question to gather any other comments participants may have on the system's output.

6.5 Other metrics

We included an open question on the survey to learn about any other review quality metrics, which participants think would be useful in an automated metareview system.

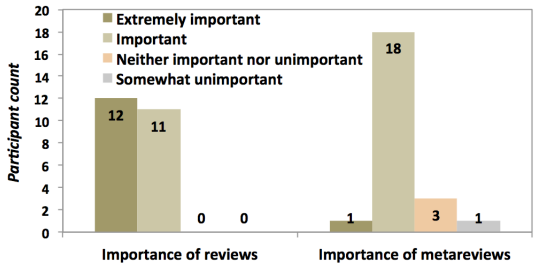


Figure 2: Participants' rating of importance of reviews and metareviews.

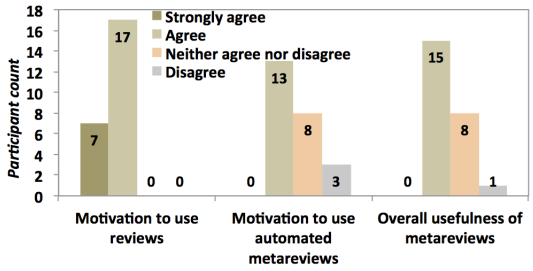


Figure 3: Participants' rating of motivation to use reviews and metareviews to improve the quality of their submission or review respectively. The chart also contains participants' estimation of usefulness of the automated metareview feature's output.

Q12: *What other information do you think might help you improve your review quality? Are there any specific review features you would like to get feedback on? e.g. language of the review, grammar, vocabulary, or nothing else*

The next section discusses our analyses on the collected data.

7. ANALYSIS OF DATA

In this section we discuss some of the findings from our data. Out of the 24 participants, 19 had prior reviewing experience. Only 7 of the participants had prior experience with the Expertiza system.

7.1 Importance of reviews and metareviews

All of the participants agreed that reviews play an important role in the decision-making process (Figure 2). A majority of the participants also agreed on the importance of metareviews (review of reviews). One participant did not respond to these questions.

We asked participants whether good quality reviews would motivate them to fix their submission. All participants agreed (7 agreed strongly) that they would incorporate suggestions from the feedback in their work (Figure 3). We asked participants whether automated feedback on their reviews would inspire them to improve their reviews. Out of the 24 participants 13 agreed that they would use the automated feedback. However 8 participants displayed doubt in the use of automated metareview feedback by answering *neither agree nor disagree*. A small number said that they would not be inclined to use the automated metareview feedback to improve their reviews.

Thus we see that as authors, participants agree that good quality feedback would motivate them to fix their work, but as reviewers they may not be inclined to use metareview feedback to update their reviews (and help other authors improve their work). The concept of automated assessment of reviews is new, and a lack of understanding of the purpose of these metrics could be one of the reasons why reviewers felt that automated metareviews may not motivate them to fix their reviews.

7.2 Importance of the review quality metrics

We analyze how participants judge each of the automated metrics' importance. The results are displayed in Figure 4. The metric, which participants rated as the most important is *relevance*. Out of the 24 participants 23 agree that relevance is important in assessing the quality of a review (3 thought it was extremely important). The next most important metric was found to be *review content*, with

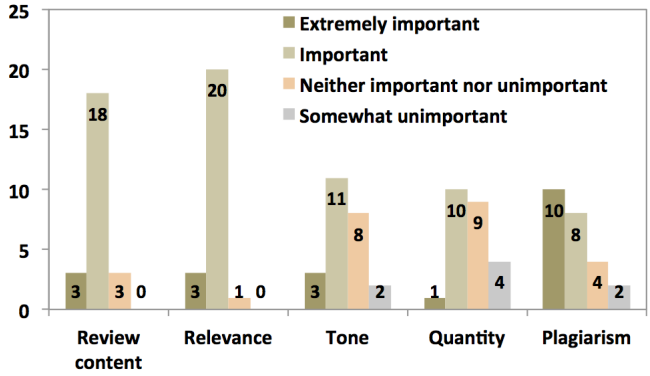


Figure 4: Participants' rating of the importance of each review quality metric.

21 of the participants agreeing on its importance (3 thought it was extremely important).

Participants found quantity to be the least important metric, with 9 of them expressing doubts on its usefulness (neither important nor unimportant) and 4 of them describing it as somewhat unimportant. Wilcoxon rank-sum test is used to determine if two metrics' ratings have identical distributions (null hypothesis) [23]. We use this test to compare metric quantity with metrics relevance and content (which have been identified as the most important metrics) at 0.05 significance level. The p value for the test on metrics quantity and relevance is 0.0003, and for metrics quantity and content is 0.002. Since these p values are < 0.05, we conclude that quantity's ratings are significantly different from those of the most important metrics - relevance and content.

Quantity contains the number of unique tokens in a review text, and is meant to motivate reviewers to write more feedback. Quantity may be obvious to a reviewer, since they are aware of the amount of feedback they have provided. Hence quantity may turn out to be the least effective, when compared with the other metrics, in conveying any new information to the reviewer. This could be why quantity is ranked as the least important quality metric.

7.3 Usefulness of system output

We questioned participants on the usefulness of the system's metareview output, to study how informative or understandable they find

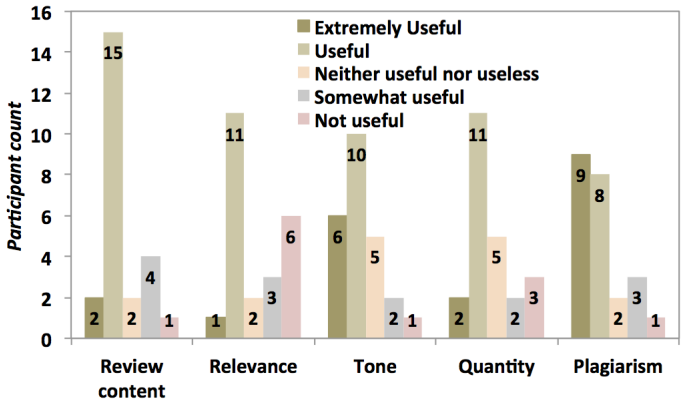


Figure 5: Participants' rating of the usefulness of each review quality metric.

it. The results of studying usefulness of metrics are displayed in Figure 5. The metrics participants rated as most useful are *plagiarism* and *review content*, with 17 of participants (9 found plagiarism extremely useful, and 2 found content extremely useful) agreeing that these metrics were useful in helping them understand where their reviews are lacking.

Tone is the second most useful metric with 16 of the participants agreeing on its usefulness, despite having 8 participants judging it to be neither important nor unimportant (from previous section). Similarly in the case of quantity, 13 of the participants found the systems' output for quantity to be useful (2 of them thought it was extremely useful), although 9 of the participants said that they thought it to be neither important nor unimportant (Figure 4).

We use the Wilcoxon test (at a significance level of 0.05) to determine if there is a significant difference (increase) in the distribution of the importance and usefulness ratings of quantity. We selected pairs, whose ratings for usefulness showed an increase from their corresponding importance ratings. The ratings have a p value of $0.03 < 0.05$, which indicates that the increase in usefulness ratings is significant. Similarly, when identifying the significance of increase between the importance and usefulness ratings of tone, we get a p value of 0.09. Although this is not < 0.05, we see that the low p value may be indicative that the improvement in usefulness ratings is not a chance occurrence (i.e., it is significant). Thus we see that although participants thought initially that tone and quantity may not be important to a metareview assessment system, they found the output from the system for these two metrics to be insightful.

Despite being judged as the most important review assessment metric only 12 of the participants found the output of the relevance metric to be useful. One of the participants expressed difficulty in interpreting the meaning of the relevance score. Our metareview feedback contains only real-valued scores in the range 0 - 1, which may not have been very useful to the reviewer in understanding the degree of relevance. This could have caused the relevance's usefulness ratings to be lower when compared to the ratings of metrics such as plagiarism, which contains true/false as output.

In the future we are planning to improve the format of the output by providing textual feedback in addition to the numeric feedback. The feedback will point to specific instances of a review that need improvement. This may make it easy for reviewers to interpret the numeric score, and maybe further motivate reviewers to use the information to improve their reviews.

7.4 Other metrics

Some of the other metrics that participants exclaimed their interest in are the *grammar and syntax* of reviews. One of the participants suggested the use of *sentence structure variability* across sentences as a means of assessing a review. The participant suggested that though short phrases may succeed in communicating the idea, they may not succeed in conveying the complete thought. The presence of well-structured sentences in a review may help the author comprehend the content of a review with ease. Well-structured sentences also indicate to authors that the reviewer put in a lot of thought and effort into writing the review. Similarly in the case of another suggested metric - *word complexity*.

Another metric suggested by a participant is *text cohesion*. Reviews sometimes contain a set of sentences, which may appear to be disconnected, i.e., lack a meaningful flow from one sentence to the next. Cohesive text help make reading and understanding reviews easier.

7.5 Usefulness of the overall automated assessment feature

We surveyed participants on the usefulness of the overall automated feedback system. Out of 24 participants 15 agreed that the feedback was useful (Figure 3), and 8 neither agreed nor disagreed.

One of the participants exclaimed concern with the use of plagiarism as a metric to assess reviews. This is likely because the participant did not see the motivation for a reviewer to plagiarize while writing reviews. Students on Expertiza are evaluated (given scores) on the quality of the reviews they write. Hence they do have a motivation to copy either other good quality reviews (available online) or chunks of text from the submission and submit them as a good quality review. Plagiarism could be caught by manual metareviewers, but may be missed by an automated system. Hence we have this additional feature to ensure that reviewers do not try to game the system by copying reviews.

8. THREATS TO VALIDITY

During the evaluation we noticed that a majority of the participants did not have prior experience in using Expertiza, which could have affected their overall performance.

We also learned, from the comments section of the questionnaire, that a few of the participants did not fully understand the meaning of some of the metrics. An understanding of the purpose of the metareview metrics is essential to assessing their importance and the output's usefulness. Hence, a lack of complete understanding of the metrics may pose as a threat to our results.

