120th ASEE Annual Conference & Exposition
FRANKLY, WE DO GIVE A D*MN
June 23-26, 2013

Paper ID #7131

Automated Text Analysis Facilitates Using Written Formative Assessments for Just-in-Time Teaching in Large Enrollment Courses

Dr. Luanna B Prevost, Michigan State University

Dr. Prevost is a postdoctoral research associate with the Center of Engineering Education at Michigan State University. Her research interests are in student writing, problem solving, and technologies that can be used to assess and teach these skills.

Dr. Kevin C Haudek, Michigan State University
Emily Norton Henry, Michigan State University
Mr. Matthew C Berry, Michigan State University
Dr. Mark Urban-Lurain, Michigan State University

Associate Professor Center for Engineering Education Research Undergraduate Studies Office College of Engineering Michigan State University

Dr. Urban-Lurain is responsible for teaching, research and curriculum development, with emphasis on engineering education and, more broadly, STEM education.

His research interests are in theories of cognition, how these theories inform the design of instruction, how we might best design instructional technology within those frameworks, and how the research and development of instructional technologies can inform our theories of cognition. He is also interested in preparing future STEM faculty for teaching, incorporating instructional technology as part of instructional design, and STEM education improvement and reform.

Automated Text Analysis Facilitates Using Written Formative Assessments for Just-in-Time Teaching in Large Enrollment Courses

Abstract

Written formative assessments can provide instructors with rich insight into students' thinking about scientific concepts. However, the time and effort involved in grading deter instructors from having students write in large courses. As large-enrollment introductory STEM courses become increasingly common, the need for innovations that facilitate the use of written assessments continues to grow. We piloted the use of automated text analysis to overcome these obstacles and facilitate the use of written formative assessment in a large-enrollment introductory biology course. Student responses to online homework on thermodynamics, metabolism, central dogma (genetics) and acid-base chemistry were collected in three 300+-person course sections. We used automated text analysis to extract and categorize concepts from student writing. Then, we used k-means cluster analysis to aggregate responses into distinct groups. From these analyses, we created feedback reports to provide instructors with an assessment of students' responses before the next class period (less than one working day), so that instructors could use this feedback to inform their instruction. We present the results of this pilot study, including a description of the feedback reports and faculty instruction in response to feedback on student writing. We also describe lessons learned to improve the use of written assessments, automated analysis, rapid feedback reports and instruction in large enrollment courses. Finally, we suggest some future directions for research based on our analysis of student writing.

Introduction

Effective assessments allow instructors to observe how learners represents their knowledge in a subject domain [1] . Formative assessment, in particular, is important for modifying instruction to improve student learning [2]. Constructed response assessments, such as writing tasks, allow students to represent their understanding in their own words, and can give faculty greater insight into student thinking compared to multiple choice assessments [3]. In addition, multiple choice assessments also introduce a significant validity threat as they introduce what may be referred to as the "either-or" forced-choice ("misconception" *vs.* scientific key concept) item preference. Indeed, the use of constructed response instruments—coupled with oral interviews—clearly reveals "mixed models" or "synthetic models" of student thinking, in which students present both scientific ideas and misconceptions, in regard to natural selection (see ref 4 for a cross-cultural example) and student thinking about acid/base chemistry [5]. Thus, assessment methods that capture holistic snapshots of students' explanatory models — such as constructed response items — are necessary for mitigating these assessment constraints and revealing these "mixed models".

However, the time and costs for grading constructed response items pose a barrier to their use, especially in large enrollment courses. Therefore innovative assessments and tools to evaluate them are needed to overcome these barriers and to encourage faculty to have students write in large enrollment STEM courses.

To address this challenge, the Automated Analysis of Constructed Response (AACR) research group has been exploring the use of feature-based lexical analysis [6] to analyze student writing about science concepts. Our research team has built a suite of constructed response (CR) questions to explore student writing about thermodynamics, metabolism, acid-base chemistry and genetics. We developed these questions using items from concept inventories (such as the Genetics Concept Assessment, ref 7), diagnostic question clusters[8, 9], and questions which have been used through several iterations in biology class[5, 10, 11]. For each question we have developed resources and statistical models to analyze students' written responses.

