EXPANDING A NATIONAL NETWORK FOR AUTOMATED ANALYSIS OF CONSTRUCTED RESPONSE ASSESSMENTS TO REVEAL STUDENT THINKING IN STEM

Mark Urban-Lurain[1], Melanie M. Cooper[2], Kevin C. Haudek[3], Jennifer J. Kaplan[4], Jennifer K. Knight[5], Paula P. Lemons[6], Carl T. Lira[7], John E. Merrill[8], Ross Nehm[9], Luanna B. Prevost[10], Michelle Kathleen Smith[11], Mary Anne Sydlik[12]

[1] Center for Engineering Education Research, Michigan State University
[2] Department of Chemistry, Michigan State University
[3] Department of Biochemistry and Molecular Biology, Michigan State University
[4] Department of Statistics, University of Georgia
[5] Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder
[6] Department of Biochemistry and Molecular Biology, University of Georgia
[7] Department of Chemical Engineering and Materials Science, Michigan State University
[8] Biological Sciences Program, Michigan State University
[9] Department of Ecology & Evolution, Stony Brook University New York
[10] Department of Integrative Biology, University of South Florida
[11] School of Biology and Ecology, University of Maine
[12] Science and Mathematics Program Improvement, Western Michigan University

Abstract

Improving STEM education requires valid and reliable instruments for providing insight into student thinking. Constructed response (CR) assessments reveal more about student thinking and the persistence of misconceptions than do multiple-choice questions, but require more analysis on the part of educators.

In the Automated Analysis of Constructed Response (AACR) Research Group (www.msu.edu/~aacr) we have developed constructed response versions of well-established conceptual assessment inventories and created computer automated analysis resources that predict human ratings of student writing about these topics in introductory STEM courses. The research uses a two-stage, feature-based approach to automated analysis of constructed response assessments. First, we design items to identify important disciplinary constructs based on prior research. The items are administered via online course management systems where students enter responses. We use lexical analysis software to extract key terms and scientific concepts from the students' writing. These terms and concepts are used as variables for statistical classification techniques to predict expert ratings of student responses. The inter-rater reliability (IRR) between automated predictions and expert human raters is as high as IRR between human experts.

We recently received another round of funding to extend our work to provide an online community where instructors may obtain scores and contribute to the library of items and resources necessary for their analyses. We provide an overview of the goals of the project and introduce the opportunities to participate in the development of a national network of faculty using these techniques.

Introduction

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].

Thus, constructed response (CR) assessments that capture students' explanatory models are needed to mitigate the constraints and reveal students' mixed models. CR assessments, for which students have to use their own language to demonstrate knowledge, are widely viewed as providing greater insight into student thinking than closed form (e.g., multiple-choice) assessments [9]. 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 [8, 10-15]. In our current work, we employ cutting-edge, lexical and computer analysis technology, focusing on the NSF DRL *Cycle of Research and Development* "hypothesize/clarify" and "design/develop/test" phases of the cycle. In this paper, we describe how we are moving to the "implement/study/improve" and "scale up/study effectiveness" phases of the cycle to enable widespread adoption of CR assessments by faculty nationwide.

Methodological Details of Our Approach

In this section, we provide an overview of our approach to developing, validating and implementing Automated Analysis of Constructed Response (AACR) assessments as background. The entire process is captured by the Question Development Cycle (QDC) shown in Figure 1. In general, we use linguistic feature-based methods [16] to extract linguistic features from students' writing [e.g., WordNet, see 17, 18] and then use those linguistic features as variables in statistical models that predict human raters' scores of the student's writing.

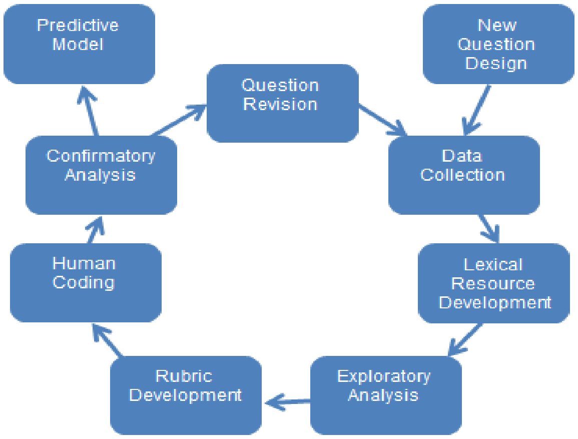


Figure 1: Question Development Cycle.

