Using Computerized Lexical Analysis of Student Writing to Support Just-in-Time Teaching in Large Enrollment STEM Courses

Mark Urban-Lurain; Luanna Prevost; Kevin C. Haudek; Emily Norton Henry; Mathew Berry
Center for Engineering Education Research
Michigan State University
East Lansing, MI 48824
urban@msu.edu (Mark Urban-Lurain)

John E. Merrill
Bioscience Program
Michigan State University
East Lansing, MI 48824

***Abstract*— We have been exploring a variety of computerized techniques for analyzing student writing in introductory biology. We achieve computer-to-expert inter-rater reliability (IRR) on par with expert-to-expert IRR (> .8). In Fall, 2012, we piloted the use of automated text analysis to facilitate the use of written formative assessment for Just-in-Time Teaching (JiTT) in a large-enrollment introductory biology course at a large public Midwestern university. A total of 12,677 student responses to 15 online homework questions were collected in three 300+ student course sections with four instructors. We used automated analysis to create feedback for instructors before the next class period (less than one working day), so that instructions could use this feedback to inform their instruction. Instructors used many of the questions pre- and post-instruction and the reports we provided to them allowed them to see how their students' answers changed as a result of their instruction. Focus groups with the instructors revealed that they already knew some of the topics that challenged students, as revealed 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.**

Keywords—large-enrollment introductory courses; constructed responses; lexical analysis; Just-in-Time Teaching (JiTT)

I. INTRODUCTION (*Heading 1*)

Developing rich, reliable, and robust measures of the composition, structure, and stability of student thinking about core scientific ideas (such as natural selection, conservation of mass and energy, and genetics) is a challenge that may be too complex to accomplish via multiple-choice assessments such as concept inventories (CIs). For example, as Nehm & Schonfeld demonstrate, the multiple-choice Concept Inventory of Natural Selection measures whether students understand "pieces" or elements of the theory of natural selection, but does not provide any measure of students' abilities to assemble the pieces into a coherent and functional explanatory structure [1, 2]. Moreover, multiple-choice CIs introduce significant validity threats as they are constrained to "either-or" forced-choice ("misconception" vs. scientific key concept) item preference and do not typically allow the detection of students who harbor "mixed models" of correct and incorrect conceptions [1, 3-8].

Multiple-choice assessments also require different cognitive processes (recognition, selection) than constructed response (CR) assessments in which students must represent their ideas in writing or by creating other models. CR assessments are widely viewed as providing greater insight into student thinking than closed form (e.g., multiple-choice) assessments [9] which encourage students to study by memorizing, rather than learning critical thinking and analytic skills that are crucial for success in all STEM disciplines [10].

Thus, CR assessments that capture students' explanatory models are needed to mitigate these constraints and reveal students' mixed models. In the past, financial and time constraints made CR assessments significantly more challenging to execute in large-enrollment courses than multiple-choice assessments. But today, advances in both technology and measurement research make it feasible to apply these techniques in instructional settings with the potential to have substantial educational impact [11].

In the Automated Analysis of Constructed Response (AACR) research group (URL) we employ cutting-edge, lexical and computer analysis technology to analyze student writing in biology and chemistry. We have been able to create statistical models of student writing that predict expert human raters' scoring with computer-to-expert inter-rater reliability (IRR) that is similar to expert-to-expert IRR (generally, > .8) [8, 12-16].

In this paper, we report on our initial efforts to move these techniques from research-to-practice in a pilot study in which we investigate applying our models to support Just-in-Time Teaching (JiTT) [(JiTT) 17] by providing instructors formative feedback about students' writing through an automated

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)..

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.

II. Methodological Details of the AACR Approach

In this section, we provide an overview of our approach to developing, validating and implementing AACR assessments as background for the work we will do in the proposed project. The entire process is captured by the Question Development Cycle (QDC) shown in Figure 1. In general, we use linguistic feature-based methods [18] to extract linguistic features from students' writing [e.g., WordNet, see 19, 20] and then use those linguistic features as variables in statistical models that predict human raters' scores of the student's writing.