In this paper we present the next phase in our research: a pilot study of the development of a Just-in-Time Teaching (JiTT, ref 12) automated analysis and reporting system which allows faculty to receive rapid feedback on students' writing to use in class the next day. We also describe some of the lessons learned and future directions for building on this model.

Pilot study design and methods

Course demographics and format

This study was carried out over the fall semester of 2012 in three sections of an introductory biology course at a large public Midwestern university. This was a large-enrollment course with sections of 309, 302, and 466 students. One section of the course was taught by two instructors; the other two sections each had one instructor. Students were mostly sophomores, but all years were represented in the classrooms (Table 1). At the beginning of the semester the mean cumulative GPA of students in each section ranged from 2.48-2.73, with one section having a significantly higher GPA than the other sections (Table 1, Kruskal-Wallis Test p=0.045). Final course grades were not significantly different across sections (Kruskal- Wallis Test p=0.748). GPAs and final course grades were not normally distributed so the nonparametric Kruskal-Wallis Test was used to compare means. Students enrolled in the course had completed, or were concurrently enrolled in, the prerequisite introductory general chemistry course.

Two of the course sections met twice a week for 1 hour and 20 minutes. The third section met three times a week for 50 minutes. All three sections had similar classroom formats. They used the same textbook (Campbell Biology 9th Edition, Pearson Benjamin Cummings, Inc) along with its online support. During class meetings, instructors used a mixture of traditional lecture interspersed with one or more of the following: 5-10 minute breakout discussions, clicker activities, and question and answer sessions. All three sections also used the university's online learning management system for sharing course content, regular multiple-choice homework assignments, and online class discussion board.

Table 1. Demographic data for introductory biology sections in which JiTT pilot study was conducted.

Section	1	2	3
Enrollment	309	466	302
Year in School (percent)			
First year	16.5	16.8	22.5
Sophomore	53.7	50.2	44.7
Junior	21.4	24.2	24.5
Senior	7.8	8.6	7.9
Gender			
Female	46	58.2	48.7
Male	54	41.8	51.3
Academic Performance			
Cum GPA start term*	2.48±1.3	2.69±1.21	2.52±1.26
Final grade in course#	2.33±1.1	2.28±1.07	2.30±1.14

*Kruskal Wallis Test p=0.045

#Kruskal Wallis Test p= 0.748

CR Questions and data collection

Overall, we collected and analyzed over 12,000 responses from students in a single semester (Table 2). Typical responses varied in length from a single sentence to a short paragraph. During the semester we administered fifteen different homework questions in three different sections of introductory biology. These fifteen questions mainly dealt with four subject areas: thermodynamics, metabolism, the central dogma of molecular biology (the flow of genetic information), and acid-base chemistry. The questions came in two forms: 1) a constructed response question (Table 2: thermodynamics and metabolism sample questions) or 2) a multiple choice option followed by a constructed response explanation (Table 2: acid-base chemistry sample question; see also ref 8). For some questions, there were multiple versions of the same question differing only in the surface features (i.e. organism or cell type) of the question, but not on the underlying concept being assessed. Instructors of the three sections were allowed to choose which questions were used in their individual sections and when they were administered. These questions were intended to be asked pre-instruction during online homework assignments, so that the responses could be analyzed and a report returned to the instructors to allow them to modify instruction during the next class period. However, some questions were also asked post-instruction, which allowed the opportunity to see if students had changed explanations due to instruction. The actual number of responses to any individual question depended on the number of sections using that question, student participation rate in completing the homework, and whether there were multiple versions of that question. All responses were collected online using the university's learning management systems and students received incentives for participation. More detail on the incentives given and their relative effectiveness is discussed below in the *Improving the Use of JiTT* section.

Table 2. Summary of responses to CR questions collected in the study.

