An Iterative Approach To Developing, Refining And Validating Machine-Scored Constructed Response Assessments

Luanna Prevost, University of South Florida, PI [1]
Andrea Bierema, Michigan State University, Post-doctoral Researcher
Jennifer Kaplan, University of Georgia, PI
Jennifer Knight, University of Colorado Boulder, PI
Paula Lemons, University of Georgia, PI
Carl Lira, Michigan State University, co-PI
John Merrill, Michigan State University, PI
Rosa Moscarella, Michigan State University, Post-doctoral Researcher
Ross Nehm, SUNY-Stony Brook, PI
Michelle Smith, University of Maine, PI
Mary Anne Sydlik, Western Michigan University, PI – Project Evaluation
Mark Urban-Lurain, Michigan State University, Overall Project PI

Need: The Automated Analysis of Constructed Response (AACR) project seeks to develop a community of faculty who use evidence based practices to improve instruction by presenting faculty with novel assessment platforms for written assessment. Written assessments provide faculty in-depth evidence of student learning as they allow faculty to gather student understanding in students' own words. However, written assessments are used infrequently in undergraduate biology courses, particularly courses with high student enrollment, because of the time and effort necessary to read and provide feedback.

Goals: The primary AACR goals are to (1) provide the means for faculty to gather evidence on student learning using formative written assessments and computerized analysis tools and (2) facilitate widespread use of these written assessments. The goal of the question development group within AACR is twofold 1) to develop a suite of formative written assessments in biology, chemistry and statistics that uncover student conceptual difficulties and 2) develop text analysis and machine learning models that automatically analyze student writing, providing faculty with immediate feedback.

Approach: Our approach is to use pre-existing concept inventories, the science education literature, and interviews with faculty to identify areas of biology, chemistry and statistics where students have persistent conceptual difficulties. We then develop questions that target these conceptual difficulties. Questions are refined based on input from faculty and data from student interviews. Questions are piloted and revised, so answers can be analyzed by computers. After we have developed a question, we use two approaches to analyze student answers: text analysis and machine learning.

---

[1] Co-authors are listed in alphabetic order after the first author

Both methods identify and extract words and phrases from student writing that are used to build models of human scoring. The models classify the key concepts or correctness of a response and do so in high agreement with human scoring. Finally, models are piloted in the classrooms of members of our faculty learning communities at six different institutions.

Outcomes: We have developed 53 questions in biology, chemistry, chemical engineering, and statistics. We have collected responses from 7854 students and provided 123 reports to faculty. We have also improved our process of question development through the use of clustering and multinomial logistic regression analyses. We also have created more interactive and user friendly feedback reports for faculty.

Broader Impacts: Currently 31 faculty are using AACR assessments and participating in our faculty learning communities. We have also recruited 12 new faculty members across our institutions to join our FLCs and use AACR assessments and resources. Additionally, we have expanded to collaborate with faculty in physics at Michigan State University and Stony Brook University and statistics at Grand Valley State University. To date, we have disseminated our findings though 37 presentations and 12 journal articles. AACR products are currently available to faculty via 2 websites.

Introduction

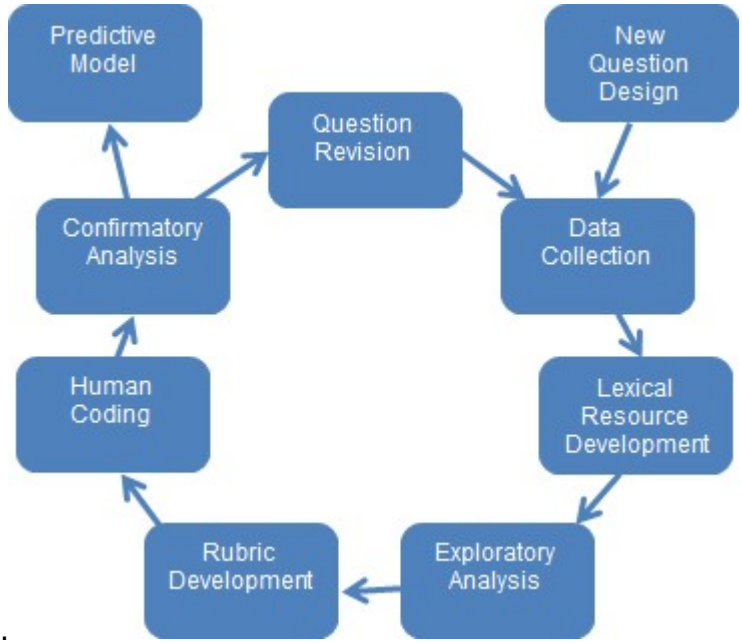
Recent evaluations of undergraduate STEM education have identified a need for appropriate assessment tools [1,2]. In biology education for example, there has been an increased development of concept inventories in response to these reports, yet there is still a need for formative assessments, assessments that can be used during the learning process. Formative assessments allow researchers to identify critical stages during learning, and allow instructors to assess student thinking and give students appropriate feedback [3].