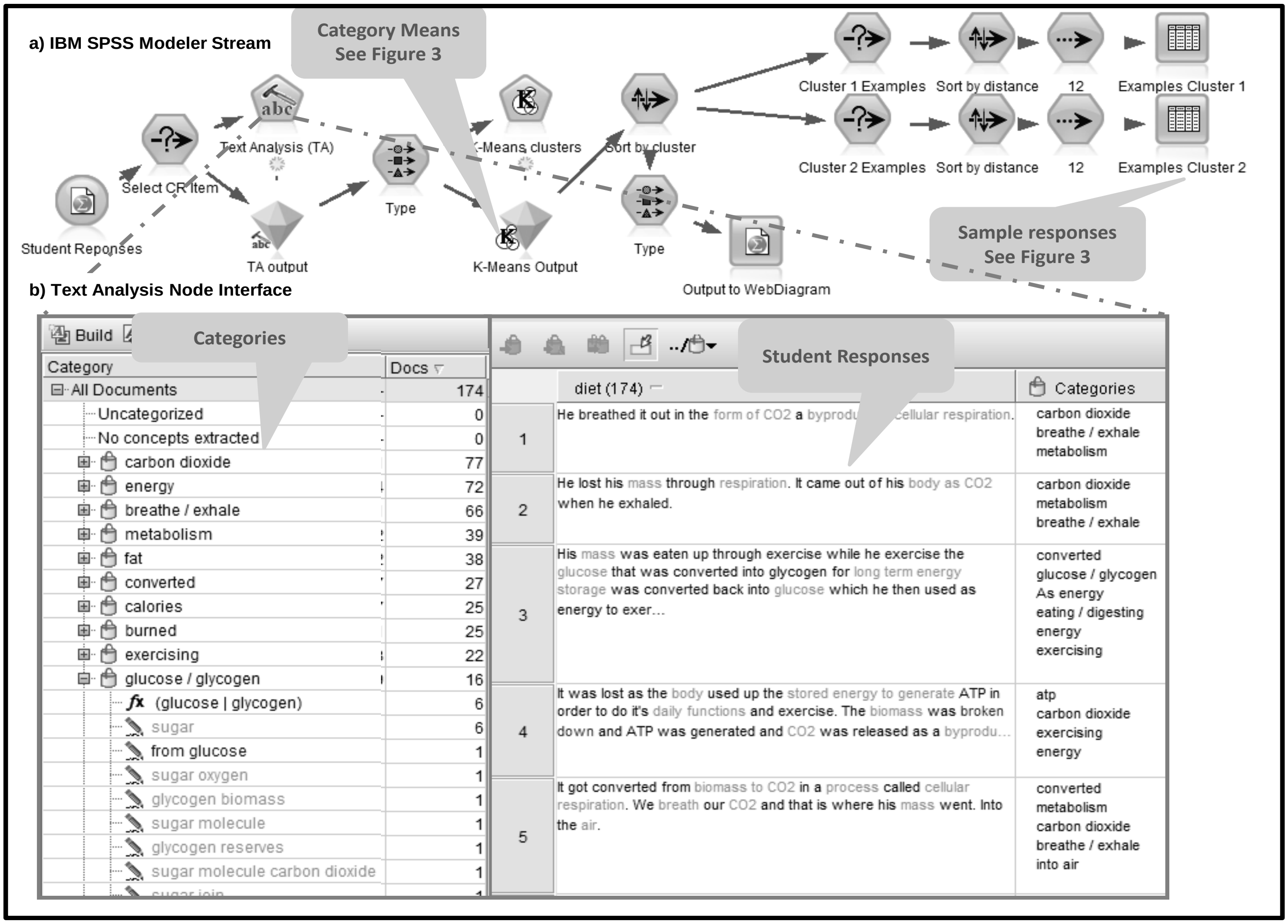
Developing AACR Questions to Assess Core Disciplinary Concepts

In the first stage of the QDC, we *Design New Questions* to measure student thinking about important disciplinary constructs. *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.

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 [19] for which we are developing AACR assessments [20-22]. 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.

Validating AACR Assessments through Lexical and Confirmatory Analysis

In this example, we describe how we use IBM SPSS Modeler [23] 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) and select 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.



Category	Docs
All Documents	174
Uncategorized	0
No concepts extracted	0
carbon dioxide	77
energy	72
breathe / exhale	66
metabolism	39
fat	38
converted	27
calories	25
burned	25
exercising	22
glucose / glycogen	16
(glucose | glycogen)	6
sugar	6
from glucose	1
sugar oxygen	1
glycogen biomass	1
sugar molecule	1
glycogen reserves	1
sugar molecule carbon dioxide	1
sugar join	1

	diet (174)	Categories
1	He breathed it out in the form of CO2 a byprodu cellular respiration.	carbon dioxide breathe / exhale metabolism
2	He lost his mass through respiration. It came out of his body as CO2 when he exhaled.	carbon dioxide metabolism breathe / exhale
3	His mass was eaten up through exercise while he exercise the glucose that was converted into glycogen for long term energy storage was converted back into glucose which he then used as energy to exer...	converted glucose / glycogen As energy eating / digesting energy exercising
4	It was lost as the body used up the stored energy to generate ATP in order to do it's daily functions and exercise. The biomass was broken down and ATP was generated and CO2 was released as a byprodu...	atp carbon dioxide exercising energy
5	It got converted from biomass to CO2 in a process called cellular respiration. We breath our CO2 and that is where his mass went. Into the air.	converted metabolism carbon dioxide breathe / exhale into air

Figure 2: 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?*

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 (grey 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 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, 11, 14]. 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.

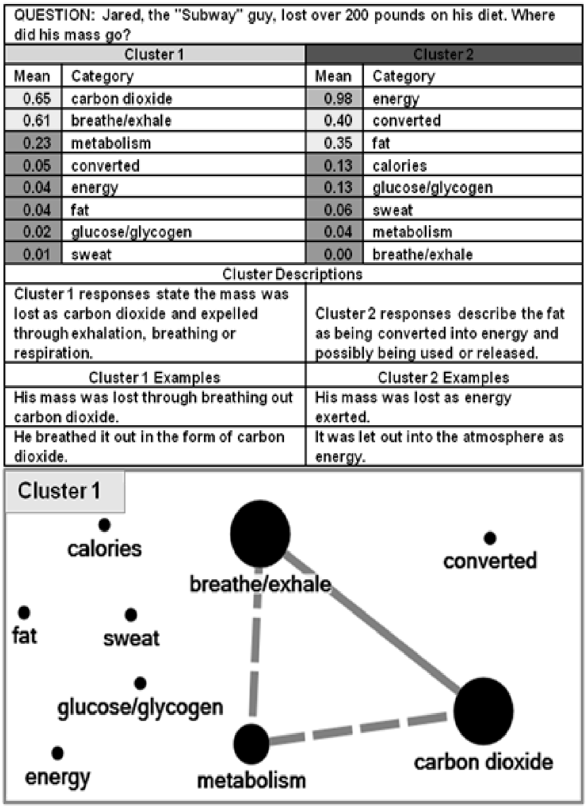
Pilot Study: Implementing AACR Questions for Just-in-Time Teaching

To enhance the infrastructure for education and improve undergraduate STEM education nationally, the tools we develop must be broadly available for faculty to use for Just-in-Time Teaching [JiTT, 24]. 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 of an introductory cells and molecules biology course for science majors at Michigan State University [25]. 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 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 frequency and 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 [20] as shown in the Cluster 2 example student answers.