Topic	No of unique questions	No of pre-responses	No of post-responses	Total responses	Sample question
Thermo-dynamics	3	2023	446	2469	A carbohydrate is composed of a string of covalently linked monosaccharides. Breaking those bonds between the monosaccharides is a chemically spontaneous reaction (ΔG for this reaction is -3.7 kcal/mol). However, this reaction occurs very slowly at room temperature. Why do you think this is so?
Metabolism	6	3351	927	4278	Each Spring, farmers plant about 5-10 kg of dry seed corn per acre for commercial corn production. By the Fall, this same acre of corn will yield approximately 4-5 metric tons of dry harvested corn. Explain this huge increase in biomass: where did the biomass come from and by what process?
Genetics	5	2484	3014	5498	There is a G to A base change at the position marked with an asterisk. Consequently, a codon normally encoding an amino acid becomes a stop codon. How will this alteration influence DNA replication? Note: the figure accompanying this question is not included here.
Acid-base chemistry	1	432	N/A	432	Consider two small, identical, organic molecules in the cytoplasm of a cell, one with a hydroxyl group (-OH) and the other with an amino group (-NH2). Which of these small molecules (A. hydroxyl, B. amino, C. both) is most likely to have an impact on the cytoplasmic pH? Explain your answer for the above question
Total	15	8290	4387	12677	

Lexical and Statistical Analysis

We performed a two-step analysis of students' written responses using IBM SPSS Modeler software version 14.2 [13]. First, we used automated text analysis to extract and categorize concepts from student writing. We used libraries that we previously had created for these topics; libraries are dictionaries of words and phrases that are relevant to the question and subject matter and that are recognized by text analysis software. Words and phrases which represent homogenous ideas are grouped into categories. For example, the category *break* in Figure 1 contains the terms *break*, *broken*, *broken down* and *break apart*. Categories are revised by an expert in the subject matter to ensure that only relevant terms are included. Once categories have been finalized, each response can be classified into one or more categories based on the words and phrases used in that response.

In our second step of analysis, we used k-means cluster analysis to group the categorized responses. Each response is classified into the cluster for which it is closest to that cluster mean, or centroid. This means that responses in a cluster were more similar to each other than to responses in other clusters. *K*-means cluster analysis allows the recombination of cases and *k* user-defined clusters over repeated iterations. Recombinations are iterated until no further change occurs. Clusters were formed based on the frequency and association of categories in and among responses. The cluster analysis was iterated for values of k=2 to 5 and clusters were homogeneous in the types of the responses they contained. A content expert in the field examined the predicted clusters to ensure that they were conceptually meaningful.

Rapid feedback reports

We used the results of lexical and cluster analyses to generate rapid feedback reports for faculty to use for JiTT. Typically, data collection on the online management system closed at midnight. Analysis and report preparation began the following morning and were completed and emailed to faculty that afternoon for use during the next class period (usually one to three days away).

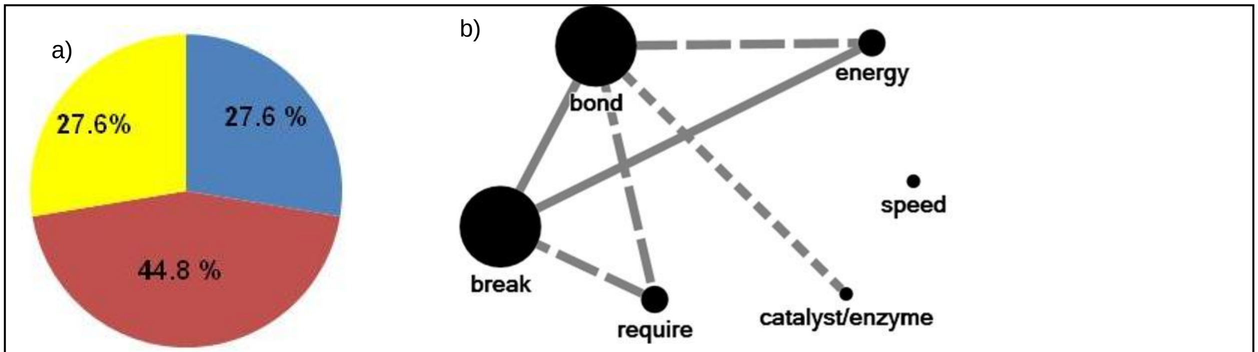
A sample report is presented in Figure 1. Reports included the distribution of clusters using a pie chart (1a), the question asked (1c), the category frequencies within each cluster (1d), cluster descriptions (1e), sample student responses that were closest to their cluster centroids (1f), and in some cases a pre-post comparison (1g) and a web diagram (1b). Reports also included category definitions (not shown in Figure 1). In most cases, the most representative k-means result grouped categorized responses into 3-5 distinct clusters. The most important categories in the predictive model (as indicated by cluster analysis results) were included in the report, along with their means (i.e., the frequency of that category in each cluster). This information was color coded for easy reference by instructors. Finally, an expert in the field (graduate student or postdoc) then reviewed the categories and responses in each cluster and provided a brief, general description of each cluster.

