Comparing Formative Feedback Reports: Human and Automated Text Analysis of Constructed Response Questions in Biology

Michele Weston[1], Joyce Parker[2], and Mark Urban-Lurain[3]
[1]Biological Sciences Program
[2]Department of Geological Sciences
[3]Center for Engineering Education Research
Michigan State University East Lansing, MI 48823

This material is based upon work supported by the National Science Foundation (DUE 1022653). 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 NSF.

Abstract

Constructed response questions can offer a detailed look into students' reasoning skills and understanding of key concepts, but take a considerable amount of time to analyze. This trade-off between the amount of time it takes to analyze constructed response questions and their ability to reveal student thinking has made them a desirable, but out-of-reach option, for instructors in large enrollment courses. Automated text analysis can potentially alleviate the time burden of constructed response questions by speeding up the scoring process, while still revealing the level of detail a human reader looks for. This report compares the quality and time needed for two different instructors' analyses of a hand-scored sample of responses to a constructed response question on cell metabolism with an analysis done using statistical modeling of automated text analysis results. We found that the automated text analysis can obtain the same information that an instructor would look for in responses. Additionally, it has the ability to summarize the entire set of responses in virtually the same amount of time. In this study the automated text analysis along with the discriminant analysis took more time than the instructors spent on their analyses, but most of the time consuming work would not need to be repeated with new data in the future.

Introduction

Two of the key purposes of assessment is to provide instructors formative feedback about students' comprehension of course material (NRC 2001) and provide a detailed look into students' reasoning skills and understanding of key concepts that can directly influence classroom instruction (Seymour 2002; AAAS 2009). However, there is a trade-off between the ability of the assessment to reveal student ideas and the amount of time it takes to analyze the data, particularly in large enrollment introductory courses. While there are many ways to obtain formative feedback, research on item types has shown that constructed-response (CR) questions can reveal a wide range of student ideas as well as misconceptions that may otherwise be hidden using multiple choice-type tests (Birenbaum & Tatsouka 1987; Kuechler & Simkin 2010; Lyons, Jones, Merrill, Urban-Lurain, Haudek 2011). Furthermore, CR questions allow students to explain their reasoning, thus giving instructors more insight into what is influencing their thinking (Smith & Tanner 2010).

Instructors who use CR questions must be willing to invest a greater amount of time in reading and evaluating the responses than would be necessary with multiple choice questions. The time difference is even more pronounced for an instructor with little experience reading student writing. In order to save time during their analyses, instructors tend to read a sample of the responses to get an impression of the ideas that students have, and to find patterns that would apply to the larger group.

One option for analyzing responses is to develop a scoring rubric that looks for common student ideas and misconceptions. Using a rubric does not speed up the review process much, but it provides a comprehensive analysis that can be re-applied on new data. If instructors are to use CR questions in large enrollment courses, they need a quicker and more thorough way to evaluate the large dataset of responses.

Automated text analysis can speed the scoring process for CR questions by summarizing the distribution of key concepts and misconceptions in student writing (Haudek, Prevost, Moscarella, Merrill, & Urban-Lurain 2012; Lyons et al. 2011; Prevost, Haudek, Urban-Lurain, & Merrill 2012; Weston et al. 2012). Text analysis identifies words and phrases that can be further analyzed with statistical modeling, including prediction of human scoring (Ha, Nehm, Urban-Lurain, Merrill 2011; Haudek et al. 2012; Nehm, Ha, & Mayfield 2012).

In this study, we investigated whether automated text analysis results can provide insight into student writing that is similar to what an instructor would look for.

Our research questions were:

1. How do instructors analyze answers to CR questions?

2. How do the results from automated text analysis techniques compare with the results from instructors' analyses?

Methods

Our research process follows the Question Development Cycle (QDC) shown in Figure 1. We create a new constructed response question, usually based on the "big ideas" from established concept inventories as they reflect the challenging issues identified by educational researchers (Smith, Wood, & Knight 2008; Wilson et al. 2006). We collect data from students, usually as homework assignments in online learning management systems. We then develop the *lexical resources* needed to analyze the question. This is an iterative step and requires content expertise to determine the optimal sets of lexical *categories* for the question. These categories are then used as variables for *exploratory statistical analyses* that reveal patterns in the student responses. The results of these analyses can help inform the development of *scoring rubrics* that experts can use to *code* student writing. The lexical categories can then be used as independent variables in a variety of statistical procedures to predict the expert score dependent variables in *confirmatory analysis*. The entire process is iterative, with feedback loops informing each of the steps until the model has inter-rater reliability (IRR) with expert scoring on par with expert-to-expert IRR.