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.	

Figure 3: 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 no link.

Pilot Study: Instructor Use of the Reports

We conducted focus groups with the four pilot study instructors throughout the semester to introduce them to the AACR assessments, explain the analyses and reports, and address difficulties they encountered. They discussed how the reports informed their awareness of students' thinking and how they used the reports to modify their instruction. Some of the seasoned instructors proceeded to create new instructional materials, such as sequences of clicker questions, to address these challenges. Others indicated that they would have preferred such materials to be provided for them. Although these instructors were previously aware of some of the concepts students found challenging, they pointed out that the written assessments provided insight as to why students struggled with these ideas. For example, in a question about genetics, one instructor noted that the reports revealed that many students thought that transcription and translation are the same process.

We learned valuable lessons from this pilot study about 1) how to improve the presentation and user-friendliness of reports; 2) how to improve the scheduling of the AACR assessments and to incentivize homework assignments; and 3) the need for professional development to support faculty use of these assessments. These lessons are reflected in the proposed activities.

Overall, our previous work shows that 1) it is possible to create scoring models that predict human scoring with IRR approaching that of well-trained expert raters across multiple topics; 2) AACR questions reveal the heterogeneity of student thinking that cannot be revealed by traditional multiple-choice items; and 3) we can capture, represent, and analyze this multidimensional information in a variety of ways that provide instructors with rich insights into student thinking. The speed with which we can now accomplish these goals means that instructors in large enrollment courses can assign AACR questions and a few days later, use the information to transform their teaching.

Current Project Scope

We have recently received five years of funding from NSF to expand our work nationally. This section describes the five project goals, how they interrelate, and how we are implementing them.

Goal: Create a community web portal for automated analysis of AACR assessments to expand and deepen collaborations between STEM education researchers and instructors

To facilitate building on-line communities of practice [26], it is critical that the software infrastructure be usable, reliable and robust, be maintainable and extensible, and be scalable as the communities grow. We will employ user-centered design to build on the technical and operational lessons we have learned to develop and test a **social collaboration community web portal**. User-centered design involves users throughout all stages of development in order to meet users' needs (http://usability.gov).

The functional structure of the proposed web portal is shown in Figure 4. We describe each of the components below.

Major **user interactions** are shown in the top row of boxes. The target portal users are faculty who teach courses in the disciplines for which we have questions and STEM education researchers who are investigating student writing and/or developing the resources to support the automated analyses. The *Web Server* provides the user interfaces for the various types of users of the system. We describe portal functionality in the context of basic *use cases* (http://www.usability.gov/methods/usecases.html) for the target users below.

Public information about the project will be accessible without an account or login. This information will include a background statement about the group, links to our publications, news, and events. Faculty who are interested in other resources will be able to create free accounts to access them.

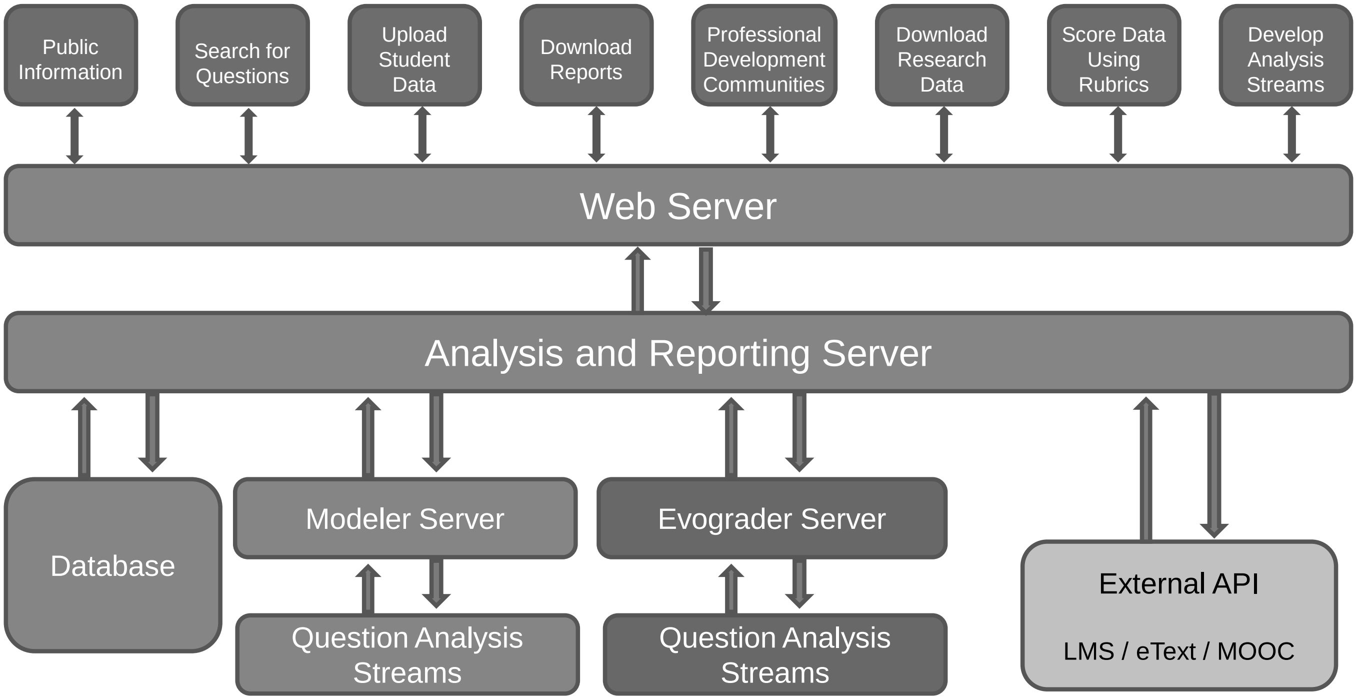


Figure 4: Community Web Portal Block Diagram.

