CBE—Life Sciences Education
Vol. 10, 149–155, Summer 2011

Feature
Meeting Report

Harnessing Technology to Improve Formative Assessment of Student Conceptions in STEM: Forging a National Network

Kevin C. Haudek,[*] Jennifer J. Kaplan,[†] Jennifer Knight,[‡§] Tammy Long,[‡||] John Merrill,[‡¶] Alan Munn,[‡#] Ross Nehm,[‡@] Michelle Smith,[‡] and Mark Urban-Lurain[‡††]**

[*]Division of Science and Mathematics Education, Michigan State University, East Lansing, MI 48824; [†]Department of Statistics and Probability, Division of Science and Mathematics Education, Michigan State University, East Lansing, MI 48824; [§]Department of Molecular Cell and Developmental Biology, University of Colorado, Boulder, CO 80309; [||]Center for Integrative Studies in General Sciences and Department of Plant Biology, Michigan State University, East Lansing, MI 48824; [¶]Biological Sciences Program, Michigan State University, East Lansing, MI 48824; [#]Department of Linguistics and Languages, Michigan State University, East Lansing, MI 48824; [@]School of Teaching and Learning, EEOB, The Ohio State University, Columbus, OH 43210; [**]Department of Genome Sciences, University of Washington, Seattle, WA 98195; and [††]Center for Engineering Education Research, Michigan State University, East Lansing, MI 48824

Concept inventories, consisting of multiple-choice questions designed around common student misconceptions, are designed to reveal student thinking. However, students often have complex, heterogeneous ideas about scientific concepts. Constructed-response assessments, in which students must create their own answer, may better reveal students' thinking, but are time- and resource-intensive to evaluate. This report describes the initial meeting of a National Science Foundation–funded cross-institutional collaboration of interdisciplinary science, technology, engineering, and mathematics (STEM) education researchers interested in exploring the use of automated text analysis to evaluate constructed-response assessments. Participants at the meeting shared existing work on lexical analysis and concept inventories, participated in technology demonstrations and workshops, and discussed research goals. We are seeking interested collaborators to join our research community.

INTRODUCTION

For more than 20 years, there have been calls for improving science, technology, engineering, and mathematics (STEM) education (Tobias, 1990; National Science Foundation [NSF], 1996; Seymour and Hewitt, 1997; National Research Council, 1999; Kardash and Wallace, 2001; Ruiz-Primo *et al.*, 2002; Seymour, 2002; Gess-Newsome *et al.*, 2003; American Association for the Advancement of Science [AAAS], 2009). A common recommendation is to move STEM instruction away from teaching and assessing the "facts" to helping students acquire deeper conceptual understanding and transferable problem-solving skills. Meaningful assessments that reveal student thinking are vital to these efforts (Pellegrino *et al.*, 2001). To this end, much work in science education has been devoted to developing concept inventories for formative assessment of students' understanding of important "big ideas" in science (D'Avanzo, 2008; Libarkin, 2008; Knight, 2010). Concept inventories are typically multiple-choice assessments in which the distracters

DOI: 10.1187/cbe.11-03-0019
[‡]These authors contributed equally to this report and are listed alphabetically.
Address correspondence to: Mark Urban-Lurain (urban@msu.edu).

are derived from common student misconceptions. These misconceptions were generated by educational research on student thinking, and alternative conceptions about the big ideas in STEM disciplines (Duit, 2009) are identified by asking students to construct explanations to questions either in interviews or in writing.

In constructed-response questions, also referred to as open-response or short-answer questions, students must write or create an answer or explanation using their own words. In some cases, constructed-response questions have been shown to better reveal students' understanding than multiple-choice questions (Birenbaum and Tatsouka, 1987; Bridgeman, 1992; Bennett and Ward, 1993; Kuechler and Simkin, 2010). In the domain of natural selection, for example, Nehm and Schonfeld (2008) showed that constructed-response scores have greater correspondence with oral interview scores than multiple-choice test scores. This greater correspondence may be explained by the fact that different cognitive processes are enlisted when providing reasoning instead of selecting an answer from several choices (Kuechler and Simkin, 2003, 2010).