Data Collection

In this study, we evaluated student responses to the following question in an introductory biology course at Michigan State University in the fall 2012 semester.

Not all cells in plants (e.g. root cells) contain chlorophyll required for photosynthesis. How do these cells get energy? (Parker et al. 2012).

This question evaluates students' ability to apply their understanding of pathways and transformations of energy and matter, one of the five core concepts in biology (AAAS 2011). We collected responses to this constructed response question on cell metabolism as homework in an online course management system. Students received credit for any attempt to answer the question and 360 out of 468 students responded. The question was given after the cell metabolism unit covering photosynthesis and cellular respiration.

Figure 1. Question Development Cycle

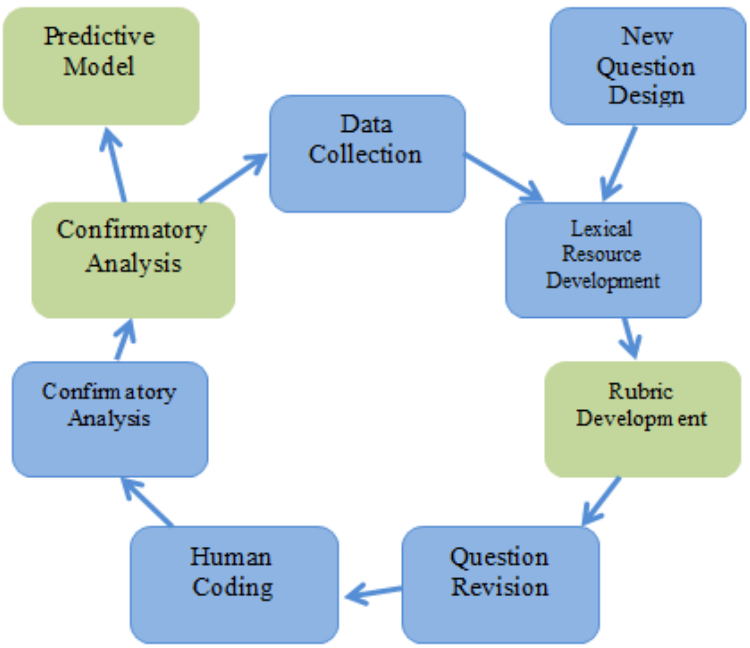

Figure 1. The steps that we use in our question development process. The instructors in this project started at data collection and ended at human coding. The green boxes apply only to the automated analysis.

A random sample of 50 responses was chosen from the data to be read separately by two instructors. The instructors were given directions to read the sample as if they were reading for formative assessment of the students' writing to assess the need for instructional intervention. Both instructors were expert scorers experienced with reading student written responses but had not seen the responses to this question before. The first instructor, Instructor A, spent 11 minutes tallying emergent ideas in the students' writing by categorizing wrong answers. She then calculated descriptive statistics for her findings. The other instructor, Instructor B, spent about 45 minutes on his analysis and developed a 10-bin analytic scoring rubric.

The researchers did an automated analysis of the full 360-response dataset using IBM SPSS Text Analytics for Surveys and the IBM SPSS Modeler 14.2 Text Mining node (IBM SPSS Text Analytics for Surveys v4; & IBM SPSS Modeler v14.2). The Text Mining node of

Modeler extracts *terms* -words and phrases- from written data. These terms are then aggregated in *categories* specified by the user that contain responses with a similar trait or traits, with an attempt to make the categories fairly homogeneous conceptually. For example, the category "*glucose*" includes responses that mention glucose, sugar, cellulose, starch, and G3P or any of their synonyms and misspellings.