Faculty will be able to *search for questions*. Associated with each question will be the concepts it tests, common student misconceptions about the topic, and technical reports about question development (i.e., numbers of students that have responded to the question; rubrics used for human scoring; and reliability and validity evidence).

After selecting questions, faculty will then use the LMS at their institutions to administer the questions and collect student responses. All LMS allow faculty to download student responses to questions in spreadsheet format. Faculty can then *upload student data* with their responses to the web server which passes them to the *analysis and reporting server* which stores the responses in the *database*, then invokes the appropriate *question analysis stream* for each question. Some questions (currently cellular biology, genetics and chemistry) are analyzed using IBM-SPSS Modeler streams created by our research team, so they are passed to the *modeler server* for analysis. Other questions are analyzed using machine learning tools in Evograder (currently natural selection and evolution), so they are passed to the *Evograder server* (http://evograder.org) for analysis. After analysis, the results are returned through the *analysis and reporting server*, to the *web server* where faculty can *download reports* with analyses of their students' writing.

The *web server* also provides a social collaboration environment for building *professional development communities*. For example, faculty may upload and download instructional materials that are aligned with each assessment. Discussion forums will allow interaction among faculty.

Researchers can *download research data*, anonymized student responses to questions, for research purposes. The *database* maintains anonymized student responses to questions along with metadata about the questions (e.g., class and institution in which they were administered) and human scoring for student responses. As more faculty use the web portal, the data repository will grow, supporting ongoing research and refinement of the analytic resources in the DRL *Cycle of Research and Development*.

Researchers can create scoring rubrics for questions and *score data using the rubrics*. The scores are stored in the *database* and can be used for confirmatory analysis on questions (see Figure 1). Researchers can also *develop analysis streams* for new or existing questions for automated analysis via the portal.

Since the primary goal of the portal is to support the transportability of the AACR research into practice among faculty, the only users we intend to support directly are faculty and researchers. There are, however, a wide range of possibilities for implementing AACR questions via on-line systems with which students directly interact and could receive immediate feedback. We will develop Application Programming Interfaces (APIs) to allow on-line systems with which students interact directly, such as LMS, publisher e-Text systems and Massively Open Online Courses (MOOCs) to interface with the *analysis and reporting server*. The APIs will allow these systems to administer AACR questions to students then submit the student responses to our analysis and reporting server and have a report returned to those systems where the students and their faculty can receive the feedback, expanding the transportability of this project.

Goal: Transport these innovations by providing faculty professional development

If you build it, they will (not necessarily) come.

Developing and testing reformed-based curricular materials and then disseminating them to faculty does NOT result in widespread adoption and implementation [27-32]. Therefore, a key feature of this proposal is **the integration of ongoing faculty professional development (PD) to support the**

transportability [33] **of the assessments and to help faculty persist in their use of the assessments.** Critical features of PD programs that successfully promote change among faculty [29] include:

1. An extended period of professional development. One-time workshops do not support sustained adoption. Successful PD usually includes ongoing support lasting a semester or more.

2. Performance evaluation and feedback. It is important for the adopting faculty to interact with the curriculum developers throughout implementation to ensure implementation fidelity and to promote faculty metacognition about how those materials fit within a larger model of learning and instruction.

3. A focus on changing faculty conceptions about teaching and learning. STEM faculty rarely have formal training in teaching or learning theory, but they often have implicit conceptions about teaching and learning [“folk pedagogies”, 34]. Therefore, faculty may need explicit support for their own conceptual change [35] to facilitate reformed teaching.

Professional development activities for faculty users of AACR assessments

To support faculty awareness, adoption, and sustained use of AACR assessments, we will provide and promote multiple PD opportunities. Our prior research has produced a suite of AACR questions in biology, so we will begin PD activities in the field of biology and extend into chemistry, chemical engineering and statistics in year 4.

1. **Instructional materials**: In our JiTT pilot, some participating instructors asked for curricular materials to address students’ misconceptions identified by the AACR assessments. Therefore, we will create instructional materials – including clicker questions, case studies, small group activities, and exam questions – that are adaptable to various classroom sizes and instructional methods. Each set of materials will go through an iterative process of development, implementation, and revision at four different institutions. To foster faculty conceptual change, we will also provide information about the most effective implementation of the instructional materials [36-42], including videos that show faculty demonstrating high fidelity implementation of the materials. We will also discuss the implementation of instructional materials during workshops and mentoring interactions (see below).

2. **Workshops**: We will offer workshops to introduce faculty to the AACR assessments and instructional materials. Since one-time workshops are unlikely to lead to effective implementation [29, 32], we will encourage participants to become involved in additional opportunities, such as mentoring and the on-line community described below. Workshops will be held at professional meetings, led by various members of the research team.

3. **Mentoring**: We will provide mentors for extended personalized support to facilitate continued participation in the assessment community. Mentors will provide one-on-one support for mentees (new faculty) throughout their first semester of implementation of the AACR assessments and instructional materials. Mentors will also discuss conceptions (folk conceptions, misconceptions) about teaching and learning, and engage mentees in evidence- and theory-based ways of thinking about AACR assessments and the broader context of teaching and learning.

4. **On-line community development**: The social collaboration environment of the web portal will support virtual communities among mentors, mentees, other faculty users, and the research team.

This collaborative environment is particularly important for peer support, so faculty can share their successes and challenges using the AACR assessments and instructional materials. Asynchronous interaction with other users in the on-line community via the web portals discussion forums and synchronous interaction via video-conference will help provide a support network for faculty.

The full selection of activities will be made available to faculty when they register on the website for the first time. Given differences in faculty members' interests and motivation [43, 44], their desire to contextualize the innovation for their own classroom conditions [30, 42], and their personal time constraints, we expect that faculty members will select PD activities according to their desired level of engagement. PD activities will be designed so that faculty at different levels of engagement will receive the support and feedback necessary to sustain their involvement.