The use of constructed-response tests is constrained by the time and expertise needed to score them, particularly in large-enrollment introductory courses. Recent advances in technology and natural language processing, however, have made computerized analysis of writing possible and may facilitate the analysis of large numbers of written responses. In this report, we describe the initial meeting of a group of science education researchers that is exploring the use of such technologies to automate the evaluation of constructed-response assessments.

The Automated Analysis of Constructed Responses (AACR; http://aacr.crcstl.msu.edu) research group consists of researchers from seven universities with backgrounds in various STEM disciplines (biology, geosciences, statistics, and technology), linguistics, and educational research. We hope this research can help us gain greater insight into student thinking about "big ideas," such as evolution, energy, and genetics. We are looking to expand the scope of this work and are seeking interested collaborators across STEM disciplines.

MEETING OVERVIEW

Members of AACR recently received NSF funding for the project Collaborative Research: Automated Analysis of Constructed Response Concept Inventories to Reveal Student Thinking: Forging a National Network for Innovative Assessment Methods (NSF DUE-1022653). On November 19–20, 2010, we held the initial meeting of this project at Michigan State University (MSU) in East Lansing. Attending the meeting were 18 participants from our 7 participating universities (Figure 1). Goals of this initial meeting included the following:

- Discuss several software options for analyzing students' written explanations and provide a hands-on workshop on using text analysis software packages
- Review existing work in automated analysis, determine ways to extend and build upon this work, and identify synergies and challenges among projects
- Identify research questions of interest across the research projects and develop a plan for collaborative data collection to address the research questions

TOOLS UNDER INVESTIGATION

Currently, AACR is investigating the utility of two different software packages for automated text analysis: IBM(R) SPSS(R) Text Analytics for Surveys (STAS; SPSS, 2009) and The Summarization Integrated Development Environment (SIDE; Mayfield and Rose, 2010a, 2010b).

STAS is commercial, lexical analysis software originally designed for marketing research to process primarily affective responses from survey questions, such as "How did you enjoy your stay in Hotel X?" Current work in AACR extends the utility of this software to analyze student responses to scientific questions. STAS software builds language categories, which typically contain multiple terms and/or functions that combine terms using Boolean logic. Categories are automatically generated by the software and then can be honed by subject matter experts. Deciding how detailed or broad categories should be requires expert input. For example, a category named "Cellular Respiration" could contain the terms *glycolysis, Krebs cycle, oxidative phosphorylation,* and *electron transport chain*. There may be other questions, however, for which the expected responses require the terms *glycolysis* and *oxidative phosphorylation* to be separated and placed in two different categories. Other types of categories are those created using students' novel and emerging ideas. These categories are usually developed iteratively through careful examination of student writing and the lexical analysis output. STAS

Figure 1. Participants in the initial meeting of Forging a National Network for Innovative Assessment Methods. Shown in the picture (from left to right): Olga Eremina, Kevin Haudek, John Merrill, Alan Munn, Jenny Knight, Ross Nehm, Michelle Smith, Mark Urban-Lurain, Jennifer Kaplan, Julie Libarkin, Merle Heidemann, Mary Anne Sydlik, Minsu Ha, Brittany Shaffer, Tammy Long, and Casey Lyons. Not pictured: Hendrik Haertig and Shauna Jones.

supports this iterative refinement, but a person must make the decisions about the categories.

An alternative approach to text analysis, SIDE uses machine learning methods to analyze text responses. SIDE is an open-source project developed by researchers at Carnegie Mellon University (www.cs.cmu.edu/~cprose/SIDE.html) to create computer scoring models that predict human expert scoring of responses. SIDE takes a set of human-scored responses (that is, a spreadsheet of responses that have been scored for the presence or absence of particular ideas) and "discovers" word patterns that account for human-generated scores. SIDE performs much of the difficult work of figuring out what elements differentiate an accurate response from an inaccurate response, or a response in which a series of words that represents a concept is present or absent. SIDE then automatically applies the rules it "learned" from human scoring to a new set of responses and determines how well the rules work using Kappa agreement values. A major strength of SIDE is that much of the rule building is automated. A weakness is that the rules are opaque; the specific reasons for categorizing responses are not described by SIDE and are based on complex algorithms.