A. *Developing AACR Questions to Assess Core Disciplinary Concepts*

In the first stage of the QDC, we *Design New Questions* to measure student thinking about important disciplinary constructs. Generally, we use concept inventories as the basis for the questions because they represent the topics that disciplinary researchers have identified as being particularly challenging for students. *Data Collection* is typically done by administering the questions via on-line course management systems into which students can enter their responses. *Lexical Resource Development* is done using lexical analysis software to extract key terms and scientific concepts from the students' writing. These terms and concepts are used as variables for *Exploratory Analysis* which aid in *Rubric Development*. We use the rubrics, both analytic and holistic, for *Human Coding* of student responses. During *Confirmatory Analysis* the *Lexical Resources* are used as dependent variables in statistical classification techniques to predict expert human coding of student responses. The entire process is iterative with feedback from the various stages informing the refinement of other components. The final product of the QDC is a *Predictive Model* that can be used to completely automate the scoring of a new set of student responses, predicting how experts would score the responses.

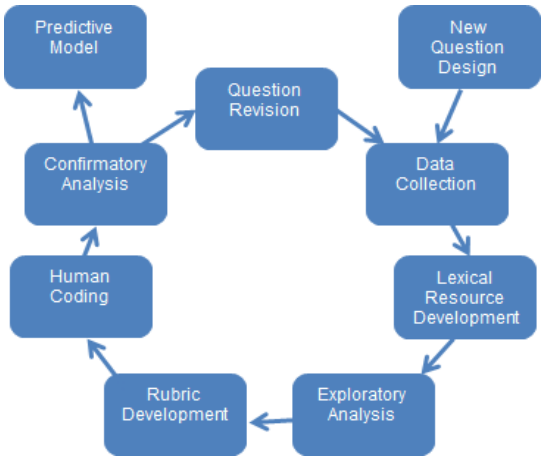

Fig. 1. Question Development Cycle (QDC)

An example of an introductory biology question for which we have completed the QDC is: **Jared, the "Subway" guy, lost over 200 pounds on his diet. Where did his mass go?** This question is designed to reveal students' ability to reason about pathways and transformations of energy and matter, one of five core biology concepts [21] for which we are developing AACR assessments [22-24]. In the following sections we elaborate on the lexical resource development and exploratory analysis phases of the QDC for the Jared problem. We first outline the process for validating the assessment and then we show how instructors can implement the AACR questions in the classroom.

B. *Validating AACR Assessments through Lexical and Confirmatory Analysis*

We use IBM SPSS Modeler [25] to perform the lexical and statistical analyses. Modeler provides data mining tools that can be used to build Modeler streams (Figure 2a) to automate analyses by assembling nodes that perform various tasks, such as accessing and merging data files, data conversions, lexical analysis, statistical analysis, machine learning, and reporting. Following the order of the nodes in Figure 2a, for example, we collect student responses (from on-line homework) to the AACR question to be analyzed, in this case the question about Jared's weight-loss. The responses are processed by the text analysis node.

Figure 2b shows some details of the text analysis node. The software extracts terms -- words and phrases in the students' responses that are relevant to the question (colored text Figure 2b, middle panel). These terms are stored in libraries (similar to dictionaries) that come with the software or were created by the researchers. Extracted terms that represent homogeneous disciplinary concepts are grouped into categories (Figure 2b, left panel), using both automated procedures and refinement by content experts. For example, the category glucose/glycogen in Figure 2b includes a number of terms (e.g., glucose, glycogen, sugar, and sugar molecules) that represent molecules that are metabolized to release carbon dioxide. Each student response is classified into one or more categories based on the terms used in that response (Figure 2b, right panel).

Continuing along the stream (Figure 2a), the text analysis categories are used as independent variables in statistical analysis or machine learning nodes. In the exploratory phase, as demonstrated in this example, we use cluster analyses to group responses that have the most similar sets of categories (Figure 3 shows example cluster results). These clusters help researchers refine the rubrics that are used for human scoring to build confirmatory models (e.g., discriminant analysis and machine learning techniques) that predict human scoring with computer-to-expert inter-rater reliability (IRR) as good as expert-to-expert IRR [8, 12, 15]. The final nodes of the stream select examples of student work most representative of the cluster, (i.e. closest to the cluster centroid). This information was used to build Just-in-Time-Teaching reports.