No textual reviews were provided by 4 of the participants, which means that the system outputs a value of 0 for each of the metareview metrics. Participants may not be able to discern the usefulness of metrics' outputs for which they have received a score of 0. These are some of the threats to the validity of our results.

9. FUTURE DIRECTIONS

In the future we plan on doing the following: (1) improve the display of metareview output to the reviewer, (2) identify the usefulness of other metareview metrics, (3) study the degree of agreement of the automated metareview ratings with human-provided metareview ratings, and (4) study improvement in reviewing skills.

In order to improve the system's metareview output we plan to highlight snippets of the review that need to be updated. Two participants suggested the need for additional information on metrics such as problem detection and solution suggestion. We plan to provide information on specific instances (of the author's work), which the reviewer needs to read and assess to identify problems or provides suggestions. Also, providing feedback to reviewers with samples of good quality reviews may help them learn how to fix their reviews.

We plan on investigating the use of other metrics such as sentence structure, cohesion and word complexity (discussed in Section 7.4) to study a review's quality. At present our graph-based representations capture sentence structure (e.g. subject-verb-object), but we do not study cohesion across sentences in a review. A study of cohesion may involve exploring other areas of natural language processing such as anaphora resolution [24].

We plan on investigating the extent to which the output from the automated metareview system, as a whole, agrees with human-provided values. This will help us determine whether the system would do as good a job of metareviewing i.e., be as good as human metareviewers in assessing reviews.

We would also like to study if reviewers who get feedback from the system show signs of improvement, i.e., if their reviewing skill improves with time. This would indicate that reviewers learn from the system's feedback to provide more specific and more useful reviews to authors. We would also like to investigate the impact a review quality assessment system has on the overall quality of the authors' submissions.

10. CONCLUSION

Assessment of reviews is an important problem in education, as well as science and human resources, and so it is worthy of serious attention. This paper introduces a novel review quality feature, which uses metrics such as review content type, relevance, tone, quantity and plagiarism to assess reviews. This feature is integrated into Expertiza, a collaborative web-based learning application. We surveyed 24 participants on the importance of the metrics and usefulness of the review quality assessment's output. Results indicate that participants found review relevance to be most important in assessing review quality, and system output from metrics such as review content and plagiarism to be most useful in helping them learn about their reviews.

11. REFERENCES

[1] K. Cho and C. D. Schunn, "Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system," *Computer Education*, vol. 48, pp. 409–426, April 2007.

[2] E. F. Gehringer, L. M. Ehresman, and W. P. Conger, S.G., "Reusable learning objects through peer review: The expertiza approach," in *Innovate: Journal of Online Education*, 2007.

[3] E. F. Gehringer, "Expertiza: Managing feedback in collaborative learning." in *Monitoring and Assessment in Online Collaborative Environments: Emergent Computational Technologies for E-Learning Support*, 2010, pp. 75–96.

[4] R. Rada, A. Michailidis, and W. Wang, "Collaborative hypermedia in a classroom setting," *J. Educ. Multimedia Hypermedia*, vol. 3, pp. 21–36, January 1994.

[5] C. KÃijhne, K. BÃűhm, and J. Z. Yue, "Reviewing the reviewers: A study of author perception on peer reviews in computer science." in *CollaborateCom'10*, 2010, pp. 1–8.

[6] P. Wessa and A. De Rycker, "Reviewing peer reviews: a rule-based approach," in *International Conference on E-Learning (ICEL)*, 2010, pp. 408–418.

[7] J. Burstein, D. Marcu, and K. Knight, "Finding the write stuff: Automatic identification of discourse structure in student essays," *IEEE Intelligent Systems*, vol. 18, pp. 32–39, January 2003.

[8] P. W. Foltz, S. Gilliam, and S. A. Kendall, "Supporting content-based feedback in online writing evaluation with LSA," *Interactive Learning Environments*, vol. 8, pp. 111–129, 2000.

[9] S. van Rooyen, N. Black, and F. Godlee, "Development of the review quality instrument (rqi) for assessing peer reviews of manuscripts," *Journal of Clinical Epidemiology*, vol. 52, no. 7, pp. 625 – 629, 1999.

[10] M. M. Nelson and C. D. Schunn, "The nature of feedback: How different types of peer feedback affect writing performance," in *Instructional Science*, vol. 27, 2009, pp. 375–401.

[11] W. Xiong, D. J. Litman, and C. D. Schunn, "Assessing reviewer's performance based on mining problem localization in peer-review data." in *EDM*, 2010, pp. 211–220.

[12] K. Cho, "Machine classification of peer comments in physics," in *Educational Data Mining*, 2008, pp. 192–196.

[13] R. Zhang and T. Tran, "Review recommendation with graphical model and em algorithm," in *Proceedings of the 19th international conference on World wide web*, ser. WWW '10, 2010, pp. 1219–1220.

[14] S. Moghaddam, M. Jamali, and M. Ester, "Review recommendation: personalized prediction of the quality of online reviews," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, ser. CIKM '11, 2011, pp. 2249–2252.

[15] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002.

[16] N. Dalvi, R. Kumar, B. Pang, and A. Tomkins, "Matching reviews to objects using a language model," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, ser. EMNLP '09, 2009, pp. 609–618.

[17] I. Titov and R. McDonald, "Modeling online reviews with multi-grain topic models," in *Proceedings of the 17th international conference on World Wide Web*, ser. WWW '08, 2008, pp. 111–120.

[18] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, ser. CIKM '10, 2010, pp. 939–948.

[19] L. Ramachandran and E. F. Gehringer, "A word-order based graph representation for relevance identification [poster]," *CIKM 2012, 21st ACM Conference on Information and Knowledge Management*, October 2012.

[20] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the web," in *Proceedings of the 14th International Conference on World Wide Web*, 2005, pp. 342–351.

[21] M. Kuniavsky, *Smart Things: Ubiquitous Computing User Experience Design: Ubiquitous Computing User Experience Design.* Morgan Kaufmann, 2010.

[22] E. L.-C. Law, V. Roto, M. Hassenzahl, A. P. Vermeeren, and J. Kort, "Understanding, scoping and defining user experience: a survey approach," in *Proceedings of the 27th international conference on Human factors in computing systems*. ACM, 2009, pp. 719–728.

[23] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[24] E. Tognini-Bonelli, "Corpus linguistics at work," *Computational Linguistics*, vol. 28, no. 4, pp. 583–583, 2002.

Effects of Automatically Generated Hints on Time in a Logic Tutor

First Author
University
Address
Address
Email@email.com

Second Author
University
Address
Address
Email@email.com

Third Author
University
Address
Address
Email@email.com

ABSTRACT

This work explores the effects of using automatically generated hints in Deep Thought, a propositional logic tutor. Generating hints automatically removes a large amount of development time for new tutors, and it also useful for already existing computer-aided instruction systems that lack intelligent feedback. We focus on a series of problems, after which, the control group is known to be 3.5 times more likely to cease logging onto an online tutor when compared to the group who were given hints. We found a consistent trend in which students without hints spent more time on problems when compared to students that were provided hints. Exploration of the interaction networks for these problems revealed that the control group often spent this extra time pursuing buggy-strategies that did not lead to solutions.

1. INTRODUCTION

Problem solving is an important skill across many fields, including science, technology, engineering, and math (STEM). Working open-ended problems may encourage learning in higher 'levels' of cognitive domains [2]. Intelligent tutors have been shown to be as effective as human tutors in supporting learning in many domains, in part because of their individualized, immediate feedback, enabled by expert systems that diagnose student's knowledge states [10]. However, it can be difficult to build intelligent support for students in open problem-solving environments. Intelligent tutors require content experts and pedagogical experts to work with tutor developers to identify the skills students are applying and the associated feedback to deliver [7].

In problem solving environments where students complete many diverse steps to solve a single problem, even labeling all correct and incorrect approaches is a large burden. There are many computer-based educational problem-solving environments, that have already been developed and can benefit from data-driven approaches to providing intelligent feedback. We hope to contribute toward data-driven techniques to automatically generate intelligent feedback based on previously recorded data from such environments, as well as methods to visualize and analyzes the large amounts of data present in student-log files.

Barnes and Stamper built an approach called the Hint Factory to use student data to build a graph of student problem-solving approaches that serves as a domain model for automatic hint generation [8]. Hint factory has been applied across domains [6]. Stamper et al. found that the odds of a student in the control group dropping out of the tutor were 3.5 times more likely when compared to the group provided with automatically generated hints [9]. The hints also affected problem completion rates, with the number of problems completed in L1 being significantly higher for the hint group by half of a standard deviation, when compared to the control group. Eagle and Barnes have abstracted this domain model into an Interaction Network for problem-solving data analysis. Their preliminary results show that applying graph mining techniques to Interaction Networks can help uncover useful clusters that represent diverse student approaches to solving a particular problem [5].

2. THE DEEP THOUGHT TUTOR

We perform our analysis on data from the Deep Thought propositional logic tutor [3]. Each problem provides the student with a set of premises, and a conclusion, and asks students to prove the conclusion by applying logic axioms to the premises. Deep Thought allows students to work both forward and backwards to solve logic problems [4]. Working backwards allows a student to propose ways the conclusion could be reached. For example, given the conclusion B, the student could propose that B was derived using Modus Ponens (MP) on two new, unjustified propositions: $A \rightarrow B, A$. This is like a conditional proof in that, if the student can justify $A \rightarrow B$ and A, then the proof is solved. At any time, the student can work backwards from any unjustified components, or forwards from any derived statements or the premises.