As part of the meeting, participants were involved in two mini-workshops: one focusing on STAS and the other on SIDE. In both workshops, participants were able to practice with sample sets of data. Typical data sets range from 100 to 1000 student responses, each of which may be from a single word to several sentences long. Both software programs are able to read data contained in spreadsheets. Data can be collected online (using a course management system or web-based survey software) or transcribed from handwritten responses. With these data sets, both programs are able to process the data in one to two minutes. Some of the lexical resources for STAS are currently available online at http://aacr.crcstl.msu.edu/resources. Likewise, tutorials on how to use SIDE and STAS are available at http://evolutionassessment.org.

REVIEW EXISTING WORK

Each research group presented their previous work and how lexical analysis might guide future directions in their research. After each presentation, meeting participants discussed implications and possible interactions among the research groups.

Cellular Metabolism

Mark Urban-Lurain and John Merrill presented the summary of the lexical analysis work in cellular metabolism that has been done by AACR at MSU (NSF DUE 07236952). AACR extended work of the Diagnostic Question Cluster research group, focusing on students' understanding of key concepts in molecular and cellular biology (e.g., tracing matter, energy, and information). These "big ideas" align with the Vision and Change recommendations (AAAS, 2009). AACR has been using the STAS software described above (SPSS, 2009).

The MSU group takes a two-stage, feature-based approach (Deane, 2006) to analyze constructed responses. First, they create items designed to identify common student conceptions based on prior research. They ask these questions in online course management systems in which students can enter their responses. They use STAS to extract key terms from the students' writing. The software places these terms into categories that are then used as variables for statistical classification techniques to predict expert ratings of student responses. The entire process is iterative with feedback from the various stages informing the refinement of other components.

Table 1. A multiple-choice question developed to assess students' ability to follow matter during cellular respiration.

	You have a friend who lost 15 pounds of fat on a diet. Where did the mass go?	
A.	**The mass was released as CO_2 and H_2O.**	44%
B.	The mass was converted into energy and used up.	23%
C.	The mass was converted into ATP molecules.	21%
D.	The mass was broken down into amino acids and eliminated from the body.	9%
E.	The mass was converted to urine and feces and eliminated from the body.	3%

Correct answer (A) is in bold font. Each distracter (B–E) represents a common student misconception. Percent of students ($n = 459$) selecting each answer is shown on the right.

Constructed-response questions may reveal a richer picture of student thinking than is possible using multiple-choice items alone. When students answer a multiple-choice question about weight loss (Wilson *et al.*, 2006), about 44% of students correctly identify the mass released as carbon dioxide and water (Table 1, choice A). The other responses are distributed across other distracters representing common student misconceptions, such as matter being converted to energy (Table 1, choice B) or mass being excreted as waste (Table 1, choice E). Automated analysis of student responses to the constructed-response version of the question (*You have a friend who lost 15 pounds of fat on a diet. Where did the mass go?*) reveals similar ideas, but also the heterogeneous nature of student ideas. For example, many students who correctly say the mass is lost as carbon dioxide and water also discuss the incorrect ideas of converting matter to energy or losing mass as waste. Examples from several other projects (Moscarella *et al.*, 2008; Haudek *et al.*, 2009; Urban-Lurain *et al.*, 2009; Urban-Lurain *et al.*, 2010) were also discussed at the meeting.

Evolution and Natural Selection

Ross Nehm, Minsu Ha, and Hendrik Haertig reviewed their recent findings from lexical analysis and related assessment research on evolution and natural selection at The Ohio State University (OSU; http://evolutionassessment.org; NSF REESE 0909999). The OSU group has been using both STAS and SIDE, but tackled the challenge of lexical analysis of constructed-response text using a slightly different approach than the other research groups. In particular, lexical analyses were begun using a "construct-grounded" approach. That is, the OSU project began by documenting the necessary and sufficient explanatory elements that experts would expect from an accurate and complete explanation of evolutionary change via natural selection (Nehm and Schonfeld, 2010). They did this by examining the professional scientific literature in

biology. Rubrics were built to identify student language that corresponded to the construct of evolutionary change based on the expert expectations. This approach is in contrast to an exploratory approach in which the rubric categories are built from topics generated from student responses. This provides an example of different strategies that may be used to develop rubrics for text analysis scoring.