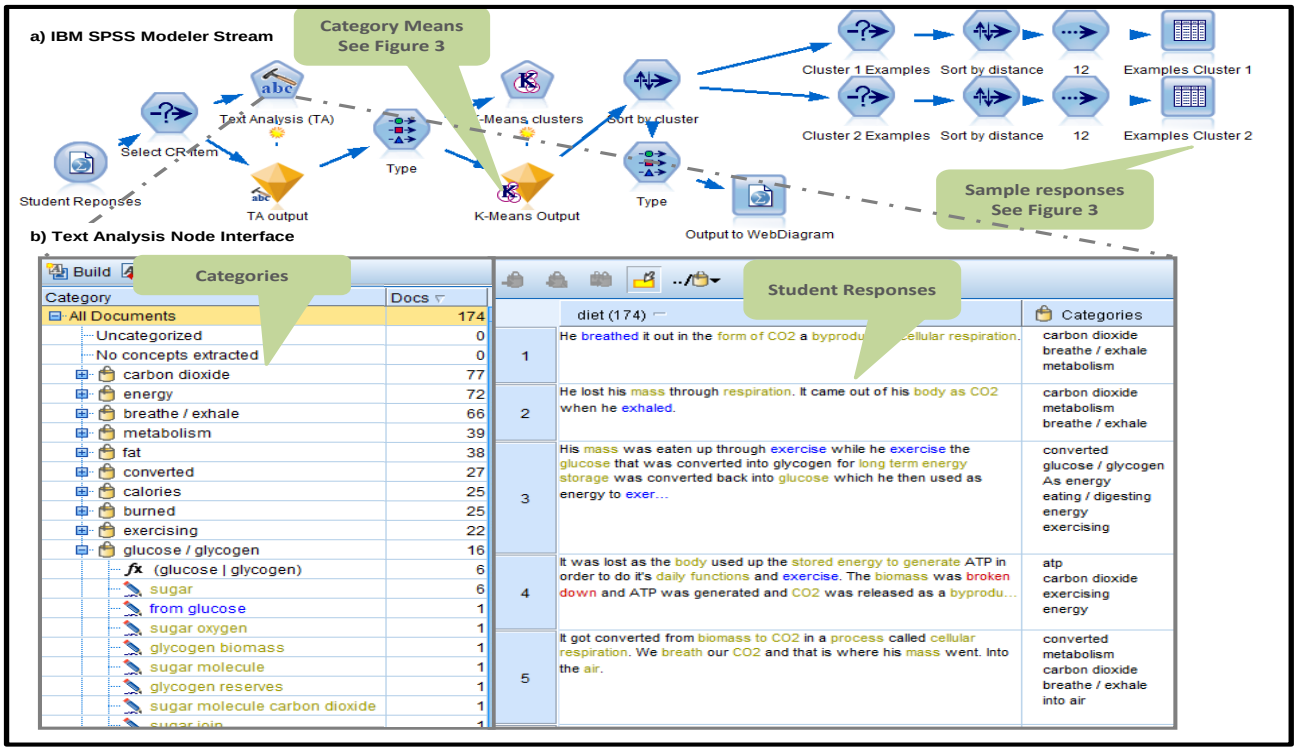

Fig. 3. IBM-SPSS Modeler showing a Report Analysis Stream (a) and Text Analysis Node (b) for the assessment question: *Jared, the "Subway" guy, lost over 200 pounds on his diet. Where did his mass go?*

III. Implementing AACR Questions for Just-in-Time Teaching

To test the feasibility of accelerating the QDC (Figure 1) and rapidly making the research results available to faculty in near real time, we conducted a JiTT pilot study during fall, 2012, in three sections (N=309; N=302; N=455) of an introductory cells and molecules biology course for science majors at MSU [26]. We administered 15 different homework questions in four subject areas: biomolecules, genetics, metabolism, and thermodynamics using the university's Learning Management System (LMS). Questions were asked pre-instruction, so that the responses could be analyzed and a report returned to the instructors to allow them to address misconceptions during the next class period. Some questions were also asked post-instruction, which allowed instructors to see how students' explanations had changed. We collected 12,677 student responses and used previously created SPSS Modeler streams (Figure 2) to generate the JiTT reports. For each question we asked, data collection closed at midnight; analysis and report preparation began the following morning; and reports were completed and emailed to instructors in the afternoon for use during the next class period.