Research on professional development

To understand the impact of the AACR project on faculty, we will investigate several research questions.

1. Which PD components influence sustained adoption of AACR assessments?

Although previous studies [29, 30, 32] have shown that PD and extended support facilitate sustained adoption and high-fidelity implementation, we have no *a priori* way to determine which components of the PD programs for the AACR assessments are most effective. Therefore, we will investigate how various components of the on-line assessment community and PD activities influence faculty participation. Using mixed-methods, we will analyze the potential impact of variables such as having a mentor, frequency of website visits, and classroom size. These data and others can be readily collected from registration for workshops or regular interaction on the portal, such as signing up for mentoring, registering to use assessments, and commenting on website content. Statistical models will be used to determine the relative contributions of these components to sustained adoption of AACR assessments.

We will also use qualitative methods, interviewing faculty to understand the factors influencing faculty adoption decisions such as 1) knowledge of how to use AACR assessments; 2) knowledge of the theoretical underpinnings of teaching tools like AACR assessments; 3) interactions with other AACR assessment users; 4) perceived barriers to, and benefits of, using AACR assessments; 5) compatibility of the AACR assessments with overall goals for their classrooms; and 6) engagement with a mentor [29, 45]. We will select faculty for a series of six semi-structured interviews over a two-year time period. As some faculty may discontinue use of the AACR assessments during the study, we will also be able to immediately investigate barriers to sustained use of the AACR assessments.

2. What characteristics of faculty PD engagement are associated with student learning?

Though improving student learning is a primary goal of STEM education research, evidence that the implementation of innovative teaching strategies actually improves student learning is sparse [29]. We are well-positioned, however, to address the relationship between faculty adoption of innovative research-based instructional activities and student learning. To determine whether faculty engagement with PD is positively associated with student learning of concepts, we will examine the relationships between student learning gains from pre-post analysis of AACR assessments and their relationships with faculty use of professional development components and instructional activities.

3. How do faculty interact in an on-line community environment?

We will study the relationships among participants in the on-line community using social network analysis to determine their patterns of interaction and information flow through the community. Data collection will be coordinated via the IBM Collaboration and Deployment Services. User interactions through the portal will be recorded and statistical analyses will be conducted using social network analysis [46] modeling features available in IBM SPSS Modeler.

Goal: Expand our basic research to chemistry, chemical engineering, and statistics

In this project, we are also adding additional collaborators to expand our disciplinary scope: 1) chemistry which is a foundational science for biology and engineering in which multiple representations are barriers to student learning; 2) chemical engineering with a focus on thermodynamics; and 3) statistics, where understanding probabilistic thinking is challenging for students across disciplines. We will support researchers in these disciplines who will explore AACR techniques through all phases of the QDC (Figure 1). We will also continue to advance our current work by 1) improving the assessments (e.g., validity, reliability) and experimental designs; 2) creating additional confirmatory models; and 3) refining our models of student reasoning.

Investigating Student Representations in Chemistry

We will adapt and develop a range of chemistry questions to probe three inter-related chemistry concepts: structure, properties, and energy changes. Previously we have developed and validated instruments to assess students understanding of the structure property relationship [47] and have been able to compare achievements for students in a new chemistry curriculum to an equivalent cohort of students in a more traditional chemistry course [48]. In this project a similar experimental design will be used to achieve three goals:

1. Compare student responses to existing questions (for which the AACR project already has a large database of responses) about acidity and basicity, with responses to questions where students must draw models and construct arguments. This comparison will provide convergent validation of student responses to the AACR questions, with their drawings, models and arguments about the same concepts.

2. Compare student responses to the existing AACR acid-base questions using two equivalent cohorts of students taken from the CLUE general chemistry and traditional general chemistry courses to investigate whether the AACR system can detect different types of responses from students.

3. Develop a set of AACR questions that probe student understanding of chemical energy, a topic that is highly problematic for students at all levels [49].

Chemical Engineering

Thermodynamics is a key concept that cuts across chemistry, biology and chemical engineering. We will coordinate with the work in chemical energy to evaluate questions in *Chemical Engineering Thermodynamics*, a course taught to sophomores and juniors in the Chemical Engineering program. The course develops students' skills for energy balances of processes with multiple units; calculation of thermodynamic properties of fluids; modeling of phase equilibriums. A similar thermodynamics course is required in all chemical engineering curricula.

Statistics

We will complete the QDC (Figure 1) for two sets of questions designed to assess student understanding in two areas of statistical thinking. The research team will write and pilot open response questions for those concepts, completing the rubric development, human coding and question modification portion of the QDC for questions related to the selected statistical topic. These data will be hand and machine scored and will be used to create scoring models that can be uploaded to the portal, completing the confirmatory analysis, predictive model, data collection and text analysis resource development phases of the QDC. Concurrently, a second set of questions designed around a second statistical concept found to be problematic to students will be piloted. The research team will complete the confirmatory analysis, predictive model, data collection and text analysis resource development for the second identified statistical concept and will load the questions associated with the first concept into the portal to start enacting the exploratory phase of the QDC. Finally, statistics instructors not associated with the research team, both at UGA and elsewhere will be invited to use the questions in the portal and the UGA research team will provide instructors with the PD necessary to incorporate the student results into classroom instruction.

Goal: Engage in ongoing project evaluation for continuous quality improvement

There is an external evaluator for the project, independent of the project research team. Project evaluation will use a formative-summative design focused on project objectives. ***Formative evaluation*** will provide timely feedback to the research and development teams to improve programming. ***Summative evaluation*** will focus on project effects on participants, including changes in faculty pedagogical content knowledge [50] and instructional and assessment strategies and in students' abilities to provide high-level explanations of their content learning.