As the construct of natural selection is generally well established, rubrics were subsequently developed by the OSU group to identify the words and phrases that students commonly use to represent each of these scientific elements (referred to as "Core Concepts"; Nehm *et al.*, 2010). In this way, Nehm and colleagues could examine how well STAS and SIDE identified the magnitude of construct coverage in students' responses (that is, the degree to which students constructed accurate and complete scientific explanations). Both STAS and SIDE were found to successfully identify the number and diversity of Core Concepts in students' written explanations relative to expert human raters. Additionally, STAS and SIDE were able to detect Core Concepts in items that differed in a variety of "surface features" or cover stories (same concept presented using different organisms such as bacteria, cheetahs, roses, and salamanders; Ha and Nehm, 2011; Nehm and Haertig, 2011). The OSU group's current work builds upon the success of Core Concept analyses and pursues the automated detection of naive ideas and misconceptions (Nehm and Haertig, 2011).

Genetics

Jenny Knight and Michelle Smith presented the assessment development process and analysis of student responses to the Genetics Concept Assessment (GCA; Smith *et al.*, 2008). Initially, they developed the 25 multiple-choice question GCA by establishing a set of learning goals that characterized the content students are required to learn in a typical genetics course, with an emphasis on larger concepts rather than details. Question stems were constructed to address these concepts; distracters were written using student ideas, using student language where possible. To examine response validity, more than 30 students were interviewed in the process of writing and rewriting distracters, such that ultimately each distracter was chosen and explained by at least 5 students. To help establish content validity, faculty reviewed the questions to determine whether they addressed specified concepts and to ensure that each question had exactly one correct answer. Finally, the questions were administered to more than 600 students at three institutions. Item discrimination and difficulty were calculated for each question, both pre- and postinstruction, to represent the range of difficulty of the items and to measure the gains that students made over the semester. The instrument was also shown to be reliable using the test/retest measure from two consecutive semesters of students at the same institution.

Since its development, the GCA has been administered at more than 250 institutions around the world. In addition, Knight and Smith continue to follow student responses on the GCA, investigating persistent misunderstandings: answers that are chosen by students both on the pre- and postinstruction administration of the GCA. This work has helped identify at least three persistent problem areas in student understanding: the nature and consequences of mutations, DNA content of cells, and allele representation on chromosomes undergoing meiosis and mitosis. Knight and Smith are exploring the use of constructed-response items in these problem areas. They are collecting data from students at MSU, the University of Colorado, Boulder, and the University of Washington, and will be analyzing the data using lexical analysis software. In addition to potentially uncovering additional ways that students think about these concepts, the analysis will allow Knight and Smith to measure whether the GCA multiple-choice items effectively capture student ideas on these topics.

Introductory Biology

Tammy Long provided an overview of a recent restructure of an introductory biology course (NSF DUE 0736928) in which systems modeling was implemented as both a pedagogical tool and an alternative assessment of student understanding. An explicit goal of the reformed course was that students would be able to communicate their understanding of biological principles through diverse means, including both verbal (text-based) explanations and visual representations (e.g., models). In a related project (NSF DRL 0910278), Long and colleagues are exploring the potential of students' self-constructed conceptual models to reveal patterns in student thinking, particularly in the context of learning about complex biological systems. They are applying principles of Structure-Behavior-Function (SBF) theory (Goel *et al.*, 1996), adapted from artificial intelligence, to inform both the model-based instructional design and the metrics for comparing and evaluating students' models. Briefly, SBF theory deconstructs systems into fundamental components: *structures* are the components or elements of the system, *behaviors* refer to the processes and mechanisms that relate structures to one another within the system, and *function* is the purpose or role of the system.

Long described the application of STAS as a tool for categorizing students' language as represented in their models. Specifically, Long and colleagues are using analyses of lexical patterns represented in students' models to: 1) determine whether students' self-generated models are reflective of understanding as represented in more traditional assessments of equivalent concepts (e.g., essays, diagnostic questions, concept inventory items, etc.); 2) identify elements within models that students appear to misunderstand most frequently; and 3) provide insight on how students use models to organize, explain, and revise their thinking over time in the context of learning complex biological problems.