Two different statistical models were used to interpret the lexical analysis results, a cluster analysis and a discriminant analysis. The cluster analysis is closely aligned to how an instructor may approach evaluating a sample of student responses. It finds natural groupings (or clusters) based on the distribution of categories in the responses. This information can be used to characterize groups of responses with similar ideas. The discriminant analysis develops a predictive model for response membership in one or more analytic rubric bins using human scoring of data from a previous semester.

Results

Exploratory Analysis: Clustering Based on Emergent Ideas

Neither Instructor 1 nor Instructor 2 had a scoring rubric to use when they read the sample of 50 responses. Without a rubric, often an instructor will begin by placing the responses into groups based on common emergent ideas. This is a useful way to summarize the ideas present in the writing and compare their relative frequencies. Instructor 1's analysis produced five mutually exclusive clusters that identified correct and incorrect ideas being used in the responses (Figure 2). For example, she placed the following response in her cluster number 2 for those that talk about the transport of energy (responses are verbatim and denoted by italics):

These cells take energy from what is around it. For example the root cells take energy from the soil that it is in. Also some cells get their energy transported from the part of the plant that contains chlorophyll.

Instructor 2 also identified key ideas in the responses and kept track of the reoccurring ones. His groups of similar ideas are also shown in Figure 2. He placed the previous example response in his cluster number 4 for responses that talk about both energy being transferred and nutrients from the soil.

Figure 2. Instructor Clusters for Sample of 50 Responses

Instructor 1		Instructor 2	
Distribution of Responses	**Description of Cluster**	**Distribution of Responses**	**Description of Cluster**
15%	1. Response gave an accurate description based on the transport of glucose and/or cellular respiration.	22%	1. Response talks about sugar being transported.
20%	2. Response talked about transport of energy.	14%	2. Response talks about energy being transferred throughout the plant.
14%	3. Response talked about the roots drawing something from the soil for energy.	22%	3. Response says that energy comes from nutrients from the soil.
2%	4. Response gave a force-dynamic style explanation such as another process brings about cell division.	4%	4. Responses talk about energy being transferred and nutrients from the soil.
16%	5. Response mentioned special processes such as C4 processes, Calvin Cycle, and Kreb's cycle.	10%	5. Response names respiration as the process involved.
		10%	6. Response names an incorrect process such as C4 photosynthesis and heterotrophy.

Figure 2. Descriptions of the clusters used by Instructor 1 and Instructor 2 to group responses with similar ideas, and the distribution of responses in those clusters. The clusters are mutually exclusive and are used as a quick way to characterize a sample of student writing

The process of grouping similar responses based on common ideas can be automated using the results of the text analysis. We used the *categories* (Table 1), or groups of related terms extracted from the responses, as independent variables in a K-means cluster analysis to find groupings in the data. We chose six clusters for our analysis because the instructors found five and six clusters. The analysis produced groupings that experts characterized as follows:

1. Response has an incorrect source of energy such as from other organisms. (16 %)
2. Response talks about sugar being transported through the plant. (13%)
3. Response mentions special processes such as the electron transport chain and heterotrophy. (16%)
4. Response talks about nutrients from the soil. (17%)
5. Response names cellular respiration as the process involved. (12%)
6. Response uses energy from parts of the plant that do photosynthesis. (23%)

The cluster analysis placed the previous example response in cluster 6, which is made up of responses that talk about energy coming from parts of the plant that do undergo photosynthesis.

Table 1.
Distribution of Text Analysis Categories in 50-Response Sample and Full Dataset

Category	Distribution in 50-Response Sample	Distribution in Full Dataset
ATP	18%	16%
Absorb	8%	13%
CAM3/CAM4	4%	0.8%
Carbon Dioxide	6%	3%
Chlorophyll	28%	26%
Dark Reactions	6%	5%
Energy	64%	70%
Fermentation	0	0.6%
Fertilizer	2%	0.6%
Glucose	22%	25%
Substances	22%	22%
Leaves	20%	13%
Light Reactions	6%	3%
Mitochondria	6%	6%
Osmosis	0	0
Other Cells	12%	12%
Parasitism	0	0.3%
Photorespiration	0	0.8%
Photosynthesis	18%	24%
Respiration/Glycolysis	20%	15%
Roots/Soil	32%	35%
Solar Radiation	12%	15%
Transport	26%	23%
Water	10%	13%

Note. Responses can be present in multiple categories, which is why they add to more than 100%.