The purpose of the evaluation will be to 1) determine the impact of the program on participating faculty and project staff, and therefore the projects potential impact on prevailing models of undergraduate STEM education; 2) provide evaluative data to staff to improve programming; 3) assess progress toward project objectives, especially in regard to implementing results in new contexts; and 4) identify strengths and limitations of the proposed project. Audiences for the evaluation will include project staff and funders. The evaluation will be framed by four key evaluation questions which are presented in Figure 5 along with benchmarks of accomplishment and data collection procedures.

Evaluation will be a collaborative effort between project staff and the project advisory board. Standard quantitative analysis methods will be used for survey and systematic observation data; appropriate qualitative methods will be used for interview and general observations. The evaluation staff will prepare data collection instruments and procedures, gather data, compile and analyze data, and prepare reports. Evaluators will serve on the project management team.

Goal: Lay the foundation for sustainability

There are opportunities for the application of these techniques to on-line and other systems with which students interact directly. The current enthusiasm for Massively Open On-line Courses (MOOCs), moves by publishers into interactive e-text books, and competition among LMS vendors to provide more detailed learning analytics and other automation, suggest that these are potential revenue sources that could be used to sustain the operation and maintenance of the Web Portal. This model would allow individual faculty members to utilize the Web Portal for no charge as we are proposing, with the infrastructure maintained by fees generated by licensing to for-profit entities. The development of the API to interface between our portal and these systems will provide the technical foundation for these interfaces. In the final two years of the project, we will approach commercial interests to explore these possibilities.