Geosciences

Julie Libarkin presented a review of her work on the Geoscience Concept Inventory (GCI; NSF DUE 0127765) and GCI WebCenter (NSF DUE 0717790; Ward *et al.*, 2010; Libarkin *et al.*, 2011). The GCI is used internationally to evaluate learning in entry-level geoscience courses and has been shown in ongoing work to correlate strongly with individual expertise in geosciences (NSF DRL 0815930). The role of Rasch analysis in validation and use of the GCI was discussed and led to the conceptual model underlying the GCI WebCenter. Libarkin's presentation also included a discussion of documented rules for multiple-choice item development and the interesting relationship between constructed-response items

and multiple-response items. The importance of aligning constructed-response and multiple-choice data sets is clear and an important mechanism for validation of concept inventory questions. Libarkin encouraged revision of existing multiple-choice items in response to student conceptual data derived from constructed-response questions.

As part of the GCI WebCenter project, Libarkin's lab is compiling a comprehensive database of student conceptions about Earth systems; thus far, 813 alternative conceptions have been compiled. This compilation is a review of existing literature and documents of exact student responses to constructed-response questions. Libarkin's group will attempt to use text analysis to analyze these data. Many papers provide 10–20 individual student responses to specific questions, offering an opportunity to analyze a small set of data for a large number of questions. This is a different approach to text analysis compared with the other research groups at the meeting, which analyze large numbers of student responses to a few questions. Libarkin's group hopes to determine whether text analysis is a viable tool for identifying alternative conceptions. If successful, these data can be used in the ongoing community-based expansion of the GCI.

Statistics

Jennifer Kaplan presented her research on lexical ambiguity in statistics. Domain-specific words that are similar to commonly used English words are said to have lexical ambiguity (Barwell, 2005). One example that is of particular importance to the other STEM disciplines is the word *random*, but there are countless others, such as *association*, *correlation*, *bias*, and *skew*. The goals of the lexical ambiguity project are to: 1) highlight specific words and document obstacles to students' comprehension that are associated with misunderstandings of those words; 2) design and implement an intervention to investigate whether and how the explicit examination of the lexical ambiguity of certain words during instruction promotes deeper understanding of statistics; and 3) assess the intervention on student learning outcomes. To meet the first goal, the research team collected data from more than 900 students at three universities using a pre- and posttest design. The data consist of student-generated sentences and definitions for the target words. At the beginning of a statistics course, students wrote one sentence and a definition for each word. At the end of the course, they provided two sentences and definitions, one for the everyday meaning and one for the statistical meaning of the target word. The research team selected a random sample of 100 responses for each word and used these to create categories for the definitions given by students. Although the interrater reliability for the hand scoring is quite high, hand scoring is both time-consuming and only provides a count of responses that fit each category. The statistics research group is interested in using software to provide efficiency in analysis and to uncover connections between concepts underlying the target words.

RESEARCH FOCI

Group discussions at the meeting focused on a number of questions and ideas that could inform research on student conceptions as well as automated analysis techniques. Major questions included those discussed below.

Are constructed-response items always needed to uncover student thinking?

A foundational tenet of our work is that writing is a powerful way to uncover student thinking. There may be multiple-choice items, however, that do a sufficient job of capturing student misconceptions, for example, due to the way these items are developed (see *Introduction* above and the Genetics discussion by Smith and Knight). A comparison of item-types (constructed-response, multiple choice, multiple true/false) with interviews as a standard for comparison will be useful to define the limitations of each type of item. It may be that only particular misconceptions or scientific ideas in particular subject domains cannot be validly diagnosed with multiple-choice concept inventories. Thus, it remains to be determined under which conditions lexical analysis will meaningfully enhance our understanding of student thinking about STEM ideas.

Are lexical analysis protocols generalizable?

To determine the depth of understanding a student has about a given scientific concept, ideally the concept should be tested across surface features or "cover stories." Given this, an important question is how effective lexical analysis will be at detecting student ideas across isomorphic items designed to test for the same idea (e.g., natural selection). Will term libraries and lexical analysis rules developed for evolution assessment items about bacteria, for example, function as effectively for items prompting equivalent responses about other organisms (Nehm and Ha, 2011)? Preliminary results using both STAS and SIDE suggest that these methods are robust to minor changes in item surface features, at least in the case of evolution and natural selection (Ha and Nehm, 2011; Nehm and Haertig, 2011). Further work in other content areas is needed to determine how generalizable this conclusion may be.