Developing an Analytic Rubric

The results of clustering, whether done by a human reader or an automated analysis, can be used as a foundation for making an analytic scoring rubric (Mueller 2010; Wiggins & McTighe 2005). The first step in creating a scoring rubric for a new question is to read through a sample of responses. Clustering helps organize the many ideas that students have into possible concepts and misconceptions to look for with the rubric. The mutually exclusive clusters can be used to create analytic rubrics where each response can scored independently on each independent rubric. Instructor 2 used the common ideas that he pulled out of the responses to quickly make an analytic rubric (Figure 3). He was then able to score the 50-response sample using his analytic rubric. Instructor 1 only used the clustering for her initial analysis, but went back later and created analytic rubrics from the clusters. Cluster 5 was removed and cluster 1 was split into two analytic rubrics. The rubrics were given new names that summarized the idea they were looking for. This step of making scoring rubrics was done by the instructors and was not automated.

Figure 3. Instructor Analytic Scoring Rubrics

Instructor 1 Analytic Bins	Instructor 2 Analytic Bins
1. Correct source 2. Incorrect source/unspecified energy 3. Incorrect source /nutrients from soil 4. Water as Source 5. Correct Process 6. Incorrect Process	1. Transport sugars 2. Transport energy 3. Transport ATP 4. Respiration 5. Dark Reactions 6. Soil/nutrients 7. From surroundings 8. Transport photosynthesis products 9. C4 photosynthesis 10. Heterotrophy

Figure 3. The analytic rubric bins made by Instructor 1 and Instructor 2. Instructor 2 made his rubric immediately after reading the sample of responses and used it in his analysis. Instructor 1

Scoring with an Analytic Rubric

Once an instructor develops an analytic scoring rubric for a question, he or she can continue using the rubric without having to recreate the clustering and rubric development work for new data. Both instructors could use their rubrics to score a new sample of 50 responses in a future semester and it would take less time than the first iteration. They can also share their rubrics with colleagues who could use the rubrics in their own classrooms once their scoring is calibrated to the expert's scoring.

This process of scoring new data with an analytic rubric can be automated using a combination of lexical analysis and discriminant analysis. The categories created in the lexical analysis can be used as independent variables in a discriminant analysis to predict the human scoring dependent variable. By using expert-scored data to create a statistical model, new data can be scored using this model to predict how experts would score the new data. We performed a discriminant analysis on the 50 response sample using Instructor 1's rubric (Table 2). The model was built using human scoring of 316 responses from a previous semester and tested on the 50-response sample. The scoring was done by Instructor 1 and an assistant after they were calibrated to an inter-rater reliability above 80%. The discriminant analysis predicts human scoring in a leave-one-out system where each case is classified using the functions derived from all other cases. The statistics in Table 2 show how the model performed on the sample of 50 responses.

Table 2.
Discriminant Analysis Results for Instructor 1's Analytic Rubric

Analytic Rubric	Description	Correctly Classified	Kappa
Correct process	Respiration or glycolysis	96%	.884
Incorrect process	Various incorrect processes	92%	.627
Correct source	Any name for a product of photosynthesis	96%	.875
Incorrect source/unspecified energy	Energy or ATP being transported	90%	.718
Incorrect source soil/nutrients	Nutrients from the soil	84%	.543
Incorrect source/water	Water without anything else or water with nutrients from the soil	96%	.730

Note. The discriminant analysis model was built using human scoring of 316 responses from a previous semester and the 50-response sample. That model was used to predict human scoring for the 50-response sample in a leave-one-out system. The correctly classified percent of responses and kappa values are calculated for the 50-response sample.

The two measures of performance in Table 2 are the percent of responses that were correctly classified and the Kappa values of agreement. The percent of correctly classified responses is done by comparing the predicted scoring to actual human scoring. The Kappa coefficient is a measure of inter-rater reliability that takes into account agreement by chance. A kappa rating above .06 is said to be "substantial" and a reading above .08 is "almost perfect" (Landis & Koch 1977).