2.1 Data

We perform our experiments on the Spring and Fall 2009 Deep Thought logic tutor dataset as analyzed by Stamper, Eagle, and Barnes in 2011[9]. In this dataset, three different professors taught two semesters each of an introduction to logic course, with each professor teaching one class with hints available and one without hints in the Deep Thought tutor. In the spring semester there were 82 students in the Hint group and 37 students in the Control group; in the fall

semester there were 39 students in the Hint group and 83 in the Control group. Students for which application log-data did not exist were dropped from the study; resulting in 68 and 37 students in the Hint group, and 28 and 70 students in the Control group for the first and second semesters respectively. This results in a total of 105 students in the Hint group and 98 students in the Control group. Students from the 6 sections of an introduction to logic course were assigned 13 logic proofs in the deep thought tutor. The problems are organized into three constructs: level one (L1) consisting of the first 6 problems assigned; level two (L2) consisting of the next 5 problems assigned; and level three (L3) consisting of the last two problems assigned. We refer to the group that received hints as the Hint group, and the group that did not receive hints as the Control group.

We are interested in the usage of hints from students in the hint group. Deep Thought has been modified to include John Stamper's Hint Factory [1], and provides four levels of automatically generated hints. The first level suggests the premise to be used, the second level provides more content, the third level provides the logic rule to be applied, and the fourth hint is the bottom-out hint explaining the exact procedure. We investigated two different components regarding hint usage in Deep Thought. The first is the average number of hints per level, per problem. That is, for example, the number of level two hints requested on problem 1-4. We also investigated hint coverage in the Deep Thought tutor as provided by the Hint Factory for each problem and the overall. In Deep Thought, the Hint Factory can either generate a hint, in which case all four levels of hints are generated or a hint cannot be generated in which case no hints will exist for some given step in the problem.

3. RESULTS

In order to investigate the increased rate of drop-out between the hint group and the control group. We concentrate on the first 5 problems from L1 of the Deep Thought Tutor. We focus here as, while the groups started with similar completion and attempt rates, after level 1 the groups diverge on both completion and problem attempt rates. Since investigation of the interaction networks for these problems revealed that the control group often pursue buggy-strategies, which do not result in solving the problem, we hypothesized that their would be differences in the amount of time spent in tutor between the groups.

We performed analysis on the student-tutor interaction logs. For each student we calculated the summation of their elapsed time per interaction. To control for interactions in which the student may have idled we filtered any interactions that took longer than 10 minutes. The descriptive statistics for this are located in table 1, Prob represents the problem number, H and C represent the Hint group and the Control group.

The large standard deviations are a sign that perhaps this data is not normal. Exploring the data with Q-Q plots reveals that the data is in fact, not normally distributed. This prevents us from performing between-group statistical tests, such as the student's t-test, as our data violates the assumption of normality. To normalize the data, we use a logarithmic transformation (common log) to make the data more symmetric and homoscedastic. Observation of the Q-Q plot and histogram of the transformed data reveal that we had addressed the normality concerns. The results are presented in table 2.

Table 1: Descriptive Statistics for Time (in seconds) Spent in Each Problem

	N		M		SD	
Prob	H	C	H	C	H	C
1.1	104	93	765.89	1245.24	956.41	1614.30
1.2	88	76	761.65	1114.37	911.24	1526.91
1.3	90	67	664.17	1086.09	733.95	2119.19
1.4	87	71	754.60	1266.39	1217.06	1808.53
1.5	84	67	710.62	1423.22	1192.43	2746.54

Table 2: Descriptive Statistics After Common Log Transformation

	N		M		SD	
Prob	H	C	H	C	H	C
1.1	104	93	2.63	2.79	0.48	0.55
1.2	88	76	2.59	2.73	0.54	0.54
1.3	90	67	2.62	2.72	0.44	0.48
1.4	87	71	2.66	2.89	0.40	0.41
1.5	84	67	2.55	2.75	0.48	0.60

To test for differences between the two groups on each problem, we subjected the common log transformed data to t-test. The results from this test are presented in table 3. There are significant differences for problems 1, 4, and 5. The ratio is calculated by taking the difference between the hint group mean and the control group mean. As $\lg(x) - \lg(y) = \lg(\frac{x}{y})$ the confidence interval from the logged data estimates the difference between the population means of log transformed data. Therefore, the anti-logarithms of the confidence interval provide the confidence interval for the ratio. We provide the C:H ratios and confidence intervals in table 4.

Table 3: Ratio Between Groups (H:C) in the Original Scale

		95% Confidence Interval			
Prob	Ratio	low	high	p-value	t
1.1	0.69	0.50	0.97	0.03	-2.18
1.2	0.72	0.49	1.06	0.10	-1.68
1.3	0.78	0.56	1.10	0.15	-1.43
1.4	0.58	0.44	0.78	0.00	-3.61
1.5	0.62	0.42	0.93	0.02	-2.31

In order to explore what these differences mean, we shall transform the data back to our original scale (seconds.) The transformed data is provided in table 5. These are the Geometric Means, which are often closer to the original median, than they are the mean. The ratios from tables 3 and 4 are easily interpreted as the log of the ratio of the geometric means. For example in problem 1.4, in the common log scale, the mean difference between hint and control group is -0.23. Therefore, our best estimate of the ratio of the hint time and control time is $10^{-.23} = 0.58$. Our best estimate of the effect of Hint is it takes 0.58 times as many seconds as the control group to complete the problem. The confidence interval reported above is for this difference ratio.

Table 4: Ratio Between Groups (C:H) in the Original Scale

		95% CI	
Prob	Ratio	low	high
1.1	1.44	1.04	2.01
1.2	1.39	0.94	2.05
1.3	1.27	0.91	1.78
1.4	1.71	1.28	2.30
1.5	1.60	1.07	2.40

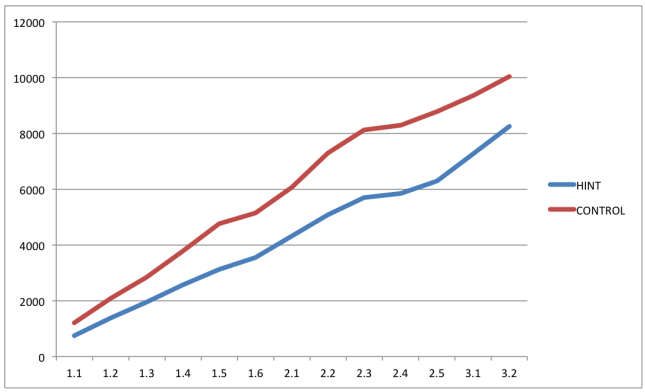


Figure 1: Cumulative average time (in seconds) per problem across the tutor.

The geometric mean of the amount of seconds needed to solve problem 1.4 for the hint group is 0.58 (95% CI: 0.44 to 0.78) times as much as that needed for students in the control group. Stated alternatively, students in the control group spend 1.71 (95% CI: 1.07 to 2.40) times as long as the Hint group in problem 1.4.

Table 5: Geometric Means and Confidence Intervals in Seconds

		95% CI			95% CI	
P	H	low	high	C	low	high
1	428.66	347.14	529.31	618.19	478.60	798.51
2	387.07	297.97	502.82	537.80	405.75	712.82
3	413.80	335.89	509.78	527.18	405.05	686.13
4	454.43	374.38	551.61	778.01	624.48	969.29
5	352.90	278.06	447.89	565.61	405.34	789.24

Exploring the total time spent between all five problems also required a log transformation. The total time spent on the first 5 problems between the hint group (M = 3.34, SD = 0.4) and the control group (M = 3.44, SD = 0.51) was not significant, $t(198) = 1.41, p = 0.16$. This corresponds to a H:C ratio of 0.81 (95% 0.60 to 1.09), and a C:H ratio of 1.24 (95% CI: 0.92 to 1.66).

In order to explore differences in overall time in tutor between the two groups, we subjected the total elapsed time on all 13 problems. The total time in tutor between the hint group (M = 3.75, SD = 0.43) and the control group (M = 3.72, SD = 0.58) was no significant, $t(200) = 0.40, p = 0.694$.

3.1 Hint Usage and Coverage

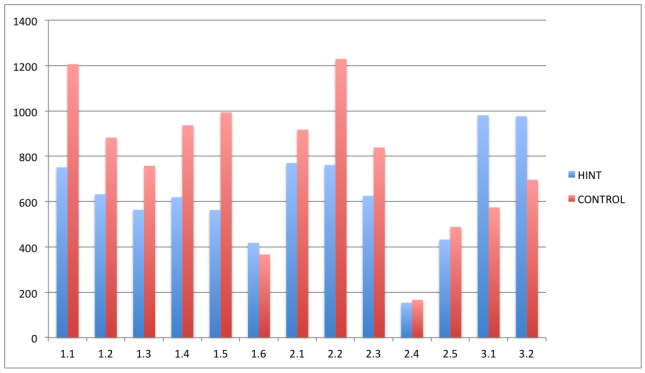


Figure 2: Average time (in seconds) spent per problem.

We investigated the average hint usage per student, per problem. Table 6 depicts the average number of hints per student for each hint level, for each problem. Note that these values are for a single problem, which requires multiple steps. This means that requesting a level four hint allows a student to skip a single step, of many, for a single problem and not an entire problem.

Table 6: Average Hint Use per Problem

Problem	H1	H2	H3	H4
1.1	1.61	0.94	0.66	0.23
1.3	1.79	1.46	1.13	0.77
1.4	2.96	1.66	1.18	0.32
2.2	3.44	2.27	2.04	1.08
2.3	5.56	3.09	2.44	1.00
2.4	1.45	0.99	0.90	0.51
2.5	3.66	1.91	1.66	0.88

In table 7 we provide the hint coverage for each problem. The hint coverage is calculated by taking the number of fulfilled hint requests divided by the number of total hint requests for a problem.

Table 7: Hint Coverage Rates

Problem	Hint Coverage
1.1	0.74
1.3	0.62
1.4	0.81
2.2	0.82
2.3	0.81
2.4	0.88
2.5	0.80
Overall	0.78

4. DISCUSSION

The results of this analysis show that students in the control group are overall not spending significantly more time in the tutor during these first five problems. However, the control does spend significantly more time in some problems compared to the hint group. Problems 1, 3 and 4 provided students with the automatically generated hints. While problem 2 and 5 had no hints for either group. We

would expect there to be differences in time to solve for the hint group, and this was the case for problem 1. We would also expect that having no hints on problem two would not display an effect, as the second problem is too early to expect differences to emerge between the groups. Problem 1.3 is interesting as this problem is the first in which the groups begin to show preferences towards different solution strategies. With the control group preferring to work backwards, and the hint group preferring to work forwards (hints are only available for solutions working forward). Problem 1.4 and 1.5, both of which showed significant differences in time spent, showed a large portion of control group student interactions to be perusing buggy-strategies.