As students progress from novice to more expert thinking about science, their mental models change. During the learning process students tend to hold mixed mental models; that is, they have both scientific and informal or incorrect ideas about scientific concepts [4,5]. Therefore student understanding at this stage contains a mix of correct and incorrect ideas. Forced-selection options such as multiple choice assessment, often limit students to selecting one answer and may not reflect the variety of ideas that students hold. Written responses also allow students to demonstrate their thinking in their own words, thus giving them the opportunity to share both correct and incorrect ideas that they hold.

Despite this advantage of written assessments, they are not commonly used in the classroom because of the time and effort required to read them and provide feedback.

This is particularly true for large undergraduate science classrooms. Our objectives are 1) to develop a suite of formative written assessments in biology, chemistry and statistics that uncover student conceptual difficulties and 2) develop text analysis and machine learning models that automatically analyze student writing, providing faculty with immediate feedback.

We are harnessing advancements in text analysis and machine learning technologies to analyze student writing about science concepts. We use an iterative approach to develop scoring models for a variety of questions in biology, chemistry, chemical engineering and statistics. This approach, the Question Development Cycle, is represented in Figure 1. Below we describe the development of the written assessments, our text analysis and machine learning approaches, and key outcomes to date



.

Figure 1. The Question Development Cycle represents the iterative process of question and predictive model development.

Development of question prompts for written assessments

In our question development stage, we identify foundational or challenging topics based on findings from concept inventories, the science education literature, and input from instructors. For example, we identified genetics concepts that were both foundational and challenging students using the Genetics Concept Assessment [6] as well as literature on persistent difficulties students have [7]. We create open ended questions based on these concepts. Questions are reviewed by faculty for content and clarity. We also use semi

structured interviews with students to determine whether students interpret questions in the manner intended. Once questions meet these criteria, we administer the question to students via their online homework management system. Students respond to the questions online and their responses are downloaded for human coding, text analysis and machine scoring.

Lexical Resource Development for Text Analysis

The AACR project uses two approaches for automated-analysis of student writing, text analysis and machine learning. For text analysis, we use SPSS Text Analysis for Surveys and SPSS Modeler to identify and extract words and phrases in student writing. We developed libraries- suites of relevant terms in biology, chemistry, and statistics- that the software can use to recognize the words in student writing. These words and phrases are assigned to categories. Each category represents one homogeneous idea. Each response can be assigned to zero to multiple categories, based on the words in the response. The categories are used for subsequent statistical analysis such as clustering (exploratory analysis) or regression analysis (confirmatory analysis).

Rubric Development and Human Coding

Student responses are also scored by disciplinary experts (faculty and postdocs) using holistic or analytic rubrics. Holistic rubrics assess the overall correctness of student responses. For example, responses may be ranked correct, partially correct or incorrect based on their content. Alternatively analytic rubrics identify the presence or absence of a concept. Thus, while using an analytic rubric, a response may be coded as having one idea present and another idea absent. Two or more experts are trained on the rubric until agreement is acceptable [8] (Cohen's Kappa > 0.8).

We have also used the results from cluster analyses of the text analysis categories to inform our rubric. Cluster analyses identify responses with similar themes. We have used these themes as criteria in our rubric development.

Confirmatory analysis

Confirmatory analysis using text analysis requires an additional step, typically regression analysis. We have used logistic regression analyses in which our text analysis categories are the binary independent variables and our human coding is our dependent variable.

An alternative approach to confirmatory automated analysis is machine learning. Machine learning uses algorithms to identify patterns in student writing that are aligned with human scoring. These patterns are used to build scoring models that can automatically score

future responses. We used the LightSide program developed by Carnegie Mellon to build predictive models of human scoring. Machine learning is similar to text analysis in its extraction of words and phrases (n-grams) from student responses. However, unlike text analysis, machine learning does not require category formation.

In our confirmatory analysis, our goal is to obtain human-computer agreement similar to human-human agreement.