What are the relative strengths and weaknesses of different automated analysis techniques?

As described above (see Tools Under Investigation), STAS and SIDE use different approaches for automated text analysis. Whereas STAS extracts lexical tokens that can be used to create categories and rules using an analysis model similar to open coding in grounded theory (Strauss and Corbin, 1998), SIDE must be provided with a previously scored training set of data in order to develop its machine learning scoring model. Currently, it appears that SIDE offers a major time advantage when one has a robust scoring rubric that has been applied to large sets of student data. STAS, however, offers the advantage of discovering novel ideas by exploring students' use of language. Thus, each program may offer advantages and disadvantages at different stages of lexical analysis research. STAS may be most useful in exploratory lexical analysis work, whereas SIDE may be most effective at later stages involving confirmatory studies. Further work on these issues is needed so that these new technological tools may be used effectively and efficiently in STEM education.

How well do these techniques predict expert scoring?

A critical benchmark for the use of automated analysis is that interrater reliability measures of computer–human scoring approach those of human–human scoring. Initial

findings suggest that different automated techniques have different computer–human interrater reliabilities. In general, however, with sufficient amounts of data (i.e., student responses), automated analysis can approach human–human interrater reliability measures (Haudek *et al.*, 2009; Nehm and Haertig, 2011). Investigation into the interactions among different statistical functions, scoring rubrics, numbers, and types of student responses on prediction accuracy should continue.

How can text analysis inform rubric creation?

Prediction of expert scoring is predicated on the existence of a valid rubric, making the development and refinement of scoring rubrics (either analytic or holistic) a key consideration for automated analysis. Analytic rubrics are used to identify the presence or absence of particular idea(s) in student writing. For example, does the student discuss the causes or presence of variation when writing about evolution? Holistic rubrics, on the other hand, judge how students construct a response as a whole (e.g., good, acceptable, poor).

The iterative nature inherent in the development and use of text analysis as part of an automated analysis supports a more grounded-theory approach of rubric creation, in that the rubric emerges from student writing as opposed to solely from expert thinking. Exploring the use of lexical analysis to improve the efficiency or speed up the rubric creation and refinement process is an area for further research.

How can linguistics enhance lexical analysis research in STEM fields?

Alan Munn and Olga Eremina, linguists in the MSU group, have begun to investigate the possibility that more nuanced linguistic information might be helpful in developing rules used in categories within STAS. It is already clear that certain linguistic aspects of questions and answers can make extraction of categories more difficult. For example, the presence of negation in part of an answer can lead to false categorization. Comparative terms such as *more than/less than* can also lead to difficulties. Adjustments to questions (for example, avoiding contrast and compare type questions and instead splitting them into two parts) can also help in this respect. The researchers are investigating whether it is possible to add limited linguistic structural information to the rules to ameliorate these problems. It is clear that linguistics may offer major insights into student conceptions in STEM fields, and further work in this area is needed.

Data collection plan

To continue to identify key concepts targeted by the various instruments and prevalent student alternative conceptions in the STEM disciplines, the meeting participants identified items from extant multiple-choice concept inventory assessments, such as the GCA and the Comprehensive Assessment of Outcomes from a First Course in Statistics (CAOS) to adapt to constructed-response formats. Some of the items were converted before the end of the meeting, but for other items, a timeline and plan for conversion were established. In addition, the participants established a plan for large-scale data collection across courses and institutions. Finally, they agreed to share scientific term libraries and scoring models for use with lexical analysis software that are necessary to automate the identification of key scientific concepts from students' written responses to the inventories.

CONCLUSION

By the end of the meeting, participants had done the following:

- Reviewed how to use STAS and SIDE for text analysis;
- Reviewed the work of the cooperating projects to identify synergies and challenges among projects; and
- Agreed on areas of research foci and developed an initial data collection plan and dissemination of resources to address the research questions.

Future work of AACR will focus on developing and refining lexical resources and scoring models, as well as guidelines for the use of different techniques. An important goal is to make these resources available to interested users without making software training necessary. One possible approach for accomplishing this goal would be to develop a web portal where users could upload their own sets of student responses and receive formative feedback in near real-time. This type of timely, powerful feedback would further the development of reformed science teaching.