Some features from a report for the Jared question are presented in Figure 3. Reports included the question asked, the category means within each cluster (the percentage of responses classified in this category within a given cluster), cluster descriptions, example student responses that were most representative (defined by the statistical distance from their cluster centroids) and a web diagram showing the relationships

QUESTION: Jared, the "Subway" guy, lost over 200 pounds on his diet. Where did his mass go?

Cluster 1		Cluster 2	
Mean	Category	Mean	Category
0.65	carbon dioxide	0.98	energy
0.61	breathe/exhale	0.40	converted
0.23	metabolism	0.35	fat
0.05	converted	0.13	calories
0.04	energy	0.13	glucose/glycogen
0.04	fat	0.06	sweat
0.02	glucose/glycogen	0.04	metabolism
0.01	sweat	0.00	breathe/exhale

Cluster Descriptions	
Cluster 1 responses state the mass was lost as carbon dioxide and expelled through exhalation, breathing or respiration.	Cluster 2 responses describe the fat as being converted into energy and possibly being used or released.
Cluster 1 Examples	**Cluster 2 Examples**
His mass was lost through breathing out carbon dioxide.	His mass was lost as energy exerted.
He breathed it out in the form of carbon dioxide.	It was let out into the atmosphere as energy.

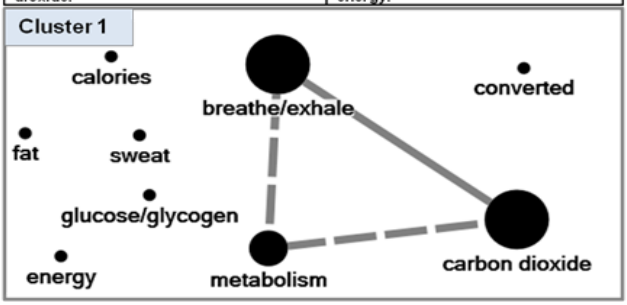

Fig. 2. Subset of Pilot Study JiTT Faculty Feedback Report Features **Legend**: Circle size corresponds to the frequency of responses containing a category. Lines indicate the percentage of shared responses. Solid lines indicate >50% shared responses.Dashed lines indicate 25-50% shared responses. Nodes < 25% shared responses were not linked.

among categories in students' answers. For most questions, responses were classified 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 the frequencies distributions of categories in each cluster.

For the analysis of the Jared question (Figure 3), we see that students in Cluster 1 write about Jared's mass being converted to carbon dioxide and expelled from the body. Student answers in this cluster had high means (frequencies) for carbon dioxide (65% of the responses in Cluster 1) and breathe/exhale (61% of the responses in Cluster 1) categories. The web diagram shows that responses in Cluster 1 have strong associations (solid line) between these two categories, as shown in the Cluster 1 example student responses, meaning that students in Cluster 1 tend to write about these ideas together. Cluster 2, however, had high means for the categories energy, converted, and fat. These students wrote that Jared's mass was converted into energy, revealing a common misconception for introductory biology students [22] as shown in the Cluster 2 example student answers.

IV. Pilot Study: Instructor Use of the 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). Instructors then used these reports in a variety of ways in their instruction.

A. Faculty focus groups

We 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.

B. 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' confusion about 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).

C. 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.

D. 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.

V. 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 [27]. Support in implementing a new practice also helps faculty adopt the practice as intended [28].