This is interesting as the control group is spending at least as much, and often more, time in tutor and yet meeting with less overall success. The control students are not becoming stuck in a single bottleneck location within the problems and then quitting, which would result in lower control group times. The control students are actively trying to solve the problems using strategies that do not work. The hint group is able to avoid these strategies via the use of the hints. The hint group students also develop a preference for solving problems forward, as that is the direction in which they can ask for hints. It is interesting to see that these preferences remain, even when hints are not available.

The effect of the automatically generated hints appear to let the hint group spend around 60% of the time per problem compared to the control group. Or stated differently, the control group requires about 1.5 times as much time per problem when compared to the hint group. These results show that the hints provided by the Hint Factory, which are generated automatically, can provide large differences in how long students need to solve problems.

Regarding average hint use, table 6 suggests that problem 2.3 is likely the most difficult as it has the highest levels of hint usage for nearly all levels. Table 6 also suggests there is little gaming behavior occurring in the Deep Thought tutor from students. As previously stated a single problem requires multiple steps, so to see level four hints at values around one and below is encouraging.

5. CONCLUSIONS AND FUTURE WORK

This paper has provided evidence that automatically produced hints can have drastic effects on the amount of time that students spend solving problems in a tutor. We found a consistent trend in which students without hints spent more time on problems when compared to students that were provided hints. Exploration of the interaction networks for these problems revealed that the control group often spent this extra time pursuing buggy-strategies that did not lead to solutions. Future work will explore other data available on the interaction level, such as errors, in order to get a better understanding of what the control group is doing with their extra time in tutor. We will also look into the development of further interventions that can help students avoid spending time on strategies that are unlikely to provide solutions.

6. REFERENCES

[1] T. Barnes and J. Stamper. Toward automatic hint generation for logic proof tutoring using historical student data. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems (ITS 2008)*, pages 373–382, 2008.

[2] B. S. Bloom. *Taxonomy of Educational Objectives: The Classification of Educational Goals*. Taxonomy of educational objectives: the classification of educational goals. Longman Group, New York, 1956.

[3] M. J. Croy. Graphic interface design and deductive proof construction. *J. Comput. Math. Sci. Teach.*, 18:371–385, December 1999.

[4] M. J. Croy. Problem solving, working backwards, and graphic proof representation. *Teaching Philosophy*, 23:169–188, 2000.

[5] M. Eagle, M. Johnson, and T. Barnes. Interaction Networks: Generating High Level Hints Based on Network Community Clustering. *educationaldatamining.org*, pages 1–4.

[6] D. Fossati, B. Di Eugenio, S. Ohlsson, C. Brown, L. Chen, and D. Cosejo. I learn from you, you learn from me: How to make ilist learn from students. In *Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, pages 491–498, Amsterdam, The Netherlands, The Netherlands, 2009. IOS Press.

[7] T. Murray. Authoring intelligent tutoring systems: An analysis of the state of the art. *International Journal of Artificial Intelligence in Education (IJAIED)*, 10:98–129, 1999.

[8] J. Stamper, T. Barnes, L. Lehmann, and M. Croy. A pilot study on logic proof tutoring using hints generated from historical student data. *Proceedings of the 1st International Conference on Educational Data Mining (EDM 2008)*, pages 197–201, 2008.

[9] J. C. Stamper, M. Eagle, T. Barnes, and M. Croy. Experimental evaluation of automatic hint generation for a logic tutor. In *Proceedings of the 15th international conference on Artificial intelligence in education*, AIED'11, pages 345–352, Berlin, Heidelberg, 2011. Springer-Verlag.

[10] K. VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221, 2011.

Providing implicit formative feedback by combining self-generated and instructional explanations

Joseph Jay Williams[1]
[1]University of California at Berkeley
joseph_williams@berkeley.edu

Helen Poldsam
Talinn Tech
hpoldsam@gmail.com

Abstract. Formative feedback for a learner typically uses human or artificial intelligence to draw an inference about a learner's knowledge state from the learner's actions, and select a learner-directed response. To tackle cases when such intelligence is not easily available, we are exploring ways of providing implicit formative feedback: A learner's action is to respond to an explanation prompt, and the learner-directed response is to provide an instructional explanation. We consider explanations for correct examples to mathematics exercises, but the exciting implications will be for less well-defined domains that are challenging for cognitive tutors to model. To motivate learners to explain and to increase implicit feedback, we also explore prompts to compare the self-generated and instructional explanations.

Keywords: explanation, self-explanation, learning, comparison, formative feedback

1 Introduction

Traditionally, feedback to students has often been *summative*, such as midterm scores and state exams, where even the application of advanced psychometric techniques leads to measures that provide a summary assessment of some attribute. The pen- and paper- tests typically administered and the time needed for another human to grade and assess places a natural delay between a student's behavior(s) and the provision of feedback about that behavior.

Now that learners' behavior is increasingly in a computerized or online environment, there are three key implications. The first is that many tests and measures typically considered as (summative) assessments can be analyzed instantaneously and automatically. The second is that online digital environments allow for the *delivery* of sophisticated instruction and formative feedback. The third is that the constant logging of data on a computer means that a much wider range of student behaviors is available as fodder for 'assessments', which can then be analyzed and used to provide *formative feedback* to students.

As evidenced by the current workshop and extensive research in the learning sciences [1] [2] [3], great progress has been made in developing formative assessments and feedback. However, the issue of providing formative feedback raises two core challenges.

The first is that providing formative feedback that helps learning seems to be constrained by how accurately an automatic system can diagnose a learner's

knowledge state, infer what instructional tactic is likely to deliver formative feedback that moves the learner to a more effective knowledge state, and ensure the learner successfully uses this instruction or formative feedback. While there have been great strides in developing the data mining and artificial intelligence capacities to achieve all three of these goals, is there a way to mitigate these constraints through a complementary approach to the problem of providing formative feedback?

The second challenge is that – even if the above issues could be solved – learners may not learn general metacognitive skills of self-regulation – to identify gaps in their knowledge, consider how to fill them or seek out new information, and engage in effective learning strategies that move their understanding forward.

One potential way to address both of these issues is to provide information from which *learners* can generate *implicit* formative feedback, and structure the instructional environment to support learners in generating and using this feedback.

This paper outlines a paradigm for doing this and reports the design of an ongoing study. Learners are asked or prompted to self-generate explanations, then are provided with normative answers or instructional explanations that respond to the same prompt, and finally are guided to compare their self-generated explanations with the instructional explanations provided. This draws together work in education and psychology on the benefits of *self-explanation* [4] [5], on how to provide appropriate instructional explanations for students [6] [7], and on the benefits of comparison for learning [8].

Context for introducing implicit feedback: Worked example solutions in Khan Academy mathematics exercises

While the general framework can be applied to many contexts, the current study examines the generation, consideration, and comparison of explanations in the KhanAcademy.org exercise framework (www.khanacademy.org/exercisedashboard). This provides a large collection of mathematics exercises with a similar format, used by tens of thousands of students. It is therefore a widely applicable context in which to develop a paradigm for providing implicit feedback from self-generated and instructional explanations.

Figure 1 shows an example of an exercise we have augmented. The typical (non-augmented) exercise starts with a statement of a problem for the student to solve, which is outlined in the box surrounded by a dark black line. When ready, students can type in a proposed answer and then receive feedback on its correctness. At any point, students can also request a hint, which reveals the next step of a worked example solution to the problem. Students have to enter the correct problem to advance, but because every problem provides on-demand "hints" which step-by-step reveal the solution, they can eventually do so (the last step is simply the answer).

This design already builds in some form of implicit feedback, if it is assumed that students first try to consider steps in the problem's solution before requesting hints. A hint or solution step can therefore give them implicit feedback about the appropriateness of what they were considering before.

Incorporating self-generated and instructional explanations

The template for Khan Academy's mathematics exercises ensures that students must generate or simply be told the correct answer by the end of each exercise. Our augmentation of the exercises all occurs after the student receives feedback that they have entered the correct answer – whether they generate it themselves, are helped by hints, or need to go to the very end of the solution to see the answer.

As shown in Figure 1, the typical Khan Academy math exercise (labeled *practice-as-usual*) is augmented using three instructional tactics: (1) Including prompts for students to *self-generate* explanations; (2) Including *instructional* explanations directed at these prompts, ostensibly from another student or teacher; (3) Asking students to *compare* their self-generated explanations to the instructional explanations.

The *self-generate* explanation prompt appears beside a solution step, in a distinctive purple font and accompanied by a text box for students to type their response. The example in Figure 1 has the prompt "Explain what this step means to you:". The *instructional* explanation can be shown in a similar position, such as "Another student explained this as:…". The *compare* judgment solicits a comparison of the student's own explanation with the *instructional* explanation which was supposedly provided by someone else: "How similar is your explanation to the other student's explanation?".