A long-term goal of AACR is to build a network of interested STEM education researchers who are willing to share assessment items, lexical resources, and statistical scoring models in order to continue to improve assessment quality in multiple disciplines. We plan to pilot items and collect data at a variety of institutions, in hope of making the items, resources, and statistical models more generalizable. Potential collaborators have the opportunity to do the following:

- Pilot constructed-response items at their institution and collect data to aid our analyses;
- Develop, evaluate, and/or apply scoring rubrics for existing items;
- Suggest other concepts, inventories, or questions that might reveal students' reasoning patterns; and
- Join online discussions about this research, student conceptions, and/or lexical analysis techniques.

If you are interested in extending this work to your college or university, becoming a participant in this network, or learning about automated analysis of constructed-responses, visit our AACR group website at http://aacr.crcstl.msu.edu or contact the corresponding author.

ACKNOWLEDGMENTS

Funding for the described research studies was provided by grants from the NSF (DUE 1022653, DUE 07236952, REESE 0909999, DUE 0736928, DRL 0910278, DUE 0127765, DUE 0717790, and DRL 0815930). 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.

REFERENCES

American Association for the Advancement of Science (2009). Vision and Change: A Call to Action, Washington, DC, 11.

Barwell R (2005). Ambiguity in the mathematics classroom. Lang Educ *19*, 117–125.

Bennett RE, Ward WC (ed.) (1993). Construction Versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment, Hillsdale, NJ: L. Erlbaum Associates.

Birenbaum M, Tatsouka KK (1987). Open-ended versus multiple-choice response formats: it does make a difference for diagnostic purposes. Appl Psychol Meas *11*, 329–341.

Bridgeman B (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. J Educ Meas *29*, 253–271.

D'Avanzo C (2008). Biology concept inventories: overview, status, and next steps. Bioscience *58*, 1079–1085.

Deane P (2006). Strategies for evidence identification through linguistic assessment of textual responses. In: Automated Scoring of Complex Tasks in Computer-Based Testing, ed. DM Williamson, Mahwah, NJ: L. Erlbaum Associates, 313–372.

Duit R (2009). Students' and Teachers' Conceptions and Science Education. www.ipn.uni-kiel.de/aktuell/stcse/ (accessed 14 March 2011).

Gess-Newsome J, Johnston A, Woodbury S (2003). Educational reform, personal practical theories, and dissatisfaction: the anatomy of change in college science teaching. Am Educ Res J *40*, 731–767.

Goel AK, de Silva Garza AG, Grue N, Murdock JW, Recker M, Govindaraj T (1996). Towards design learning environments. I. Exploring how devices work, Third International Conference on Intelligent Tutoring Systems, Montreal, Canada: Springer.

Ha M, Nehm R (2011). Comparative Efficacy of Two Computer-Assisted Scoring Tools for Evolution Assessment, Orlando, FL: National Association for Research in Science Teaching.

Haudek K, Moscarella RA, Urban-Lurain M, Merrill J, Sweeder R, Richmond G (2009). Using Lexical Analysis Software to Understand Student Knowledge Transfer between Chemistry and Biology, National Association of Research in Science Teaching Annual Conference, Garden Grove, CA.

Kardash CA, Wallace ML (2001). The perceptions of science classes survey: what undergraduate science reform efforts really need to address. J Educ Psychol *93*, 199–210.

Knight JK (2010). Biology concept assessment tools: design and use. Microbiol Aust *31*, 5–8.

Kuechler WL, Simkin MG (2003). How well do multiple choice tests evaluate student understanding in computer programming classes? J Inf Syst Educ *14*, 389–399.

Kuechler WL, Simkin MG (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. Decis Sci J Innovative Educ *8*, 55–73.

Libarkin JC (2008). Concept inventories in higher education science, Council Promising Practices in Undergraduate STEM Education Workshop 2, Washington, DC: National Research Council.

Libarkin JC, Ward EMG, Anderson SW, Kortemeyer G, Raeburn S (2011). Revisiting the Geoscience Concept Inventory: A Call to the Community, GSA Today, *in press*.