Figure 1 (a - f) represents the results of the lexical and cluster analysis for 173 responses to a question about reaction thermodynamics given to students post-instruction. Cluster 1 contained responses that indicated energy was required to break bonds of the carbohydrate molecule. This one idea was expressed in 27.6% of the students' responses. Cluster 3 contained responses (27.6%) that indicated an enzyme was needed to speed up the rate of the reaction at room temperature, but few of these responses explained the mechanism by which the enzyme would do so. Cluster 2 (44.8%) was the most heterogeneous cluster. Cluster 2 grouped together responses which expressed a variety of incorrect ideas; each of these ideas were only present in a few responses. For example, fewer than 10% responses in Cluster 2 expressed the idea that there was a lack of energy or repeated the question but did not provide any additional information. Fewer than 5% responses explained that the reaction was endergonic or the form of the carbohydrate molecule caused the slow reaction rate. Because these responses are so infrequent, they did not result in separate clusters but were aggregated in Cluster 2. This type of cluster with variety of incorrect and infrequent ideas was also found in the analysis of other questions.

In addition to the cluster percentages and descriptions, some reports contained web diagrams for each cluster, giving a visual representation of the frequencies of category and the association among categories within that cluster (as shown in Figure 1). Larger nodes (circles) indicated higher frequencies, while associations were represented by the lines between nodes (Figure 1b). In Cluster 1, as indicated by its description, the categories *break* and *bond* are most common and are highly associated.

For each cluster, five sample responses were included in the report. Within a cluster, each response was assigned a value based on its distance to the cluster centroid, and the five responses closest to the centroid in each cluster were included in the report. The sample responses gave instructors an opportunity to examine student work directly, have references to discuss with students in the next class, and make an independent interpretation of the cluster if they so desired.

In the cases where pre- and post-instruction assessments were given, instructors were also provided with a summary of how students performed based on the movement among clusters. For example, in Figure 1g, 23.3% of the students who responded to both the pre- and the post-assessment had improved their explanation by moving from Cluster D on the pre-assessment to Cluster 1 on the post-assessment. Cluster D responses were collected prior to instruction and were vague, repeating the question without answering it. Cluster 1 responses were collected after instruction and contained more meaningful explanations. The table shows that 23.3% of students who had vague pre-instruction responses (Cluster D) produced a post-instruction response that was assigned to Cluster 1. In the report, improvements were highlighted in green, while responses that expressed more incorrect ideas post-instruction were highlighted in red.



🟦 Cluster 1 – energy needed to break bonds
🟥 Cluster 2 – miscellaneous
🟨 Cluster 3 – catalyst/enzyme

Web diagram of categories (circle) and links for Cluster 1. Circle size corresponds to the frequency of responses containing a category (see 1 d). Lines indicate the percent shared responses. Solid lines 35% shared responses; dashed lines 25-34, dotted lines 15-24%. Categories with fewer than 15% shared responses were not linked.

c) QUESTION: A carbohydrate is composed of a string of covalently linked monosachharides. Breaking those bonds between the monosaccharides is a chemically spontaneous reaction (ΔG for this reaction is -3.7 kcal/mol). However, this reaction occurs very slowly at room temperature. Why do you think this is so?

d)

Cluster 1		Cluster 2		Cluster 3	
Mean	**Category**	**Mean**	**Category**	**Mean**	**Category**
0.94	bond	0.37	energy	1.00	catalyst/ensyme
0.94	break	0.15	speed	0.38	speed
0.48	energy	0.10	require	0.13	energy
0.33	require	0.04	break	0.10	bond
0.21	catalyst/ensyme	0.03	bond	0.08	break
0.10	speed	0.00	catalyst/ensyme	0.08	require

e) **Description**

Cluster 1 responses state that energy is needed to break the bonds in the carbohydrate	Cluster 2 responses include a mixture of ideas, including that there is a lack of energy, the reaction is endergonic, and the form of carbohydrate affects the rate of the reaction.	Cluster 3 responses state that an catalyst/enzyme is needed to speed up the reaction. However, few talk about the how the enzyme functions

f)