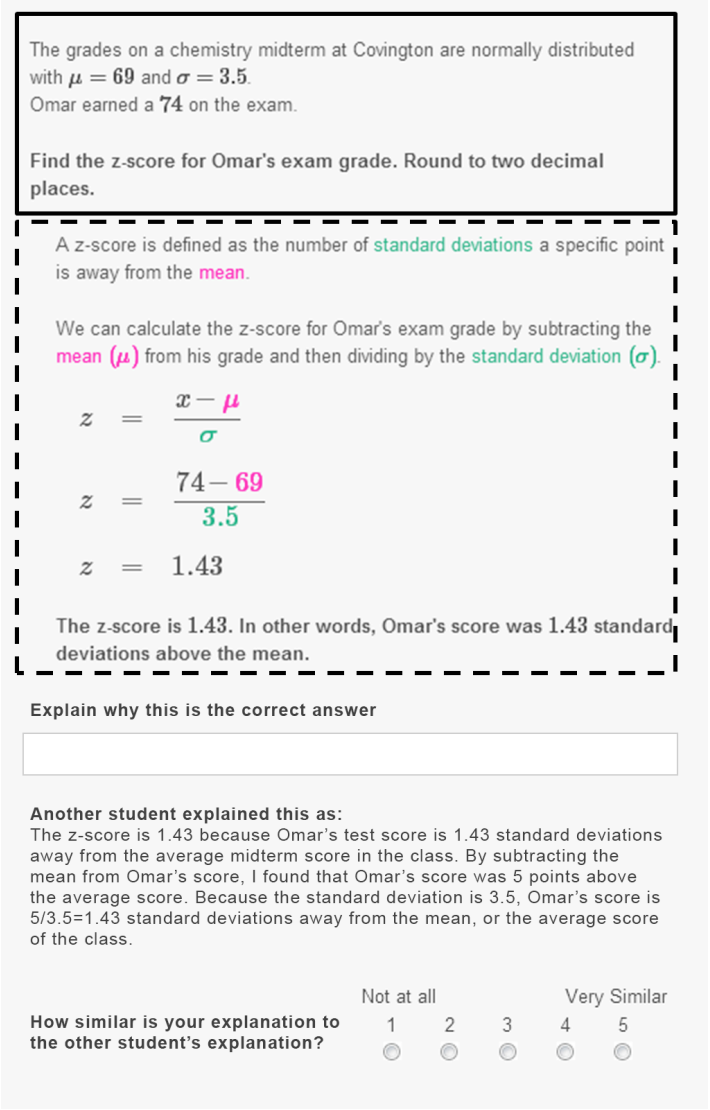


Figure 1: Illustration of worked example solution in typical Khan Academy exercise, and how the problem can be augmented with: (1) a prompt to self-generate an explanation for the correct answer, (2) An instructional explanation, ostensibly from another student, (3) A request for a learner to compare his/her explanation with the instructional explanation. The *practice-as-usual* exercise can be found at https://www.khanacademy.org/math/probability/statistics-inferential/normal_distribution/e/z_scores_1

Experiment

The ongoing study will be conducted using a convenience sample of adults recruited from Amazon Mechanical Turk, as well as undergraduate students. The goal is to investigate this paradigm in a controlled laboratory setting, and then extend it to a realistic educational environment with students in a high school, or introduce it on the actual Khan Academy platform, in an extension of an ongoing collaboration with Khan Academy.

The study independently manipulates whether or not learners are prompted to *self-generate* explanations for the correct answer (once it is obtained), and whether or not they are provided an *instructional* explanation for the correct answer. This results in four conditions:

Practice-as-usual with the typical Khan Academy exercise and no self-generated or instructional explanation.

Self-generated explanation (but no instructional explanation) which includes the prompt to explain why the answer is correct.

Instructional explanation (but no prompt to self-generate an explanation) which provides an explanation that is supposed to come from another student.

Self-generated and *instructional* explanations. This condition is key to evaluating whether learning can be improved through using explanations to provide implicit formative feedback for learners. As described in the next section, several variables are manipulated in this condition to investigate the most effective means of combining self-generated and instructional explanations.

Self-generated and instructional explanations: Order & Comparison

To further investigate the learning benefits of self-generated and instructional explanations, the condition in which participants receive both a self-generated and instructional explanation is made of four nested conditions. These are generated by experimentally manipulating the *order* of self-generated and instructional explanation (self-generated prompt first, then instructional explanation, vs. instructional then self-generated) and whether or not a *comparison* is requested (no comparison prompt, vs. a comparison prompt). The comparison prompt asks learners to rate similarity of self-generated and instructional explanations, such as can be seen in Figure 1: "How similar is your explanation to the other student's explanation?", rated on a scale from 1 (not at all) to 5 (very similar).

It should be noted that the self-generated and instructional explanation are never onscreen at the same time, to avoid simple copying or rote responses. Whichever is presented first simply disappears on the appearance of whichever is presented second.

The design therefore produces four conditions: *Self-Instructional*, *Instructional-Self*, *Self-Instructional-Compare*, *Instructional-Self-Compare*. The manipulations that produce these conditions allow us to investigate whether and when learners receive implicit formative feedback from generating explanations, receiving instructional explanations, and engaging with prompts to compare these explanations.

Summary

The study outlined here aims to investigate whether the proposed combinations of self-generated and instructional explanations have a beneficial impact on learning. The study can shed light on how to design a learning environment to provide implicit formative feedback, by examining how accuracy and speed in exercises is influenced by the relative effects of self-generating explanations, receiving instructional explanations, doing both, and comparing one's self-generated effort with an instructional explanation. More generally, the software adaptation of the Khan Academy exercise framework provides a setting to ask an even broader range of issues: such as changing the type of explanation prompts, features of the instructional explanations, the kinds of comparison prompts used (listing vs. rating, analyzing differences vs. similarities, contrasting explanation quality by identifying pros & cons of each, or by grading or rating different explanations).

References

1. Nicol, D. J., & Macfarlane‐Dick, D. (2006). Formative assessment and self‐regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, 31(2), 199-218.
2. Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. Computers & Education, 57(4), 2333-2351.
3. Shute, V. J. (2008). Focus on formative feedback. Review of educational research, 78(1), 153-189.
4. Fonseca, B. & Chi, M.T.H. 2011. The self-explanation effect: A constructive learning activity. In: Mayer, R. & Alexander, P. (Eds.), *The Handbook of Research on Learning and Instruction* (pp. 270-321). New York, USA: Routledge Press.
5. Williams, J. J., Walker, C. M., Lombrozo, T.: Explaining increases belief revision in the face of (many) anomalies. In: N. Miyake, D. Peebles, & R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 1149-1154). Austin, TX: Cognitive Science Society (2012)
6. Wittwer, J. & Renkl, A. (2008). Why instructional explanations often do not work: a framework for understanding the effectiveness of instructional explanations. *Educational Psychologist, 43*, 49-64.
7. Renkl, A. (2002). Worked-out examples: Instructional explanations support learning by self-explanations. Learning and instruction, 12(5), 529-556.
8. Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. Journal of Educational Psychology, 95(2), 393-408.

An architecture for identifying and using effective learning behavior to help students manage learning

Paul Salvador Inventado*, Roberto Legaspi, Koichi Moriyama, Ken-ichi Fukui and Masayuki Numao
The Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka, Japan, Osaka, 567-0047
{inventado,roberto,koichi,fukui}@ai.sanken.osaka-u.ac.jp, numao@sanken.osaka-u.ac.jp

ABSTRACT

Self-regulated learners are successful because of their ability to select learning strategies, monitor their learning outcomes and adapt them accordingly. However, it is not easy to measure the outcomes of a learning strategy especially while learning. We present an architecture that allows students to gauge the effectiveness of learning behavior after the learning episode by using an interface that helps them recall what transpired during the learning episode more accurately. After an annotation process, the profit sharing algorithm is used for creating learning policies based on students' learning behavior and their evaluations of the learning episode's outcomes. A learning policy contains rules which describe the effectiveness of performing actions in a particular state. Learning policies are utilized for generating feedback that informs students about which actions could be changed or retained so that they can better adapt their behavior in future learning episodes. The algorithms were also tested using previously collected learning behavior data. Results showed that the approaches are capable of building a logical learning policy and utilize the policy for generating appropriate feedback.

Keywords

delayed feedback, self-regulated learning, profit sharing

1. INTRODUCTION

Students often learn on their own when they study for tests, make assignments and perform research as part of their academic requirements. They also learn by themselves when they investigate topics which may not be directly related to class discussions but are interesting to them. When students learn alone, they encounter many challenges related to the learning task, as well as challenges that are meta-cognitive and affect related.

Students who can self-regulate are capable of overcoming these challenges better compared to those who cannot. One reason for this is that self-regulated students know how to select and adapt their learning strategies depending on the current situation. However, this is a complex task because it requires attention and sophisticated reasoning to know which learning strategies to apply, to monitor the outcomes of a learning strategy and to know when a strategy needs to be changed [13].

In this research, we discuss an architecture for helping students manage their learning behavior by helping them become aware of the outcomes of the learning strategies they employed and by helping them identify which strategy is effective in a particular situation.

2. RELATED WORK

Self-regulated learners can be differentiated from less self-regulated learners by looking at the learning behaviors they exhibit. They are characterized by their diligence and resourcefulness, their awareness of the skills they possess, their initiative to seek out information and their perseverance to continue learning and find ways to overcome obstacles [13].

Research such as that of Kinnenbrew, Loretz and Biswas [8] has shown these differences in behavior. In their work, they investigated students' learning behavior while using Betty's Brain, a computer-based learning environment in the science domain that helped students develop learning strategies. They processed log data from student interactions and mapped them to canonical actions. Action sequences were then mined using sequential pattern mining and episode mining to discover learning behaviors. Their results showed that high performing students showed systematic reading behavior and frequent re-reading of relevant information which was not seen in low performing students.

In the work of Sabourin, Shores, Mott and Lester [10], the authors also observed differences in the students' behavior as they interacted with Crystal Island, a game-based learning environment developed for the microbiology domain. While interacting with the environment, students were prompted to report their mood and status. These were later processed and used to categorize the students' goal setting and goal

*also affiliated with: Center for Empathic Human-Computer Interactions, College of Computer Studies, De La Salle University, Manila, Philippines

reflection behavior. They were then given an overall self-regulated learning (SRL) score based on their reports and assigned into low, medium or high SRL category. Students in the high SRL category frequently used in-game resources that provided task-related information and resources that allowed them to record notes. They also spent less time using resources for testing their hypothesis and had higher learning gains.

