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

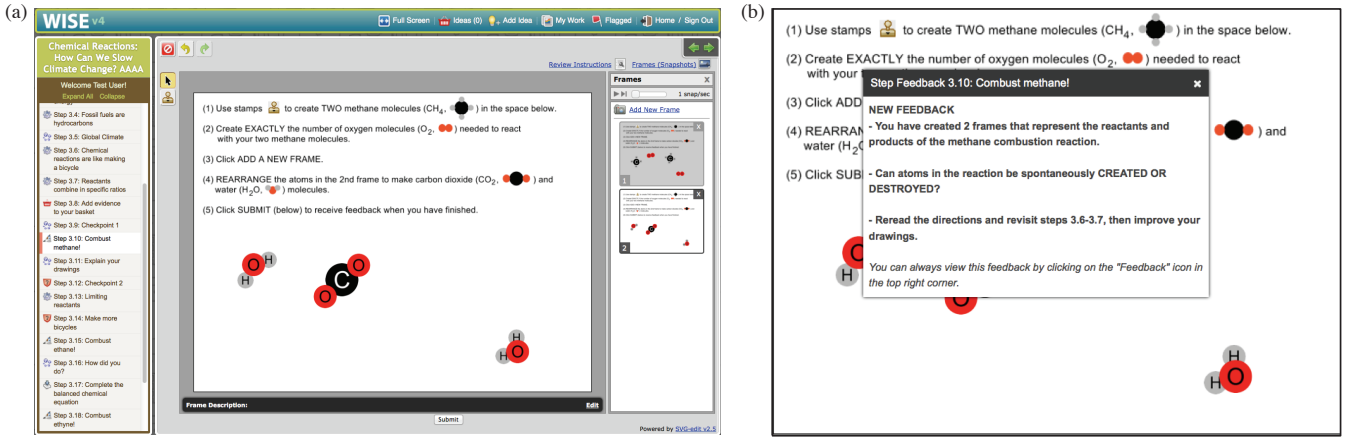

Figure 1: The WISE drawing environment. (a) A screenshot of a student drawing. Students place atom stamps on the central canvas to show the molecules at the beginning and end of a chemical reaction. On the right side of the screen, the two frames that the student has created are shown. (b) The student 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.

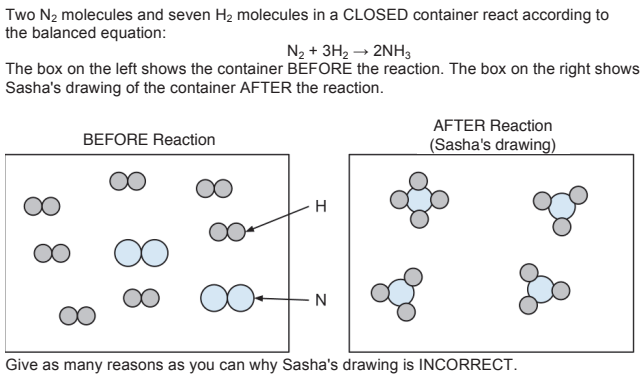

Figure 2: Item from the pre- and post-test related to drawing chemical reactions. Students are asked to examine Sasha's drawing and explain why the drawing is incorrect. The drawing task is similar to those in the unit, but asks students to evaluate rather than generate the drawing and requires integrating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

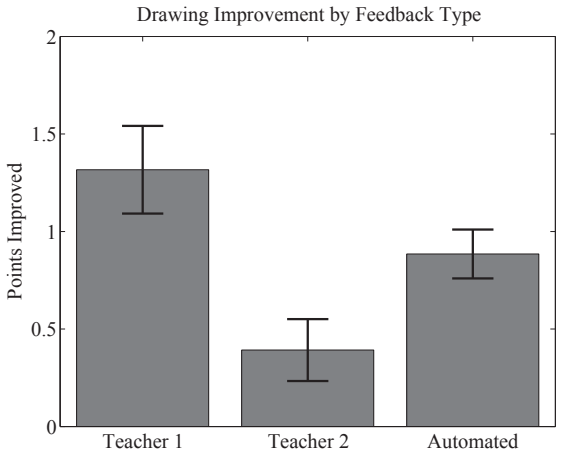

Figure 3: Improvement on drawing scores based on type of feedback received. Error bars indicate one standard error.

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.