Cluster 1 Examples	Cluster 2 Examples	Cluster 3 Examples
It could be because it takes a lot of energy to break down those bonds so at room temperature the energy isn't enough making it a slow reaction.	This reaction occurs very slowly at room temperature because the reaction is only -3.7 kcal/mol which is a low reaction.	Because there is no enzyme present to speed up the reaction.
The reaction occurs very slowly at room temperature because it is relatively colder than the conditions that the breaking of the bonds between the monosaccharides usually happens. Heat changes the rate of the	Due to the saturation of carbs at room temperature.	because the process doesn't have an enzyme catalyzing the reaction
because it takes a lot of energy to break these bonds.	because it is and endergonic reaction	Because a monosaccharide is not a protein so there are no enzymes to catalyze the reaction for it to speed up.

g)

Pre - Instruction Cluster	Post-Instruction Cluster		
	1. energy needed to break bonds	2. miscellaneous	3. catalyst/enzyme
A. higher temp needed to break bonds	21.6	45.1	33.3
B. higher temp need	32.0	40.0	28.0
C. covalent bonds difficult to break	47.1	35.3	17.6
D. answer repeats the question	23.3	46.7	30.0

Green → moved to a cluster expressing correct or more correct ideas Red → moved to a cluster expressing incorrect ideas
Clusrer Notes: "miscellaneous" contains a mixture of low frequency responses
"repeats question" contains a responses that only included words in the question stem with no further explanation

Figure 1. Sample report from automated lexical and statistical analysis

Faculty focus groups

We also held four 1- 2 hour focus groups with the four participating faculty during which we discussed their participation in this pilot study. The early-semester focus group introduced faculty to the constructed response assessments, text analysis and the utility of the report. We also interviewed faculty about what aspects of the report they would find useful in their classrooms. We met with faculty mid-semester to identify difficulties that they had encountered using the report, allowing us to address those issues. During both the mid-semester and end-of-semester focus groups, faculty described how the report informed their awareness of students' thinking, including prior knowledge, misconceptions, and gaps in their knowledge. Faculty also discussed how they had used the information provided in the feedback report to modify their instruction. Based on these focus group discussions, we describe faculty instruction based on the analysis and report of student writing in the following section.

Faculty interventions and instruction in response to JiTT feedback

Faculty instructors were interested in determining students' prior knowledge about several topics prior to instruction, and identifying student misconceptions or ideas that were challenging to students. After reading the report, the instructors provided students feedback in several different ways. Some instructors created instructional materials, such as a sequence of clicker questions to address these challenges. For example, one instructor created clicker questions to use over multiple class sessions to emphasize the concept that energy is required to break bonds, and how reactions are coupled within biological systems to create a favorable reaction. This coupling is often implicit or overlooked in biology instruction at the introductory level, leaving students with the idea that breaking phosphate bonds in ATP is solely responsible for the energy released during metabolic processes. This faculty member used student sample responses from, or responses similar to those in, the feedback reports as multiple-choice options for the clicker questions. This exercise was designed to help student identify responses options that expressed ideas similar to their own homework responses. Students had the opportunity to discuss their options in groups with their classmates and then groups shared their response with the entire class, which allows them to express their ideas and get feedback from both the instructor and their peers.

Before assigning CR questions, instructors were already familiar with some of the ideas that challenged students as they had encountered these problems in previous semesters with multiple-choice examinations. However, the instructors pointed out that the written assessments were particularly important for gaining insight as to *why* students have struggled continuously with these ideas. One instructor was aware of students' struggles with central dogma concepts, but was finally able to identify that students had not grasped that transcription and translation were different processes using the responses to the CR questions.

Faculty also used materials that were already prepared to address misconceptions such as the conversion of matter to energy in metabolic reactions. Our questions on metabolism were developed from multiple choice items from a diagnostic question cluster (DQC) [8, 9]. Pre-existing clicker questions, created in response to the DQC project, were used by some instructors to revisit misconceptions about photosynthesis and conservation of matter during respiration.