Outcomes

Biology

To date we have developed a suite of biology questions that explore a range of topics including genetics, ecology, evolution and photosynthesis. Using these questions and scoring models, we have identified student conceptual difficulties in biology [9], explored how question structure elicits student understanding [10], and explored the use of analytic and holistic rubric for automated scoring [11]. Our questions and scoring models are being used by 31 biology faculty at six institutions. These faculty administer the questions and receive feedback reports detailing their class outcomes. At each institution, the faculty form faculty learning communities to discuss the implementation of the questions, their student outcomes and how these can be used to inform their teaching. For additional information on the organization of and outcomes from these faculty learning communities at these institutions see the papers *A Community of Enhanced Assessment Facilitates Reformed Teaching* and *Expanding a National Network for Automated Analysis of Constructed Response Assessments to Reveal Student Thinking in STEM.*

Chemistry

Within chemistry, four questions and their scoring rubrics have been developed.

Statistics

Research in statistics education has focused on creation of analytic models of students' interpretation of histograms. Two questions and scoring models have been developed, along with faculty feedback reports for these questions.

Chemical engineering

Our chemical engineering group has developed six questions about thermodynamics. Questions have been revised based on input from faculty at four institutions and from student interviews. Text analysis libraries and categories are currently under development for this suite of questions.

1 American Association for the Advancement of Science, “Vision and change in undergraduate biology education: A call to action,” 2011.
2 N. R. C. NRC, *BIO2010: Transforming Undergraduate Education for Future Research Biologists*. Washington, D.C.: The National Academies Press, 2003.
3 R. H. Nehm and H. Haertig, “Human vs. Computer Diagnosis of Students’ Natural Selection Knowledge: Testing the Efficacy of Text Analytic Software,” *Journal of science education and technology*, vol. 21, no. 1, pp. 56–73, 2012.
4 M. T. H. Chi, P. J. Feltovich, and R. Glaser, “Categorization and representation of physics problems by experts and novices,” *Cognitive science*, vol. 5, no. 2, pp. 121–152, 1981.
5 J. E. Opfer, R. H. Nehm, and M. Ha, “Cognitive foundations for science assessment design: Knowing what students know about evolution,” *J. Res. Sci. Teach.*, vol. 49, no. 6, pp. 744–777, Aug. 2012.
6 M. K. Smith, W. B. Wood, and J. K. Knight, “The Genetics Concept Assessment: A New Concept Inventory for Gauging Student Understanding of Genetics,” *CBE Life Sci Educ*, vol. 7, no. 4, pp. 422–430, 2008.
7 M. K. Smith and J. K. Knight, “Using the Genetics Concept Assessment to document persistent conceptual difficulties in undergraduate genetics courses,” *Genetics*, vol. 191, no. 1, pp. 21–32, 2012.
8 J. R. Landis and G. G. Koch, “An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers,” *Biometrics*, pp. 363–374, 1977.

9 Prevost, L. 2015. Assessing student ecological understanding using text analysis and machine learning. Ecological Society of America, 100th Annual Meeting, Baltimore, MA

10 M. Weston, K. C. Haudek, L. Prevost, M. Urban-Lurain, and J. Merrill, “Examining the Impact of Question Surface Features on Students’ Answers to Constructed-Response Questions on Photosynthesis,” CBE Life Sci Educ, vol. 14, no. 2, p. ar19, Jun. 2015.

11 Moscarella, R.A., J.R., Stoltzfus, J. Merrill, K.C. Haudek, and M.Urban-Lurain.2015. Creation of scoring rubrics assisted by Computerized Text Analysis. Create for STEM Mini Conference. Michigan State University.

Biographical Information
Author Biographies: co-authors are listed in alphabetic order after the first author

Luanna Prevost, University of South Florida, PI
Dr. Prevost is an Assistant Professor in the Department of Integrative Biology at the University of South Florida. She is interested in exploring undergraduate student thinking in biology. Her research employs written assessment, automated analysis tools, and game design to explore student understanding of biology. She is also interested in how these approaches can be used to foster active learning environments in undergraduate biology classrooms.

Andrea Bierema, Michigan State University, AACR Research Associate
Dr. Bierema is a research associate in the Center for Engineering Education Research, College of Engineering, Michigan State University. Her training is in both biological and science education research. Research interests include undergraduate student conceptions and sense-making of fundamental biological concepts, the portrayal of fundamental biological concepts in curriculum, and student group discourse.

Jennifer J. Kaplan, UGA PI
Dr. Kaplan is an Associate Professor in the Department of Statistics at the University of Georgia. Her research interests are undergraduate student learning in statistics, the pedagogical and content knowledge needs of instructors of statistics, particularly of GTAs, and the types of professional development that will lead to gains in knowledge for both instructors and students.