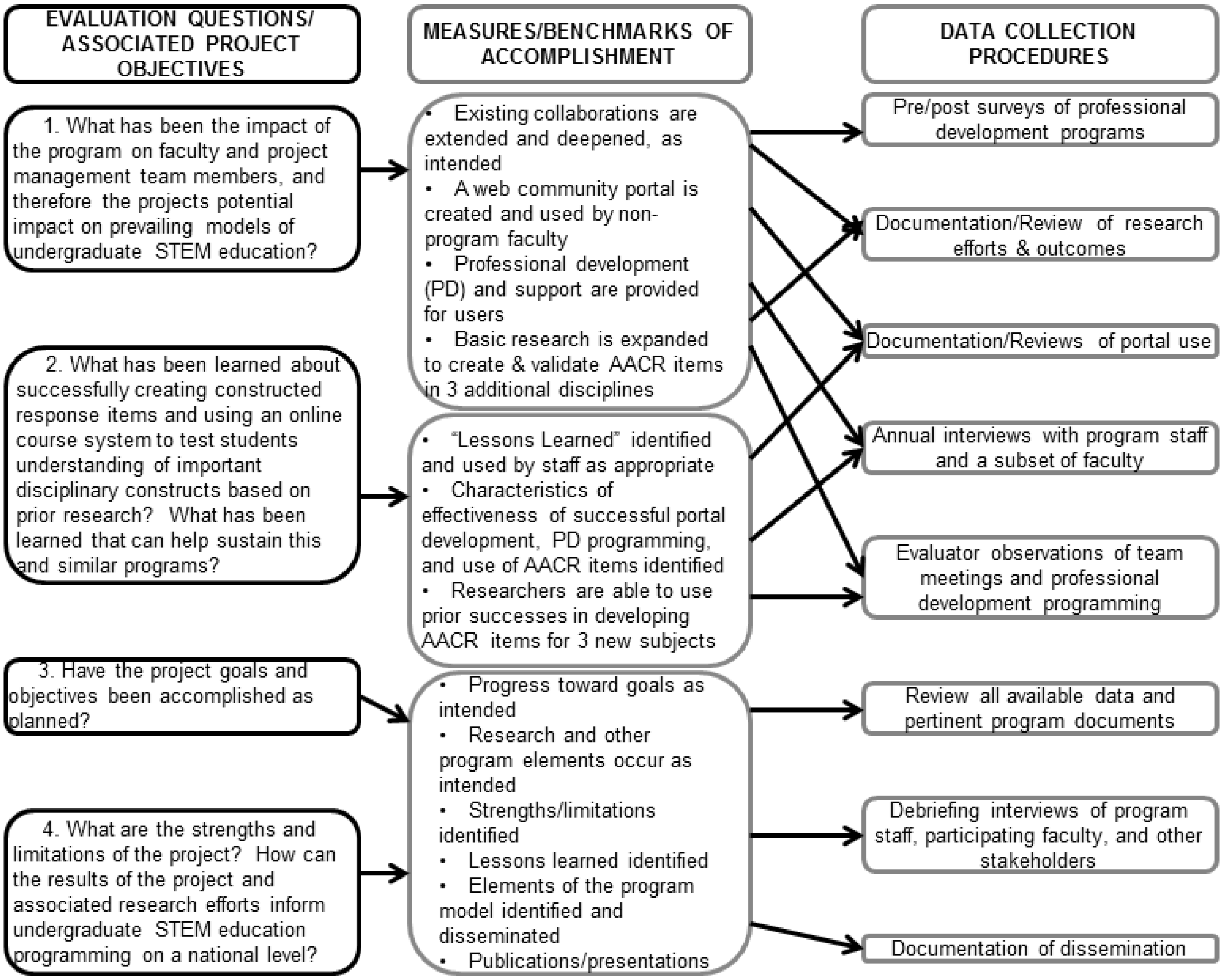

EVALUATION QUESTIONS/ ASSOCIATED PROJECT OBJECTIVES
MEASURES/BENCHMARKS OF ACCOMPLISHMENT
DATA COLLECTION PROCEDURES
1. What has been the impact of the program on faculty and project management team members, and therefore the projects potential impact on prevailing models of undergraduate STEM education?
2. What has been learned about successfully creating constructed response items and using an online course system to test students understanding of important disciplinary constructs based on prior research? What has been learned that can help sustain this and similar programs?
3. Have the project goals and objectives been accomplished as planned?
4. What are the strengths and limitations of the project? How can the results of the project and associated research efforts inform undergraduate STEM education programming on a national level?
• Existing collaborations are extended and deepened, as intended
• A web community portal is created and used by non-program faculty
• Professional development (PD) and support are provided for users
• Basic research is expanded to create & validate AACR items in 3 additional disciplines
• "Lessons Learned" identified and used by staff as appropriate
• Characteristics of effectiveness of successful portal development, PD programming, and use of AACR items identified
• Researchers are able to use prior successes in developing AACR items for 3 new subjects
• Progress toward goals as intended
• Research and other program elements occur as intended
• Strengths/limitations identified
• Lessons learned identified
• Elements of the program model identified and disseminated
• Publications/presentations
Pre/post surveys of professional development programs
Documentation/Review of research efforts & outcomes
Documentation/Reviews of portal use
Annual interviews with program staff and a subset of faculty
Evaluator observations of team meetings and professional development programming
Review all available data and pertinent program documents
Debriefing interviews of program staff, participating faculty, and other stakeholders
Documentation of dissemination

Figure 5: Evaluation Plan Logic Model.

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. This project builds upon our extensive research in science education, assessment and student cognition to expand a collaborative network to create and automatically analyze constructed response conceptual assessments that provide insight into student thinking and learning about key STEM concepts. The results of this project will be available for use by STEM faculty and science educators to gauge the efficacy of instruction or instructional change or to give an instructor a unique perspective on students' conceptual understanding, even in large enrollment STEM courses. By integrating a faculty professional development program that is grounded on research in faculty change, this project will facilitate the NSF DRL *Cycle of Research and Development* by bringing together STEM education researchers, faculty interested in discipline-based education research (DBER), Scholarship of Teaching and Learning (SoTL) and instructors who teach foundational STEM courses and meets the desired societal outcome of improving STEM education and educator development at the undergraduate level.

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 National Association for Research in Science Teaching Conference, Anaheim, CA, 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 Transforming Undergraduate Biology Education: Mobilizing the Community for Change, Washington, D.C., 2009.

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

11. 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.

12. 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.

13. 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.

14. 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.

15. 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 Frontiers in Education, San Antonio, TX, 2009.

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

17. C. Fellbaum, *WordNet: An electronic lexical database,* Cambridge, Mass.: MIT Press, 1998.

18. G. A. Miller, “WordNet: a lexical database for English,” *Communications of the ACM,* vol. 38, no. 11, pp. 39-41, November, 1995.

19. 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, 2011.

20. 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.

21. 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.

22. 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." p. 19.

23. IBM, "IBM SPSS Modeler Version 14.2," 2011.

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

25. 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 American Society for Engineering Education, Atlanta, 2013.

26. E. Wenger, *Communities of practice: Learning, meaning, and identity,* New York: Cambridge University Press, 1998.

27. A. L. Beach, C. Henderson, and N. D. Finkelstein, “Facilitating change in undergraduate STEM education,” *Change: The Magazine of Higher*

Learning, vol. 44, no. 6, pp. 52-59, December 10, 2012.

28. M. H. Dancy, and C. Henderson, “Barriers and promises in STEM reform,” in National Academies of Science Promising Practices Workshop, Washington, DC, 2008.

29. C. Henderson, A. Beach, and N. Finkelstein, “Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature,” *Journal of Research in Science Teaching*, vol. 48, no. 8, pp. 952-984, 2011.

30. C. Henderson, and M. H. Dancy, “Physics faculty and educational researchers: Divergent expectations as barriers to the diffusion of innovations,” *American Journal of Physics*, vol. 76, no. 1, pp. 79-91, 2008.

31. C. Henderson, and M. H. Dancy, *Increasing the impact and diffusion of STEM education innovations*, National Academy of Engineering, New Orleans, LA, 2011.

32. 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.

33. J. Feser, M. J. Borrego, R. Pimmel, and C. K. Della-Piana, "Results from a survey of National Science Foundation Transforming Undergraduate Education in STEM (TUES) program reviewers." in American Society for Engineering Education San Antonio, TX, 2012.

34. J. Bruner, "Folk pedagogy," *The culture of education*, pp. 44-65, Cambridge, MA: Harvard University Press, 1996.

35. S. Vosniadou, “International handbook of research on conceptual change,” New York: Routledge, 2008,

36. E. Mazur, *Peer instruction: A users' manual*, First ed., Upper Saddle River, NJ: Prentice Hall, 1996.

37. J. E. Caldwell, “Clickers in the Large Classroom: Current Research and Best-Practice Tips,” *CBE Life Sci Educ*, vol. 6, pp. 9-20, 2007.

38. M. K. Smith, W. B. Wood, W. K. Adams, C. Wieman, J. K. Knight, N. Guild, and T. T. Su, “Why peer discussion improves student performance on in-class concept questions,” *Science*, vol. 323, pp. 122-124, January 2, 2009.

39. S. T. Kalinowski, M. J. Leonard, and T. M. Andrews, “Nothing in evolution makes sense except in the light of DNA,” *CBE Life Sciences Education*, vol. 9, no. 2, pp. 87-97, Summer, 2010.

40. T. M. Andrews, S. T. Kalinowski, and M. J. Leonard, ““Are Humans Evolving?” A classroom discussion to change student misconceptions regarding natural selection,” *Evolution: Education and Outreach*, vol. 4, no. 3, pp. 456-466, September, 2011.

41. S. T. Kalinowski, T. M. Andrews, M. J. Leonard, and M. Snodgrass, “Are Africans, Europeans, and Asians different "races"? A guided-inquiry lab for introducing undergraduate students to genetic diversity and preparing them to study natural selection,” *CBE Life Sciences Education*, vol. 11, no. 2, pp. 142-151, Summer, 2012.

42. K. Smith, “Lessons learnt from literature on the diffusion of innovative learning and teaching practices in higher education,” *Innovations in Education and Teaching International*, vol. 49, no. 2, pp. 173-182, 2012.