MetaTutor is a hypermedia learning environment developed for the biology domain that identifies students' SRL processes and also helps them use these processes [2]. Students who used the system indicated the SRL processes they used by selecting it from the list of SRL processes in the system's interface. Pedagogical agents also gave them prompts to use certain SRL processes depending on the current situation (i.e., student information, time on page, time on current sub-goal, number of pages visited relevance of the current page to the sub-goal, etc.) and also gave them feedback regarding how they used these processes. Students who used the version of the system with prompts and feedback were reported to have higher learning efficiencies compared to students who used a version of the system without prompts and feedback.

3. SYSTEM ARCHITECTURE

Learners often have difficulty in selecting, monitoring and adapting learning strategies because of its high cognitive load requirement. This is especially true for complex domains such as science, math, engineering and technology. The approach we take in this work involves helping students understand the outcomes of their learning behavior better by helping them recall what transpired in a recently concluded learning episode. The advantage of recalling is that after the learning episode, students do not need to worry about the learning task and can focus on analyzing their learning behavior. Students will also have a more complete and accurate measurement of their learning behavior's effectiveness because they can observe both short and long term effects on learning. This information will be useful for students in future learning episodes because when they monitor and adapt learning strategies, they can base their decisions on the current context as well as their predictions of what could happen according to their reflections from previous learning episodes.

Asking students to recall a recently concluded learning episode presents two issues. First, students will not be able to completely remember what transpired during the learning episode. We addressed this in our previous work wherein we developed a tool called Sidekick Retrospect, which took screenshots of the students' desktop and video frames from a video of their face during a learning episode [7]. Students who used the software in our experiment reported that they were able to discover things about their behavior that they were previously unaware of. It was also enough to help them reflect on what transpired so that they were able to identify problems with their learning behavior and think of probable solutions. Figure 1 shows a screenshot of the system's interface which are presented to the students after the learning episode. A timeline of the entire learning episode is shown together with desktop and webcam video screenshots relative to the mouse's position in the timeline.

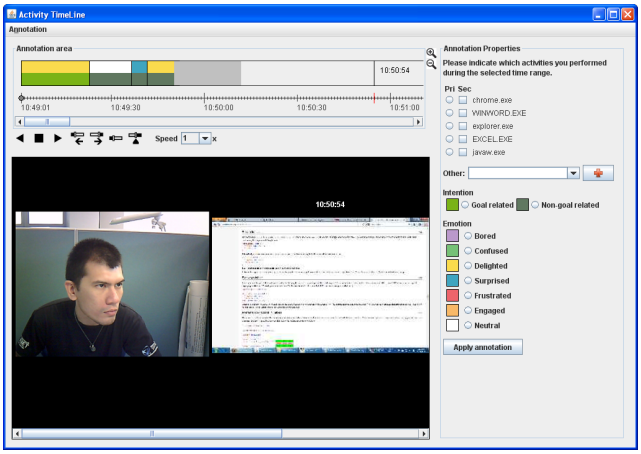


Figure 1: Sidekick Retrospect Annotation Interface

An issue we encountered from our previous work was that students who used the software seemed to focus only on the most significant aspect of the learning episode. They did not reflect as much on other instances during the learning episode even when they employed other learning strategies that also had an impact on their learning. This may have been the case because students were already too tired to spend more time analyzing each event in depth.

The architecture presented in Figure 2, integrates the methodology we used in our previous work with our current approach for helping students recall what transpired during the learning session and helping them discover more insights about their learning behavior. We designed our system so that students would not be bound by a specific environment or domain and keep the learning environment as natural as possible. Students were allowed to learn using any tool or application on or off the computer. However, they had to stay in front of the computer so it could take desktop and webcam video screenshots of their activities and so they could annotate the data after the learning episode. The entire process was split into three phases which are each discussed in the following subsections.

3.1 Interaction Phase

The interaction phase begins by first asking students to input their learning goals for the current learning episode. Data collection starts right after students finish inputting their goals. The system then starts logging the applications used by the students, the title of the current application's window and the corresponding timestamps. Screenshots of the desktop and the webcam's video feed are also taken and stored using the same timestamp as that of the log data.

3.2 Annotation Phase

In the annotation phase, students are asked to annotate their *intentions*, *activities* and *affective states*. Intentions can either be goal related or non-goal related relative to the goals that were set at the start of the learning episode. Activities referred to any activity the student did while learning which could either be done on the computer (e.g., using a browser) or out of the computer (e.g., reading a book). Two sets of affect labels were used for annotating affective states

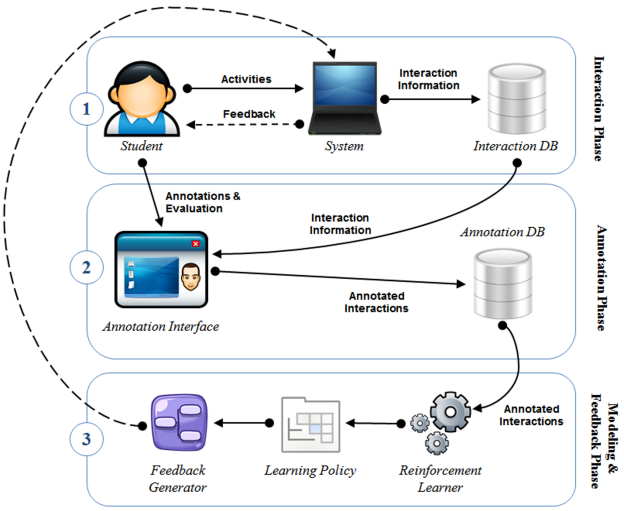


Figure 2: System architecture

wherein goal-related activities were annotated as either delighted (DEL), engaged (ENG), confused (CNF), frustrated (FRS), bored (BRD), surprised (SRP) or neutral (NUT) and non-goal related activities were annotated as either delighted (DEL), sad (SAD), angry (AGY), disgusted (DIS), surprised (SRP), afraid (AFR) or neutral (NUT). Academic emotions [4] are used for annotating goal related intentions because they give more contextual information about the learning activity. However, academic emotions might not capture other emotions outside of the learning context so Ekman's basic emotions [5] were used to annotate non-goal related intentions.

The system's annotation interface helps students recall what transpired during the learning episode by showing desktop and webcam screenshots depending on the position of the mouse on the timeline. The actual annotations can be created by using the mouse to select a time range then clicking on the corresponding intention, activity and affective state buttons. Students are also allowed to input a description of the activity when it was done outside of the computer.

While annotating, students inherently recall what transpired allowing them to identify the appropriate annotation. Going through the entire learning episode sequentially also helps the students annotate more accurately because they can see how and why their activities change. Furthermore, they also see the outcomes of these activities. It is possible that students might not annotate the data correctly for fear of judgment or lower scores. However, reassuring them that the results will not be used as part of their grades or telling them that accurately annotating their data will help them become more self-regulated and effective could help minimize these cases.

After the annotation process, students are asked to give a learning effectiveness rating between one to five, indicating how good they felt the learning episode was. This rating is likely to be accurate because of the level of detail in which students reviewed their learning episode.

3.3 Modeling and Feedback Phase

In the modeling and feedback phase, students' data are analyzed to create and update the student's list of effective learning behavior or policy. Students' behavior in the current learning episode can be compared to the policy to identify effective and ineffective behavior that can be adapted in succeeding episodes.

3.3.1 Learning policy creation

Self-regulation can be viewed as cyclic phases of forethought, performance and self-reflection [14] wherein reflections about the outcomes of behavior after a learning episode can be used to increase the effectiveness of future learning episodes (e.g., discarding or modifying ineffective behavior). The ideal effect of would be for learning outcomes to continually improve over time.

We fit this incremental perspective of adapting behavior into a reinforcement learning (RL) problem in machine learning which searches for the best actions to take in an environment (i.e., learning behavior) to maximize a cumulative reward (i.e., learning effectiveness) [11].

Profit sharing is a model-free RL approach that is capable of converging even in domains that do not satisfy the Markovian property [1]. We decided to use this approach primarily because we deal with human behavior in a non-deterministic and uncontrolled environment. Profit sharing's reinforcement mechanism allows it to learn effective, yet sometimes non-optimal, policies quickly compared to other algorithms. This is ideal for our situation because we need to give policy-based feedback using minimal data.

Profit sharing differs from other RL techniques because it reinforces effective rules instead of estimating values from succeeding sequential states. A rule consists of a state-action pair (O_t,A_t) which means performing A_t when O_t is observed. We consider these rules as learning behaviors. An episode n is a finite sequence of rules wherein the entire sequence is awarded the reward R based on its outcome. After each episode, the weights of each rule in the sequence is updated using (1) where function $f(R,t)$ is a credit assignment with t being the rule's distance from the goal. Note that it is possible for a rule's weight to be updated more than once if it appears more than once in a sequence. The set of all rules and their corresponding weights is called a policy. A policy is rational or guaranteed to converge to a solution when the credit assignment fuction satisfies the rationality theorem (2) with L being the number of possible actions. In our work, we used a modified version of the rational credit assignment function (3), which was adapted from [1] so that the rules' weights will be bound by the reward value.

$$W_{n+1}(O_t, A_t) \leftarrow W_n(O_t, A_t) + f(R, T) \tag{1}$$

$$\forall t = 1, 2, 3..., T. \quad L\sum_{j=0}^{t} f(R, j) < f(R, t) \tag{2}$$

$$f_{n+1}(R, t) = (R - W_n(O_t, A_t))(0.3)^{T-t} \tag{3}$$

According to Winne's [12] SRL model, students adapt their

strategies based on the results of metacognitive monitoring and evaluation. When the outcome of a task satisfies a student's expectations, then they may continue performing the current task or proceed to the next task. On the other hand, when a task does not achieve its expected outcomes, students can adapt their strategies accordingly. Unfortunately, we did not have access to students' metacognitive evaluations in our data. However, Carver and Scheier's [3] model theorized that the results of metacognitive evaluations can be observed in students' emotion. When the outcome of a task is according to a student's expectation, then neutral affect is experienced. However, when the outcome does not satisfy expectations then negative affect is experienced. On the other hand, when the outcome exceeds expectations then positive affect is experienced. Based on these assumptions, we represented our states using the triple <activity, affect, duration>. Apart from affect which approximated students' metacognitive evaluation, we included activity to indicate the task performed by the student and duration to indicate how long it was performed by the student.