Often with pre-instruction administration of the CR questions, a large fraction of the class was unable to give a correct or relevant response. In some instances the items reviewed material covered in the prerequisite chemistry course (e.g. exergonic reactions). Few introductory science courses have writing practice, and this may be the first attempt for many students to construct a representation of their understanding. Therefore, more opportunities to practice writing may be needed, which could be facilitated by automated analysis. Faculty also proposed future in-class activities to improve student writing skills, including critiques of poorly- and well-written responses gathered from CR questions and opportunities to write in class and turn in work for credit (e.g. minute papers).

Improving the Use of JiTT analysis of constructed response assessments.

During this pilot project, we learned valuable lessons on 1) how to improve the presentation and user-friendliness of reports, 2) how to improve the scheduling and incentivizing of homework assignments, and 3) the need for professional development to support faculty use of these assessments.

Improving report format

In each report, we present a good amount of detail about the ideas that students express in their responses. Faculty found that such detail could be overwhelming especially when first becoming familiar with the reports. They suggested an approach where they are presented with a brief description of the clusters and have the option go on to investigate more details about each cluster or category if they so choose. Based on this discussion, reports now begin with a pie chart (Figure 1a) showing the distribution of responses among clusters with short phrases describing the clusters. Often, just knowing for example, that 49% of students fall into a cluster and hold a particular misconception, is sufficient information for faculty to begin instruction. However, faculty also indicated that they would like details about clusters to be available so they can see what ideas students are using in their responses (Figure 1e,f) and how these ideas are associated within clusters, or differ among clusters Figure (1b,d). This detail is also useful for reflection on one’s teaching at the end of the semester.

Additionally faculty reported that 3-5 clusters were optimal for interpretation. Although the analysis can generate more clusters, with each cluster describing a more fine-grained type of response, we aimed to customize to the instructors' needs and typically presented faculty with 3-5 clusters. Faculty reported that they would only be able to address a few common incorrect ideas or misconceptions in class, and therefore, they would like the report to focus on the major correct and incorrect ideas. Thus, instead of many small clusters, the reports sometimes included one cluster which contained various incorrect ideas, each of which occurred in fewer than 10% of the responses. We also gave faculty examples of responses containing these ideas, in case they did have the opportunity to address these ideas in class.

Encouraging student participation

Each of the three course sections used a different type of incentive to encourage participation. We found that the two sections which gave regular homework credit had better participation (53-83%) than the section which gave extra-credit points (22-46% participation). Additionally, in the section with low participation, there were significant differences in the GPA and course grade of students who participated in homework assignments for extra credit and those who did not (Mann Whitney U –test; $p< 0.005$). In the low-participation section, students who answered CR questions on average entered the course with a higher GPA (2.56±1.37) and obtained a higher grade at the end of the course (2.62±1.06) compared to students who did not participate in the CR online homework (average GPA at start 2.49±1.15; average grade in course 2.00±1.31). This suggests that students who perform more poorly do not often take the opportunity to complete extra credit work and do not get the benefit of the additional practice. Therefore, we suggest instructors using these homework assignments should make them a required part of the regular coursework.

Scheduling

Automated analysis and the generation of reports within a few hours allows faculty to have data about their students' learning immediately available to them. Generally the online homework assignments were due around midnight, analysis began at 9am and reports were ready for faculty before the end of the work day for use in class the next day. The faculty reported that they needed more time than the overnight period to digest the contents of the report and modify their lesson plan. Often this was because faculty had prepared their instructional material days or weeks in advance.

We can address this in three ways
1) The homework assignments could be given earlier: one week or more in advance, especially in the case of pre-instruction assessments. This would give the faculty sufficient time to modify their lesson plans. This approach is less efficient for post-instruction assessments where immediate feedback to students during the next class meeting would be ideal.

2) During the pilot, we usually gave sets of two to six questions for each online homework assignment. Alternatively, faculty could assign just one question that targets a particular misconception. Faculty could modify their lesson plan to address this one misconception, and have material prepared beforehand in the event that there is a considerable fraction of students whose responses suggest that they hold this misconception.
3) A third option would be to design instructional material and provide support to inform faculty instruction based on the results of the constructed response assessments in their classroom. Plans for faculty professional development are discussed in more detail in the following section.