43. 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.

44. F. D. Roberts, C. L. Kelley, and B. D. Medlin, “Factors influencing accounting faculty members' decision to adopt technology in the classroom,” *College Student Journal*, vol. 41, no. 2, pp. 423, June, 2007.

45. E. M. Rogers, *Diffusion of innovations*, 4th ed., New York: Free Press, 1995.

46. W. R. Penuel, W. Sussex, C. Korbak, and C. Hoadley, “Investigating the potential of using social network analysis in educational evaluation,” *American Journal of Evaluation*, vol. 27, no. 4, pp. 437-451, December, 2006.

47. M. M. Cooper, S. M. Underwood, and C. Z. Hilley, “Development and validation of the implicit information from Lewis structures instrument (IILSI): do students connect structures with properties?,” *Chemical Education Research and Practice*, vol. 13, pp. 195-200, 2012.

48. M. M. Cooper, S. M. Underwood, C. Z. Hilley, and M. W. Klymkowsky, “Development and assessment of a molecular structure and properties learning progression,” *Journal of Chemical Education*, vol. 89, no. 11, pp. 1351-1357, 2012.

49. M. M. Cooper, and M. W. Klymkowsky, “The trouble with chemical energy,” In Press.

50. L. S. Shulman, "Knowledge and teaching: foundations of the new reform," *Teaching as community property: Essays on higher education*, pp. 83-114, San Francisco, CA: Josey-Bass, 2004.

Biographical Information

Mark Urban-Lurain is an Associate Professor and Associate Director of the Center for Engineering Education Research at Michigan State University. He is the lead PI and project director of the AACR project. He is responsible for teaching, research and curriculum development, with emphasis on engineering education and, more broadly, STEM education.

Melanie Cooper is the Lappan-Phillips Professor of Science Education and Professor of Chemistry at Michigan State University. Her research has focused on improving teaching and learning in large enrollment general and organic chemistry courses at the college level, and is co-author of the NSF supported ”Chemistry, Life, the Universe & Everything”. She is a Fellow of the American Chemical Society and the American Association for the Advancement of Science, a member of the Leadership team for the Next Generation Science Standards (NGSS), and the National Research Council advisory Board on Science Education (BOSE) and has received a number of awards.

Kevin Haudek is a Research Specialist in the Center for Engineering Education Research at Michigan State University. He is a member of the AACR research group. His research interests are in student understanding and application of chemistry in biological contexts and strategies to increase student writing in undergraduate STEM courses.

Jennifer Julia Kaplan is an Assistant Professor in the Department of Statistics at the University of Georgia. Her research field is statistics education: in particular, undergraduate learning of statistics. In addition to her work on the AACR project she is studying the effects of language on student learning in statistics, student understanding of histograms and changes in pedagogy in undergraduate service courses that improve student learning in statistics.

Jenny Knight is a Senior Instructor in the Department of Molecular, Cellular and Developmental Biology at the University of Colorado, Boulder, where she has also been the departmental coordinator of the Carl Wieman Science Education Initiative. She also directs the National Academies Regional Mountain West Summer Institute on Undergraduate Education in Biology. Her research focuses on developing meaningful biology concept assessments, uncovering student misconceptions in genetics and introductory biology, and most recently, on the nature of in-class student discussions and their impact on student reasoning and learning.

Paula Lemons is an Associate Professor in the Department of Biochemistry and Molecular Biology at the University of Georgia. She does research on problem solving among college biology students. She also investigates the process by which college biology instructors make changes to their teaching.

Carl T. Lira teaches thermodynamics at all levels, chemical kinetics, and material and energy balances. He has been recognized with the Amoco Excellence in Teaching Award, and multiple presentations of the MSU Engineering College Withrow Teaching Excellence Award. He is co-author of a widely-used textbook, Introductory Chemical Engineering Thermodynamics. He has active research in phase equilibria, kinetics, alternative fuels, and reactive distillation. He has M.S. and Ph.D. degrees from the University of Illinois, Champaign-Urbana, and a B.S. from Kansas State University.

John E Merrill is an Associate Professor, Department of Microbiology and Molecular Genetics, Michigan State University. He is director of the Biological Sciences Program which administers the core undergraduate biology courses in service to multiple life science departments within the College of Natural Science at MSU. His current research interest is in finding better ways to assess his students' learning of core biology concepts.

Ross Nehm is Associate Professor in the Department of Ecology & Evolution, and Core Faculty in the Ph.D. Program in Science Education at Stony Brook University (SUNY). He has authored or coauthored 50 journal articles and book chapters and presented more than 100 conference talks and papers. He serves on multiple journal editorial boards. His major awards include a NSF CAREER award, a teaching award from Berkeley, a mentoring award from CUNY and an Education Mentor in the Life Sciences from the National Academies.

Luanna B. Prevost is an Assistant Professor in the Department of Integrative Biology at the University of South Florida. Her research interests are in the assessment of student learning, automated analysis of student writing in biology, problem solving in biology and graduate student teaching professional development.

Michelle Kathleen Smith is an Assistant Professor at the University of Maine in the School of Biology and Ecology and the Research in STEM Education (RiSE) Center. Her laboratory engages undergraduate and graduate students, K-12 teachers, and university faculty in research on teaching and learning. She has authored or co-authored 20 articles and book chapters. Dr. Smith is a PI or co-PI on several NSF awards funded through the TUES, WIDER, and Noyce programs.

Mary Anne Sydlik is the Director of the Science and Mathematics Program Improvement (SAMPI) Center, an outreach division of the Mallinson Institute For Science Education at Western Michigan University. She is the external evaluator for the AACR project. Her interests are in supporting efforts to improve the educational experiences and outcomes of undergraduate STEM students. She has been the lead external evaluator for a number of STEM and NSF funded projects. She also currently serves as the internal evaluator for WMU's Woodrow Wilson Fellows project and the institution's Howard Hughes Medical project.