The data showed that students performed similar activities but used different applications (e.g., browsing websites with Google Chrome vs. Mozilla Firefox). Instead of treating these separately, we categorized the students' activities into six types: information search [IS] (e.g., using a search engine), view information source [IV] (e.g., reading a book, viewing a website), write notes [WN], seek help from peers [HS] (e.g., talking to a friend), knowledge application [KA] (e.g., paper writing, presentation creation, data processing) and off-task [OT] (e.g., playing a game).

Durations were even more varied ranging from one second (e.g., clicking a link from a search results page) to 53 minutes (e.g., watching a video). Using this directly will result in a large state space so we categorized them into short, medium or long duration. The duration values were positively skewed so evenly partitioning the data according to the number of elements or frequency would cause both short and medium groups to have small and similar values. The long duration group on the other hand, would have values with high variation. We decided to use k-Means to categorize the duration values into three clusters (i.e., $k = 3$) and using a Euclidean distance formula as described in [6]. Clustering produced groups with elements having similar duration values and whose values were different from the other groups. Elements in the cluster with the smallest values were labeled short duration, elements in the cluster with the biggest values were labeled long duration and the elements in the remaining cluster were labeled with medium duration. The centroids identified by k-means for short, medium and long durations were 69.4 seconds (1.15 minutes), 614.5 (10.2 minutes) seconds and 1999.4 seconds (33.3 minutes) respectively. 90.83% of the duration values were short, 8.17% were medium and 0.10% were long.

In the learning context, actions would refer to changing from one activity to the other. So, we used the same eight activity categories as actions. However, we added a change information source [CS] action to handle cases when students would either view a different website or change to or from a physical information source (e.g., book, printed conference paper).

In this representation, there would be no consecutive rules with states having the same values unless they were paired with different actions. Otherwise, these rules were merged and their durations added. An example of a rule would have the form (<IV, CNF, short>, CS).

The student's rating of the learning episode's effectiveness can directly be used as the reward value. Data from learning episodes can then be converted into rule sequences and be used to update each rule's weight incrementally using (1) with the corresponding reward values. The rules' weights are expected to converge to the reward value it is commonly associated with.

3.3.2 *Learning policy-based feedback*

According to Pressley, Levin and Ghatala [9], adult students who were given information regarding the utility of two learning strategies and a chance to practice them were capable of validating its outcomes and were reported to use the more effective strategy. In our case, the utility of performing an action in a certain state is its weight value (i.e., applying the rule will likely lead to a learning effectiveness rating that is at least the weight value). Information about the utility of two or more competing rules (i.e., rules referring to the same state but with different actions) can be used to give students feedback at the end of a learning episode so they can verify and adapt them accordingly in succeeding episodes. When students used more effective rules, it is assumed to result in better learning effectiveness ratings which will reinforce the rule in the learning policy.

As more rules are observed and added into the learning policy, some rules may not be relevant to a particular learning episode. The rules with their corresponding utilities should first be filtered before they are presented to the student. In the first learning episode, the learning policy will still be empty so feedback will be unavailable. When a policy already contains rules, each rule employed in the current learning episode can be compared to the rules in the learning policy and provide relevant feedback. The pseudo code presented below describes how three types of feedback can be given to the student. First, when students perform an action with a worse utility based on the policy, the system can remind the student to select the better action. Second, if the student performs an action which isn't in the policy but has lower utility than the best action in the policy, the student is told that the action may be ineffective. Lastly, if the student performs an action which isn't in the policy but has a higher utility than the best action, the student is informed that a better action has been found compared to the previous best action. Whenever a student performs the best action according to the policy, feedback is no longer given because it is assumed that the student already knows this and is the reason why the action was selected. In cases when the student performs an action in an unknown state, feedback cannot be given as well because of insufficient information.

Initialize set of weighted rules X
Copy old policy P into P'
For each (O_t, A_t) in the current learning episode
    Update $W(O_t, A_t)$ in P' using (1)

```
    For each (O_p, A_p) in policy P
        If O_t = O_{p,i}
            Add W(O_{p,i}, A_{p,i}) into X
        End
    End
End
For each (O_t, A_t) in the current learning episode
    If (O_t, A_t) not in X
        Unknown utility
    Else if (O_t, A_t) not in P
        If W(O_t, A_t) < max(W(O_{p,i'}, A_{p,i'})) in X
            Inform student that A_{p,i'} > A_t
        Else
            Inform student that A_t > A_{p,i'}
        End
    Else
        If A_t <> A_{p,i'} where max(W(O_{p,i'}, A_{p,i'})) in X
            Inform student that A_{p,i'} > A_t
        End
    End
End
```

A cause for concern is that the learning policy might not have converged yet resulting in incorrect feedback (e.g., telling the student to perform an action which is actually ineffective). Again according to Pressley *et. al.* [9], despite being given incorrect utility information adults are able to select better strategies after practice wherein they are able to observe the strategy's actual utility. As students constantly select effective actions (i.e., as a result of their own evaluation), the policy will be updated to reinforce these actions and decrease the chance of providing incorrect feedback. This emphasizes the need for students in this environment to explore other actions so that they can find the best actions which will also be reflected in the policy. It also then becomes necessary for other mechanisms to encourage exploration such as looking at other students' learning policies for possible actions or using expert knowledge.

4. LEARNING BEHAVIOR DATA

The methodology described in the interaction and annotation phases of the architecture was used in collecting the data in our previous work [7]. The data was collected from four students aged between 17 and 30 years old, conducting research as part of their academic requirements. Three of the students were taking Information Science while one student was taking Physics. During the data collection period, two of the students were writing conference papers and two made power point presentations about their research. They all processed and performed experiments on their collected data, searched for related literature and created a report or document. Although their topics were different, they performed similar types of activities. Two hours of annotated learning behavior data in five separate learning episodes were collected from each student over a one week period. The annotation data was processed using the method described in Section 3.3.1 resulting in five separate learning episodes for every student and each episode consisting of the sequenced rules. On average, students used 54.35 rules per session (N=20; σ=27.71) including repeated rules.

Table 1: Rule Categories

#	Type	State	Action	Reward
1	PRL	ENG, IV, short	KA	0.360000
2	PRL	ENG, IV, short	CS	0.004154
3	CDH	CON, IV, short	CS	0.441939
4	CDH	CON, IV, short	KA	2.34E-05
5	CDH	CON, IV, short	OT	9.16E-15
6	RLX	ENG, KA, long,	OT	1.830000
7	RLX	ENG, KA, long,	HS	0.009720
8	RLX	ENG, KA, long,	IV	2.13E-06
9	RSL	DEL, OT, short	KA	0.389484
10	RSL	DEL, OT, short	IV	2.00E-18
11	RSL	DEL, OT, short	HS	9.57E-26

5. RESULTS AND ANALYSIS

The learning policies generated by the profit sharing algorithm on the learning behavior data consisted of rules based on the state and action representation used. There were many rules due to our selected state-action space, but we observed four categories after analyzing the data– Prolonged learning (PRL), Cognitive disequilibrium handling (CDH), Relaxation (RLX) and Resumed learning (RSL). Table 1 presents examples of each category which were taken from the learning policy of the doctoral physics student who was experimenting with her data and used its results for writing a conference paper.

PRL rules refer to states wherein students feel engaged while performing a learning-related activity and switch to another learning-related activity. It describes how long a certain type of activity could be effective and what other activities may complement it. Taking the physics student's data as an example, let us consider that she was looking into different concepts for data manipulation because she needed it for writing her conference paper. According to rules 1 and 2, it was better for her to try and run an experiment on her data (i.e., apply knowledge), before shifting to a different concept (i.e., view information source). This would allow her to have a better understanding of the concept and allow her to write the paper more easily.

CDH rules refer to states wherein students adapt their behavior to handle negative affect (e.g., confusion or boredom) while learning. These give an idea how long to stay in a confusing or bored learning state before shifting to an activity that will probably alleviate the problem. For example, rule 3 indicates that it is probably better to find a different information source if it is confusing instead of spending a lot time trying to understand it. Rule 5 also indicates that it is not a good idea to just engage in off-task activities when it is difficult to understand a certain information source.

RLX rules refer to states wherein students relax or shift to off-task activities after learning. According to rule 6, it was effective for the student to relax after spending a long time learning. This supports claims that off-task activities or relaxation are important for continued learning [7].

RSL rules refer to cases wherein students shift back to learning from an off-task activity. It seemed that the utility for performing actions in this category are context-dependent.

Table 2: Rule correctness over learning episodes

Ep	+	-	New^+	New^-	Unknown	Reward
2	0	0	1	0	3	4
3	1	0	2	1	1	3
4	12	0	5	0	1	4
5	4	51	0	1	6	2

For example, according to rule 9, it was more effective to apply knowledge probably because the goal was to write a conference paper. Spending too much time reading information sources would help, but not directly lead to the achievement of the goal. This effect is important to consider because if students change their goals, the policy may not be directly applicable to the new goal. A separate experiment needs to be conducted to observe how the architecture will handle such scenarios. We think however that the speed in which the algorithm adjusts the learning policy is a good factor that can make it capable of handling such changes.

After a student completes a learning episode, an updated learning policy can now be used to generate feedback. The feedback will be based on five cases: the student chooses the best action according to the policy (+), the student does not choose the best action according to the policy (-), the student tries a new action which has better results than the best action in the policy (New^+), the student tries a new action which has worse results than the best action in the policy (New^-) and the student performs the only action associated to a state in the policy or the student performs an action in an unknown state for the first time such that the policy will not be able to identify if there is a better action (Unknown).

We simulated how feedback would be generated for these five cases by testing the algorithm on data from the same student. The student's actions in the first learning episode were used to build an initial policy. No feedback was generated at this point because learning policy would only contain rules based on the current episode. Feedback for the second episode could now be generated because it can be compared with the learning policy created using data from the first learning episode. The learning policy was updated using data from the second episode, and was used to generate feedback for the third learning episode. This was repeated for all remaining learning episodes. Table 2 presents the number of times each case is encountered as new learning episodes are experienced by the student.