VI. 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 faculty 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.

VII. Conclusion

If we are to heed the call for promoting higher order student thinking and providing more opportunities for students to write, while at the same time containing costs, we must find ways to leverage technology in the service of supporting and evaluating constructed response assessments. The approaches outlined in this paper demonstrate the feasibility of using off-the-shelf analytic software to allow instructors to 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 (URL). Please visit our site if you are interested in learning more about computerized text analysis in STEM Education.

Acknowledgment

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

References

[1] R.H. Nehm and I.S. Schonfeld, “Measuring knowledge of natural selection: A comparison of the CINS, an open-response instrument, and an oral interview,” Journal of Research in Science Teaching, vol. 45, (no. 10), pp. 1131-1160, 2008.

[2] R.H. Nehm and I.S. Schonfeld, “The future of natural selection knowledge measurement: A reply to Anderson et al (2010),” Journal of Research in Science Teaching, vol. 47, (no. 3), pp. 358-362, 2010.

[3] R.H. Nehm and L. Reilly, “Biology majors’ knowledge and misconceptions of natural selection,” BioScience, vol. 57, (no. 3), pp. 263 - 272, March 2007.

[4] M. Ha and H. Cha, “Pre-service teachers' synthetic view on Darwinism and Lamarckism,” in Book Pre-service teachers' synthetic view on Darwinism and Lamarckism, Series Pre-service teachers' synthetic view on Darwinism and Lamarckism, Editor ed.^eds., City, 2009.

[5] R.H. Nehm, H. Haertig, and J. Ridgway, “Human vs. computer diagnosis of mental models of natural selection: Testing the efficacy of lexical analyses of open response text,” in Proc. Transforming Undergraduate Biology Education: Mobilizing the Community for Change, 2009, pp. Pages.

[6] R.H. Nehm, S.Y. Kim, and K. Sheppard, “Academic preparation in biology and advocacy for teaching evolution: Biology versus non-biology teachers,” Science Education, vol. 93, (no. 6), pp. 1122-1146, 2009.

[7] R.H. Nehm, T.M. Poole, M.E. Lyford, S.G. Hoskins, L. Carruth, B.E. Ewers, and P.J.S. Colberg, “Does the segregation of evolution in biology textbooks and introductory courses reinforce students’ faulty mental models of biology and evolution? ,” Evolution: Education and Outreach, vol. 2, (no. 3), pp. 527-532, September 2009.

[8] K.C. Haudek, L.B. Prevost, R.A. Moscarella, J.E. Merrill, and M. Urban-Lurain, “What are they thinking? Automated analysis of student writing about acid/base chemistry in introductory biology,” CBE - Life Sciences Education, vol. 11, (no. 3), pp. 283-293, September 2012.

[9] M. Birenbaum and K.K. Tatsouka, “Open-ended versus multiple-choice response formats - It does make a difference for diagnostic purposes,” Applied Psychological Measurement, vol. 11, pp. 329-341, 1987.

[10] K.F. Stanger-Hall, “Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes,” CBE-Life Sciences Education, vol. 11, (no. 3), pp. 294-306, September 4, 2012 2012.

[11] R.E. Bennett, “Moving the field forward: Some thoughts on validity and automated scoring,” in Automated scoring of complex tasks in computer-based testing, D. M. Williamson, I. I. Bejar and R. J. Mislevy eds., Mahwah, N. J.: Lawrence Erlbaum Associates, 2006, pp. 403-412.

[12] K.C. Haudek, J.J. Kaplan, J. Knight, T. Long, J. Merrill, A. Munn, R. Nehm, M. Smith, and M. Urban-Lurain, “Harnessing technology to improve formative assessment of student conceptions in STEM: Forging a national network,” CBE - Life Sciences Education, vol. 10, pp. 149-155, Summer 2011.

[13] R.H. Nehm, M. Ha, and E. Mayfield, “Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations,” Journal of Science Education and Technology, vol. 21, (no. 1), pp. 183-196, February 2012.

[14] J.E. Opfer, R.H. Nehm, and M. Ha, “Cognitive foundations for science assessment design: Knowing what students know about evolution,” Journal of Research in Science Teaching, vol. 49, (no. 6), pp. 744-777, 2012.