Jennifer Knight, University of Colorado Boulder PI.
Dr. Knight is an Associate Professor in the Department of Molecular Cellular and Developmental Biology (MCDB). She has a Ph.D. in Neuroscience, and previously worked as a developmental biologist and geneticist. Her research now focuses on developing and using active learning materials and concept assessments, and studying the factors that influence students' in-class discussion. Dr. Knight coordinated the MCDB Science Education Initiative for 7 years, and is actively involved in CU's Center for STEM Learning, as well as other national organizations devoted to science education research.

Paula P. Lemons, UGA PI and Co-PI. http://sites.bmb.uga.edu/lemonslab/
Dr. Lemons is an Associate Professor in the Department of Biochemistry and Molecular Biology at the University of Georgia. Her research interests are in faculty development, with a focus on the process by which faculty change their teaching beliefs and practices while engaged in activities like faculty learning communities. She also studies problem solving among biology undergraduates, focusing on students' application of threshold concepts in biochemistry to problems involving visual representations.

Carl T. Lira, Michigan State University, Co-PI,
Dr. Carl T. Lira, Associate Professor of Chemical Engineering at Michigan State University, integrates computer technology into classroom learning through assignments, interactive classroom activities, clickers. His participation on the project concerns the development of constructed response questions in engineering, initially focusing on energy concepts.

John E. Merrill, Michigan State University, PI and Co-PI.
Dr. Merrill is Associate Professor of Microbiology and Molecular Genetics (College of Osteopathic Medicine) and Director of the Biological Sciences Program (College of Natural Science). Primary research interests include assessment of student learning in foundational undergraduate biology courses. Previous work on concept inventory type assessment instruments led to an interest in finding better ways to explore student thinking about important biological concepts.

Rosa A. Moscarella, Michigan State University AACR Research Associate.
Dr. Moscarella is a Research Associate in the Center for Engineering Education Research, College of Engineering, Michigan State University. Formally trained as a biologist, she is interested in understanding students' learning obstacles in biology and more specifically in genetics. Her research focuses on three main aspects: 1) developing assessments and diagnostic tools that better reveal students' thinking in biology, 2) understanding the basis of students' learning difficulties and misconceptions in biology, and 3) designing a learning progression for college genetics.

Ross Nehm, Stony Brook University, PI, Co-PI
Dr. Nehm is Associate Professor of Ecology & Evolution and Associate Director of the Ph.D. Program in Science Education. He studies student thinking about biological concepts such as natural selection and evolution. Additional work has examined novice and expert reasoning strategies, psychometric evaluation of education instruments, science teacher belief revision and professional development, conceptual structuring of scientific understanding, and the comparative efficacy of educational innovations. Currently,

several projects are focusing on developing and evaluating machine-learning models for automated assessment of complex scientific practices, such as biological explanations.

Michelle Smith, UMaine PI, http://umaine.edu/center/directory/faculty-page/michelle-smith/
Dr. Smith is an Assistant Professor in the School of Biology and Ecology at the University of Maine and holds the C. Ann Merrifield Professorship in Life Sciences Education. Her research laboratory engages undergraduate and graduate students, postdocs, K-12 teachers, and university faculty in research on teaching and learning. Together they focus on: 1) developing tools to understand student conceptual difficulties and conduct classroom observations, 2) studying what aspects of peer discussion make it an effective learning tool, and 3) understanding what factors influence faculty members' decisions about teaching.

Mary Anne Sydlik, Western Michigan University. Dr. Sydlik is the Director of the Science and Math Program Improvement (SAMPI) Center, an outreach division of Mallinson Institute for Science Education. SAMPI specializes in evaluation, research, and technical assistance for higher education institutions and K-12 schools. She is the external evaluator for AACR III. Dr. Sydlik has been the lead external evaluator for a number of STEM and NSF-funded projects. Her interests are in adding to efforts to improve the educational experiences and outcomes of undergraduate STEM students.

Mark Urban-Lurain, AACR Project PI. (www.msu.edu/~urban)
Dr. Urban-Lurain is an Associate Professor and Acting Director of the Center for Engineering Education Research in the College of Engineering at Michigan State University. His research interests are in theories of cognition, how these theories inform the design of instruction, how we might best design instructional technology within those frameworks, and how the research and development of instructional technologies can inform our theories of cognition. He is also interested in preparing future STEM faculty for teaching, incorporating instructional technology as part of instructional design, and STEM education improvement and reform. Much of his research has focused on incorporating technology in the context of instructional design and using technology to provide assessments for formative feedback in the improvement of instruction.