Professional Development for faculty using CR questions and JiTT reports

Faculty were very enthusiastic about using the CR online homework assessments to get students writing and the JiTT reports as a means of evaluating student writing. Because of the quick turnaround time between administering questions, generating the report and having the next class meeting the following day, faculty requested assistance in modifying their instruction to address areas of difficulty for students as identified in the report. Having a suite of materials that would address misconceptions identified by each question would reduce the prep-time required, which is especially important for faculty to make use of the JiTT feedback.

Therefore we are developing materials to accompany questions, so that faculty will have those available when planning their instruction. We are building a community of science education researchers and instructors who will design and test these materials, and make them available for widespread use. These faculty will be part of an online community interested in using constructed response assessments in their classrooms. Faculty will also be able to share resources they have created for their own classroom, such as those developed by faculty who participated in our pilot study. The web portal that will host these online activities is described in the Future Directions section below.

Additionally, we held two meetings with faculty early in the semester to get them familiar with the reports. We will continue to provide this support to faculty, especially as they first begin to use the assessments and instructional material. Faculty who receive support are more likely to continue with the use of innovative research-based instructional materials [14]. Support in implementing a new practice also helps faculty adopt the practice as intended [15].

Future Directions

We are currently investigating the feasibility of developing an automated web-portal, where faculty could upload their own students' responses and receive a feedback report similar to what we have described in this paper. We envision this portal as place where CR questions with developed analytic resources are available for faculty to download and use in their own courses or learning-management systems. An instructor could upload student responses in electronic form and, in a matter of moments, be presented with a feedback report. These reports could contain various levels of detail about the entire class performance or individual students based on the interest of the faculty. Critical to this web-portal idea is the development of a completely automated analysis procedure, which is hidden behind the user interface. A key step in moving in this direction is the validation of clusters/models by both additional student responses, as well as discipline experts. In addition to the CR questions and generated reports, we envision faculty contributing their own experiences or classroom materials in order to address the student difficulties highlighted via the feedback reports. In this way, the portal will facilitate the building of a community of practice: faculty interested in improving their own teaching along with researchers investigating students' learning of science.

Another feature we are considering for the future is how to best return direct feedback to students. Students have expressed interest in learning whether their submitted response was "correct" or "incorrect", in order to gauge their own learning. Although we do not advocate using automated analysis to assign points or "correctness" to individual responses, we may be able to provide students with formative feedback in one of two forms. We may provide a direct report to students that include which concepts they used in their explanation, which cluster their response was placed in or which other responses were most similar to their own, along with information about an "expert" or target answer. Alternatively, each response is assigned a probability of being grouped into a particular cluster, and we can use this information to guide student feedback. In the case of responses with high probabilities of being grouped into a cluster, we may report directly to students the clusters into which their responses fell. In the case of responses with low probabilities, we can recommend that an instructor review these responses before the results are reported to students. This will greatly reduce the number of responses that an instructor will have to read while still providing direct feedback to students. Providing feedback to students may be a key factor in keeping participation rates for the online homework high throughout the semester.

In addition to building an automated analysis web-portal, we want to continue to explore research questions that deal with teaching and learning. We have an assessment structure in place to capture student ideas both before and after instruction. In this way, we can measure change in student ideas and ask questions about whether completing the homework or a particular classroom intervention had an effect on student knowledge. We are also interested in exploring which student difficulties are the most resistant to change. Can we identify common "conceptual-paths" students take as they develop from naive ideas or misconceptions to more sophisticated ideas or scientific ideas? In addition to the change in student thinking, we would like to continue studying what faculty are doing in the classroom. Specifically, what exactly are faculty changing in their instruction, if anything, due to information about their students' responses contained in the feedback report? Does addressing these problems in class or via additional assignments make a difference in student learning? What methods are most effective for addressing these problems? What parts of the feedback report are most meaningful in determining whether to and how to change instruction? We see these as important questions in making progress towards rigorous, reformed science teaching that promote the best outcome for students.