Discussion

In large enrollment courses, time constraints prevent instructors from reading every student's written response to constructed response questions. This discourages instructors from having students write, even though constructed response assessments provide richer insight into student thinking and conceptual challenges (Bennett 1993; Birenbaum & Tatsouka 1987; Clauser 2000; Hancock 1994; Hogan & Murphy 2007; Kuechler & Simkin 2010; Martinez 1999). The most time consuming steps for instructors are the exploratory analysis and applying the scoring rubric to new data. Instructor 1 spent 7 minutes and Instructor 2 spent about 45 minutes on the exploratory analysis of their 50-response sample. Their time was spent reading through the responses and tallying emergent ideas. Creating the text analysis categories took about 45 minutes and would not need to be repeated with new data. The automated cluster analysis using text analysis results took 15 minutes for the entire 360-response dataset and was comparable to the amount of time spent on it by human readers. Once a rubric is made, it can be applied to new data and takes about the same amount of time to code responses as the initial reading. The discriminant analysis will predict human scoring of the rubric and only takes seconds to score new data, giving instructors the fast turnaround time they need for instructional intervention.

Without automated text analysis, the best option that instructors have for handling large datasets is to read a sample of the responses. Reading and scoring a sample still takes a considerable amount of time and the sample may not be representative of the whole group (Table 1). A concept or misconception could seem more or less frequent in the population than it actually is, interfering with instructional intervention. Once a discriminant analysis model has been created, it allows instructors to evaluate the entire dataset in moments. Instructors can then sample individual responses to read based upon their placement in analytic rubrics or their text analysis categorization.

Human raters face other challenges. One is contextual influences. What an instructor has been teaching in the recent past influences how s/he categorizes responses for formative assessment purposes. Automated analysis can provide more stable scoring that allows for better comparison across courses or semesters. Another challenge human raters face is interpreting responses with unusual wording or poor grammar. It is easy to be distracted by unusual use of language and misunderstand what the student is trying to say. The automated analysis is less dependent on response wording since it analyzes the presence or absence of different combinations of categories in each response, but ignores grammar, reducing one form of bias that can impact human interpretations.

Conclusion

Two steps of our assessment development process, exploratory analysis and coding with a rubric, can be automated using lexical analysis and statistical modeling. The results of the exploratory cluster analyses where comparable to those of human readers. The clusters contained similar ideas to what the instructors found in their exploratory work. The discriminant analysis predicted human scoring accurately for each of Instructor 1's analytic scoring rubrics, with at least 80% agreement for each rubric and at or over 90% agreement on five of the rubrics. The kappa values were all above the .60 threshold for "substantial" agreement, except for the rubric *Incorrect Source Soil/Nutrients* and two of the bins was above the .8 threshold for "almost perfect" agreement. These performance values can improve when more data is available to be added to the training dataset. Once a discriminant analysis model has been developed, it can score written responses with accuracy comparable to human scorers in just a few minutes.

Our first research question was: How do instructors analyze answers to CR questions? In this case, both instructors started by looking for emergent ideas in the students' responses. They identified similar ideas (see Figure 2), but made finer or larger categories.

Our second research question was: How do the results from automated text analysis techniques compare with the results from instructors' analyses? The automated cluster analysis, when set to find a similar number of clusters, identified clusters that overlapped those of the instructors. This implies that the automated cluster analysis would be useful to these instructors for planning just-in-time teaching or other types of feedback to students. The high degree of agreement between the automated discriminant analysis and the human scorers would be useful for comparison of large datasets from course sections experimenting with novel teaching strategies or other course improvements.

Additional Resources

Our question packages, term libraries, and text analysis packages can be found at our website, www.msu.edu/~aacr.

Acknowledgements

The authors would like to thank John Merrill for his assistance creating an instructor rubric and the anonymous reviewers whose comments helped improve the paper.

References

American Association for the Advancement of Science. (2011). Vision and change in undergraduate biology education: A call to action (pp. 11). Washington, DC: American Association for the Advancement of Science.