[15] M. Ha, R. Nehm, M. Urban-Lurain, and J.E. Merrill, “Applying computerized scoring models of written biological explanations across courses and colleges: Prospects and limitations,” CBE - Life Sciences Education, vol. 10, (no. 4), pp. 379-393, Winter 2011.

[16] M. Urban-Lurain, R.A. Moscarella, K.C. Haudek, E. Giese, D.F. Sibley, and J.E. Merrill, "Beyond multiple choice exams: Using computerized lexical analysis to understand students' conceptual reasoning in STEM disciplines " in Book Beyond multiple choice exams: Using computerized lexical analysis to understand students' conceptual reasoning in STEM disciplines vol. 39, Series Beyond multiple choice exams: Using computerized lexical analysis to understand students' conceptual reasoning in STEM disciplines Editor ed.^eds., City: ASEE/IEEE 2009.

[17] G.M. Novak, A. Gavrini, W. Christian, and E. Patterson, Just-in-time teaching: Blending active learning with web technology: Addison-Wesley, 1999.

[18] P. Deane, "Strategies for evidence identification through linguistic assessment of textual responses," in Automated scoring of complex tasks in computer-based testing, D. M. Williamson, I. I. Bejar and R. J. Mislevy eds., Mahwah, N. J.: Lawrence Erlbaum Associates, 2006, pp. 313-372.

[19] C. Fellbaum, WordNet: An electronic lexical database, Cambridge, Mass.: MIT Press, 1998.

[20] G.A. Miller, "WordNet: a lexical database for English," Communications of the ACM, vol. 38, (no. 11), pp. 39-41, November 1995.

[21] American Association for the Advancement of Science, "Vision and change in undergraduate biology education: A call to action," American Association for the Advancement of Science, Washington, DC, July 15-17 2011.

[22] C. Wilson, C.W. Anderson, M. Heidemann, T. Long, J. Merrill, B. Merritt, G. Richmond, D. Sibley, and J. Parker, "Assessing students' ability to trace matter in dynamic systems in cell biology," Cell Biology Education, vol. 5, pp. 323-331, 2006.

[23] G. Richmond, B. Merritt, M. Urban-Lurain, and J. Parker, "The development of a conceptual framework and tools to assess undergraduates' principled use of models in cellular biology," Cell Biology Education, vol. 9, (no. 4), pp. 441-452, 2010.

[24] M. Urban-Lurain, R.A. Moscarella, K.C. Haudek, E. Giese, J.E. Merrill, and D.F. Sibley, "Insight into Student Thinking in STEM: Lessons Learned from Lexical Analysis of Student Writing," in Book Insight into Student Thinking in STEM: Lessons Learned from Lexical Analysis of Student Writing, vol. 83, Series Insight into Student Thinking in STEM: Lessons Learned from Lexical Analysis of Student Writing, Editor ed.^eds., City: NARST, 2010, pp. 19.

[25] IBM, "IBM SPSS Modeler Version 14.2," in Book IBM SPSS Modeler Version 14.2, Series IBM SPSS Modeler Version 14.2, Editor ed.^eds., City, 2011.

[26] L.B. Prevost, K.C. Haudek, E. Norton, M. Berry, J.E. Merrill, and M. Urban-Lurain, "Automated text analysis facilitates using written formative assessments for just-in-time teaching in large enrollment courses," in Book Automated text analysis facilitates using written formative assessments for just-in-time teaching in large enrollment courses, Series Automated text analysis facilitates using written formative assessments for just-in-time teaching in large enrollment courses, Editor ed.^eds., City: IEEE, In Review.

[27] C. Henderson, M.H. Dancy, and M. Niewiadomska-Bugaj, "Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process?," Physical Review Special Topics - Physics Education Research, vol. 8, (no. 2), pp. 020104-1 - 020104-15, July 31 2012.

[28] D.L. Penberthy and S.B. Millar, "The "Hand-off" as a flawed approach to disseminating innovation: Lessons from chemistry," Innovative Higher Education, vol. 26, (no. 4), pp. 251-270, Summer 2002.