Research Implications

The ability to clearly communicate one's ideas and understanding is an important skill for success in STEM careers[16]. The development of validated CR questions and scoring models for student writing can help change practice in the classroom. With such resources, instructors can include written assessments as a regular form of assessment, even in 400 person classrooms, and students can get practice representing their thinking in their own words. Furthermore, this innovation is highly applicable to other large-scale teaching environments such as the rapidly developing massive open online courses (MOOCs).

The text analysis resources (libraries and categories) used to conduct this study and other analyses of student writing in science can be freely downloaded (with a registered account) on our AACR group website (www.msu.edu/~aacr/). Please visit our site if you are interested in learning more about computerized text analysis in STEM Education.

Acknowledgements

We would like to thank the four instructors who participated in the study and the students in their classes who completed these assignments.

This material is based upon work supported by the National Science Foundation under award 1022653 (DUE). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).

References

1 Pellegrino, J. W., Chudowsky, N., and Glaser, R. (2001) Knowing what students know: The science and design of educational assessment, National Academies Press.

2 Bell, B. and Cowie, B. (2001). The characteristics of formative assessment in science education. Science education, 85, 536–553.

3 Birenbaum, M., and Tatsuoka, K. K. (1987) Open-Ended Versus Multiple-Choice Response Formats—It Does Make a Difference for Diagnostic Purposes, Applied Psychological Measurement 11, 385 –395.

4 Ha, M., and Cha, H. (2009) Pre-service teachers' synthetic view on Darwinism and Larmarckism. In National Association for Research in Science Teaching Conference, Anaheim, CA.

5 Haudek, K. C., Kaplan, J. J., Knight, J., Long, T., Merrill, J., Munn, A., Nehm, R. Smith M., and Urban-Lurain, M. 2011. Harnessing Technology to Improve Formative Assessment of Student Conceptions in STEM: Forging a National Network. CBE-Life Sciences Education 10,149–155.

6 Deane, P. (2006) Strategies for evidence identification through linguistic assessment of textual responses. In Automated scoring of complex tasks in computer based testing (Williamson, D. M., Bejar, I. I., and Mislevy, R. J., Eds.), pp 313–371.

7 Smith, M. K., Wood, W. B., and Knight, J. K. (2008) The Genetics Concept Assessment: A New Concept Inventory for Gauging Student Understanding of Genetics, CBE Life Sciences Education 7, 422–430.

8 Parker, J.M., Anderson, C.W., Heidemann, M., Merrill, J., Merritt, B., Richmond, G., et al. (2012). Exploring Undergraduates' Understanding of Photosynthesis Using Diagnostic Question Clusters. CBE-Life Sciences Education, 11, 47–57.

9 Wilson, C.D., Anderson, C.W., Heidemann, M., Merrill, J.E., Merritt, B.W., Richmond, G., et al. (2006). Assessing students' ability to trace matter in dynamic systems in cell biology. CBE-Life Sciences Education, 5, 323–331.

10 Haudek, K. C., Prevost, L. B., Moscarella, R. A., and Merrill, J. E. (2012). What are they thinking? Automated analysis of student writing about acid/base chemistry in introductory biology. CBE Life Sciences Education 11, 283-293.

11 Prevost, L. B., Haudek, K. C., Merrill, J. E., & Urban-Lurain, M. (2012). Examining student constructed explanations of thermodynamics using lexical analysis. Proceedings of the Frontiers in Education Conference, Seattle, Washington.

12 Novak, G., Gavrin, A., Christian, W., and Patterson, E. (1999) Just-In-Time Teaching: Blending Active Learning with Web Technology 1st ed., Addison-Wesley.

13 IBM (2011) IBM SPSS Modeler 14.2.User's Manual.

14 Henderson, C., Dancy, M., and Niewiadomska-Bugaj, M. (2012) Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process?, Physical Review Special Topics-Physics Education Research 8, 020104.

15 Penberthy, D. L., and Millar, S. B. (2002) The "hand-off" as a flawed approach to disseminating innovation: Lessons from chemistry, Innovative Higher Education 26, 251–270.

16 NRC -CUSE. (1999) Transforming undergraduate education in science, mathematics, engineering, and technology, National Academies Press.