The table shows that the student implemented a few rules in episode two which was caused by the student spending a long time performing an activity. We see that her learning policy was updated with three new rules as well as a new effective action (i.e., performing an off-task activity after spending a long time experimenting with data). The high reward value indicates that the student did well because all actions, including those unknown actions, were effective. This was confirmed by checking her updated learning policy generated in the fifth episode. The unknown actions were in fact the best actions in their corresponding states (i.e., performing an off-task activity after spending some time experimenting with data, resume data experimentation after a short off-task activity and consulting a friend about the experiment after a short off-task activity). The student also performed few actions in the third episode but gave it a smaller reward value probably because she spent too much time talking to a friend even though the other actions were effective (i.e., resuming data experimentation after a short off-task activity and viewing a paper after some time experimenting with data). In the fourth episode, the student constantly performed effective actions and even discovered a new action which probably caused the increase in reward. Finally in the fifth episode, the student performed a lot of ineffective actions which probably caused the big decrease in the reward value. Specifically, as we have discussed earlier, she spent short amounts of time repeatedly viewing different information sources. The policy indicated that it would have been better for her to apply knowledge, which in her context would mean either writing the paper or experimenting with her data. This could in fact be an effective strategy because she could verify and learn more about the concept by applying it rather than moving on to another concept right away.

Our results also showed that there was a relationship between the number of times students correctly followed rules in their learning policy and their learning effectiveness rating. Figure 3 presents graphs corresponding to each student showing this relationship. The learning effectiveness ratings were expressed as ratios relative to the highest rating (i.e., five) and the number of correct actions were expressed as ratios relative to the total number of actions in the learning episode. The trend indicates that the learning policy was able to identify effective actions from the students' behavior such that when the students selected more effective actions (i.e., based on the learning policy), they also had a more effective learning episode. This means that if the student will be able to follow the feedback provided by the system in succeeding learning episodes, it is likely for them to have more effective learning experiences.

6. CONCLUSION AND FUTURE WORK

We have presented an architecture for collecting students' learning behavior data, uncovering effective learning behaviors and using them to help students manage their learning. The approach does not require a specific learning environment so the student's behavior is naturalistic and captures how he/she actually learns. However, it does require students to annotate their data. Annotation is done after learning so it does not require additional cognitive load during the learning episode. Desktop and web cam screenshots can help students recall the context in which they learned and can likely improve annotation accuracy.

The profit sharing algorithm was used for building learning policies that contained rules describing an action's effectiveness in a particular state. Learning policies generated from previous learning episodes can be compared with data from the current learning episode to identify which actions were effective or ineffective and generate feedback accordingly. Feedback about possible improvements can be useful for students to adapt their actions in future learning episodes.

Simulations from actual data showed that updating the learning policy also changed the resulting feedback such that

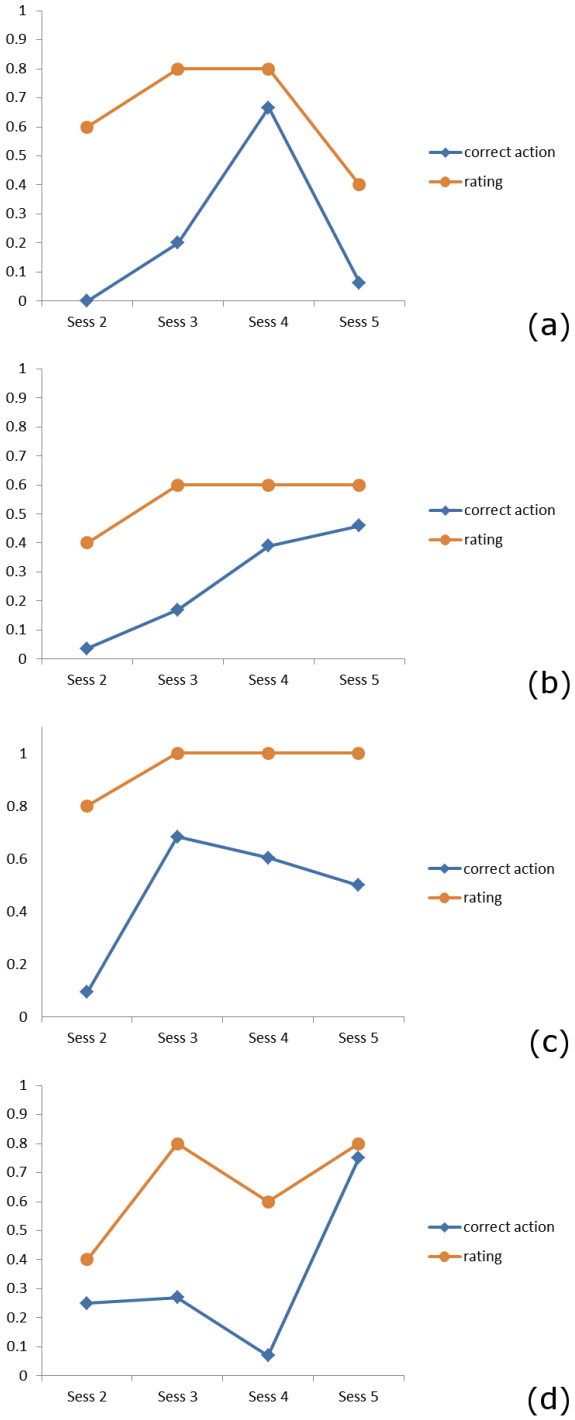


Figure 3: Relationship between action correctness and student rating

newer, more effective actions were presented to the student. This helps ensure that the student will always be prompted to select the most effective learning behavior. The relationship between the number of effective rules followed by the student and their learning effectiveness ratings indicate that the learning policy-based feedback will have a good chance of helping students learn more effectively.

The architecture we have designed still has some issues that need to be addressed. Our state representation did not contain information regarding students' metacognitive evaluations. Although we used emotions to approximate these evaluations, asking students to annotate them will be more accurate and create better policies. The reward values we used were based on students' self-evaluations and it would be interesting to see the difference when using learning gains instead (e.g., asking students to take a pretest and posttest). Combining both learning gains and self-evaluation to create the reward value may be a better measurement because it will consider both the student's preferred learning behavior and knowledge gained.

Our architecture also faces a common problem in RL called the exploration-exploitation problem. In order for the policy to be optimal, students need to try as much actions as possible. Due to the approach's reliance on the student's learning behavior, it cannot suggest actions outside of the current learning policy. This would require mechanisms for suggesting actions not in the learning policy such as using other students' learning policies or using expert knowledge.

Even though the approach can create policies that span across learning episodes, it has only been tested with learning episodes having the same goal. In the case of our data, students were either writing a conference paper or creating a power point presentation. It will be more useful if it could also be used across different learning goals. The current approach needs to be tested to see how well it fares in such a case and necessary modifications need to be applied accordingly.

The data we used was collected from adult learners and may be effective for them. However, according to Pressley *et. al.* [9], children have difficulty in verifying learning strategy utility even after practice. It is possible that additional feedback may be needed to fit this approach to younger learners.

Acknowledgements

This work was supported in part by the Management Expenses Grants for National Universities Corporations from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) and JSPS KAKENHI Grant Number 23300059. We would also like to thank all the students who participated in our data collection.

7. REFERENCES

[1] S. Arai and K. Sycara. Effective learning approach for planning and scheduling in Multi-Agent domain. In *6th International Conference on Simulation of Adaptive Behavior*, pages 507–516, 2000.

[2] R. Azevedo, R. S. Landis, R. Feyzi-Behnagh, M. Duffy, G. Trevors, J. M. Harley, F. Bouchet, J. Burlison, M. Taub, N. Pacampara, M. Yeasin, A. K.

M. M. Rahman, M. I. Tanveer, and G. Hossain. The effectiveness of pedagogical agents' prompting and feedback in facilitating co-adapted learning with MetaTutor. In *Intelligent Tutoring Systems*, pages 212–221, 2012.

[3] C. S. Carver and M. F. Scheier. Origins and functions of positive and negative affect: A control-process view. *Psychological Review*, 97(1):19–35, 1990.

[4] S. D. Craig, A. C. Graesser, J. Sullins, and B. Gholson. Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*, 29(3):241–250, 2004.

[5] P. Ekman. Are there basic emotions? *Psychological Review*, 99(3):550–553, 1992.

[6] G. Gan, C. Ma, and J. Wu. *Data clustering: theory, algorithms, and applications*, volume 20. Society for Industrial and Applied Mathematics, 2007.

[7] P. S. Inventado, R. Legaspi, R. Cabredo, and M. Numao. Student learning behavior in an unsupervised learning environment. In *20th International Conference on Computers in Education*, pages 730–737, 2012.

[8] J. S. Kinnebrew, K. M. Loretz, and G. Biswas. A contextualized, differential sequence mining method to derive students' learning behavior patterns. *Journal of Educational Data Mining*, in press.

[9] M. Pressley, J. R. Levin, and E. S. Ghatala. Memory strategy monitoring in adults and children. *Journal of Verbal Learning and Verbal Behavior*, 23(2):270–288, 1984.

[10] J. Sabourin, L. R. Shores, B. W. Mott, and J. C. Lester. Predicting student self-regulation strategies in game-based learning environments. In *Intelligent Tutoring Systems*, pages 141–150, 2012.

[11] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*. A Bradford Book, 1998.

[12] P. H. Winne. Self-regulated learning viewed from models of information processing. *Self-regulated learning and academic achievement: Theoretical perspectives*, 2:153–189, 2001.

[13] B. J. Zimmerman. Self-regulated learning and academic achievement: An overview. *Educational psychologist*, 25(1):3–17, 1990.

[14] B. J. Zimmerman. Becoming a Self-Regulated learner: An overview. *Theory Into Practice*, 41(2):64–70, 2002.