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 1-27). Hillsdale, N.J.: L. Erlbaum Associates.

Birenbaum, M., & Tatsouka, K. K. (1987). Open-ended versus multiple-choice response formats - It does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11,* 329-341.

Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement, 24(4),* 310-324.

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *Journal of Experimental Education, 62(2),* 143-157.

Haudek, K.C., Prevost, L.B., Moscarella, R.A., Merrill, J.E., & Urban-Lurain, M. (2012). What are they thinking? Automated Analysis of student writing about acid/base chemistry in introductory biology. *CBE Life Sci Educ, 11,* 283-293. doi: 10.1187/cbe.11-08-0084.

Ha, M., Nehm, R.H., Urban-Lurain, M., & Merrill J.E. (2011). Applying computerized-scoring models of written biological explanations across cultures and colleges: Prospects and Limitations. *CBE Life Sci Educ,10,* 379-393. doi: 10.1187/cbe.11-08-0081.

Hogan, T. P., & Murphy, G. (2007). Recommendations for preparing and scoring constructed-response items: What the experts say. *Applied Measurement in Education, 20(4),* 427-441.

Kuechler, W. L., & Simkin, M. G. (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. *Decision Sciences Journal of Innovative Education, 8(1),* 55-73. doi: 10.1111/j.1540-4609.2009.00243.x

Lyons, C., Jones, S., Merrill, J., Urban-Lurain, M., & Haudek, K.C. (2011). *Moving across scales: Using lexical analysis to reveal student reasoning about Photosynthesis.* Paper presented at National Association for Research in Science Teaching International Conference, Orlando, FL.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34(4),* 207-218. doi: 10.1207/s15326985ep3404_2

Mueller, J. (2010). Rubrics Retrieved July 8, 2010, from http://jonathan.mueller.faculty.noctrl.edu/toolbox/rubrics.htm

National Research Council. (2001). Knowing what students know: The Science and Design of educational assessment. Washington, DC: National Academy Press.

Nehm, R.H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Educational Technology 21,*183. doi: 10.1007/s10956-011-9300-9.

Parker, J.M., Anderson, C.W., Heidemann, M., Merrill J., Merritt, B., Richmond, G., & Urban-Lurain, M. (2012). Exploring undergraduates' understanding of photosynthesis using Diagnostic Question Clusters. *CBE Life Sci Educ, 11,* 47-57. doi: 10.1187/cbe.11-07-0054.

Prevost, L.B., Haudek, K.C., Urban-Lurain, M., & Merrill, J.E. (2012). *Examining student constructed explanations of thermodynamics using lexical analysis.* Paper presented at Frontiers in Education, Seattle, WA.

Seymour, E. (2002). Tracking the processes of change in US undergraduate education in science, mathematics, engineering, and technology. *Science Education, 86(1),* 79-105.

Smith, J.I., & Tanner, K. (2010). The problem of revealing how students actually think: Concept inventories and beyond. *CBE: Life Science Education, 9(1),* 1-5.

Smith, M. K., Wood, W. B., & Knight, J. K. (2008). The genetics concept assessment: A new concept inventory for gauging student understanding of genetics. *CBE Life Sci Educ, 7(4),* 422-430. doi: 10.1187/cbe.08-08-0045.

SPSS Modeler (version 14.2) [computer software]. IBM.

SPSS Text Analytics for Surveys (version 4) [computer software]. IBM.

Weston, M., Haudek, K.C., Prevost, L.B., Lyons, C., Urban-Lurain, M., & Merrill, J.E. (2012). *How do biology undergraduates "explain" photosynthesis? Investigating student responses to different constructed response question stems.* Paper presented at National Association for Research in Science Teaching International Conference, Indianapolis, IN.

Wiggins, G., & McTighe, J. (2005). Understanding by design. Alexandria, VA: Association for Supervision and Curriculum Development.

Wilson, C., Anderson, C. W., Heidemann, M., Long, T., Merrill, J., Merritt, B., . . . Parker, J. (2006). Assessing students' ability to trace matter in dynamic systems in cell biology. *Cell Biology Education, 5,* 323-331.