Mayfield E, Rose CP (2010a). An Interactive Tool for Supporting Error Analysis for Text Mining, Human Language Technologies. Los Angeles, CA: North American Association for Computational Linguistics.

Mayfield E, Rose CP (2010b). SIDE: The Summarization IDE. www.cs.cmu.edu/~cprose/SIDE.html (accessed 14 March 2011).

Moscarella RA, Urban-Lurain M, Merritt B, Long T, Richmond G, Merrill J, Parker J, Patterson R, Wilson C (2008). Understanding Undergraduate Students' Conceptions in Science: Using Lexical Analysis Software to Analyze Students' Constructed Responses in Biology, National Association for Research in Science Teaching 2008 Annual International Conference, Baltimore, MD.

National Research Council (1999). Transforming Undergraduate Education in Science, Mathematics, Engineering, and Technology, Washington, DC: National Academy Press.

National Science Foundation (1996). Shaping the Future: New Expectations for Undergraduate Education in Science, Mathematics, Engineering and Technology, Washington, DC: National Science Foundation, 76.

Nehm RH, Ha M (2011). Item feature effects in evolution assessment. J Res Sci Teach *48*, 237–256.

Nehm RH, Ha M, Mayfield E (2011). Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. J Sci Educ Technol, *in press*.

Nehm R, Ha M, Rector M, Opfer J, Perrin L, Ridgway J, Mollohan K (2010). Scoring Guide for the Open Response Instrument (ORI) and Evolutionary Gain and Loss Test (EGALT). 40. http://evolutionassessment.org (accessed 14 March 2011).

Nehm RH, Schonfeld IS (2008). Measuring knowledge of natural selection: a comparison of the CINS, an open-response instrument, and an oral interview. J Res Sci Teach *45*, 1131–1160.

Nehm RH, Schonfeld IS (2010). The future of natural selection knowledge measurement: a reply to Anderson *et al.* (2010). J Res Sci Teach *47*, 358–362.

Pellegrino JW, Chudowsky N, Glaser R (2001). Knowing What Students Know: The Science and Design of Educational Assessment, Washington, DC: National Academy Press.

Ruiz-Primo MA, Shavelson RJ, Hamilton L, Klein S (2002). On the evaluation of systemic science education reform: searching for instructional sensitivity. J Res Sci Teach *39*, 369–393.

Seymour E (2002). Tracking the processes of change in U.S. undergraduate education in science, mathematics, engineering, and technology. Sci Educ *86*, 79–105.

Seymour E, Hewitt NM (1997). Talking About Leaving: Why Undergraduates Leave the Sciences, Boulder, CO: Westview Press.

Smith MK, Wood WB, Knight JK (2008). The genetics concept assessment: a new concept inventory for gauging student understanding of genetics. CBE Life Sci Educ *7*, 422–430.

SPSS (2009). SPSS Text Analysis for Surveys 3.0 User's Guide, Chicago, IL.

Strauss A, Corbin J (1998). Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, Thousand Oaks, CA: Sage Publications.

Tobias S (1990). They're Not Dumb, They're Different: Stalking the Second Tier, 94th ed., Tucson, AZ: Research Corporation.

Urban-Lurain M, Moscarella RA, Haudek KC, Giese E, Sibley DF, Merrill JE (2009). Beyond Multiple Choice Exams: Using Computerized Lexical Analysis to Understand Students' Conceptual Reasoning in STEM Disciplines, San Antonio, TX: Frontiers in Education, ASEE/IEEE.

Urban-Lurain M, Moscarella RA, Haudek KC, Giese E, Merrill JE, Sibley DF (2010). Insight into Student Thinking in STEM: Lessons Learned from Lexical Analysis of Student Writing, National Association for Research in Science Teaching Annual International Conference, Philadelphia, PA.

Ward EMG, Libarkin JC *et al.* (2010). The Geoscience Concept Inventory WebCenter provides new means for student assessment. eLearning Papers, 1–14. http://elearningpapers.eu/en/download/file/fid/19521 (accessed 12 May 2011).

Wilson CD, Anderson CW, Heidemann M, Merrill JE, Merritt BW, Richmond G, Sibley DF, Parker JM (2006). Assessing students' ability to trace matter in dynamic systems in cell biology. CBE Life Sci Educ *5*, 323–331.