Exploring Computerized Lexical Analysis to Predict Calibrated Peer Review Ratings of Student Writing in Chemistry

Kevin C. Haudek*, Arlene A. Russell^ and Mark Urban-Lurain*

*Center for Engineering Education Research, Michigan State University
^ Department of Chemistry and Biochemistry, University of California – Los Angeles

Paper presented at the annual meeting of the National Association for Research in Science Teaching, Rio Grande, Puerto Rico, April 6-9, 2013.

Correspondence concerning this article should be addressed to:

Mark Urban-Lurain, Center for Engineering Education Research, Michigan State University, Room 1410B, 428 S. Shaw Lane, East Lansing, MI 48824
Email: urban@msu.edu.

Abstract

Writing in undergraduate science courses represents an authentic scientific task and allows insight into student thinking, but is often limited in large enrollment courses due to resource constraints. We are investigating the combination of two approaches for evaluating student writing to overcome these constraints: using Calibrated Peer Review (CPR) and computerized text analysis. We investigate the possibility of computerized analysis of long (approximately 1700 characters), highly-structured essays, which have been scored by multiple trained, student peer reviewers. We extended and revised resources created in previous lexical projects for an assignment about buffer systems given in a general chemistry course. Our analysis revealed that students used many ideas in their writing. Over 100 lexical categories were created to capture the ideas in student writing, with each student's response placed into about 40 categories. We used these lexical categories as independent variables in statistical models to predict peer ratings. Also, we investigated the addition of limited semantic information (word proximity) and results of writing-quality rubric criteria as independent variables. Addition of these variables resulted in a better performing statistical model, with R^2=0.551. This scoring model used 24 of the independent variables: 20 lexical categories, 2 rubric criteria, one word proximity pair and response length. The lexical categories selected by the scoring model align well with the scoring rubric used by the peer reviewers, providing face validity for the lexical analysis. In addition, both rubric criteria included as variables were selected by the statistical model, indicating that writing quality was an important consideration of peer review.

Introduction

Ideally, assessment data should provide information to instructors about their students' thinking so that instructors can design appropriate instructional interventions (Bransford, 2000; Von Glasersfeld, 1994). In large enrollment courses, typical of college-level introductory STEM courses, assessments are usually limited to multiple choice instruments that can be machine scored. However, there is evidence that students may correctly answer multiple-choice questions but still harbor misconceptions, or *conceptual barriers*, which seriously compromise their learning (Nehm & Schonfeld, 2008).

Constructed response assessments, in which students have to use their own language to demonstrate their knowledge, can provide good insight into student thinking (Birenbaum & Tatsouka, 1987). These assessments also allow the uncovering of *mixed models* of student thinking (Haudek, Prevost, Moscarella, Merrill, & Urban-Lurain, 2012). Many students do not just have exclusively right or wrong ideas, but a mixture of these ideas, connected in some fashion. In addition to revealing student understanding, the exercise of writing in STEM courses correlates with increased learning and interest, leading to renewed calls for *writing-to-learn* strategies (Reynolds, Thaiss, Katkin, & Thompson, 2012). The call for writing-to-learn in science suggests that writing is an authentic task of scientists and is required to be an effective communicator and researcher (Moore, 1994). In addition, studies on expository writing in STEM courses show that students that write report improved understanding of the course material (reviewed in Rivard, 1994). Although widely seen as more reflective of student thinking and a more authentic scientific task, constructed response or writing assessments are significantly challenging to use in large-enrollment courses given existing resource constraints for their evaluation.

One approach to overcome this problem of evaluation is the use of trained student peer reviewers. The Calibrated Peer Review (CPR) system was designed to increase the amount of writing and evaluation done by students in a chemistry course (Russell, Chapman, & Wegner, 1998). Briefly, students in the CPR system are "calibrated" as reviewers by evaluating example essays with a rubric supplied by the instructor. After a calibration exercise, the student peer reviewers are given feedback on their evaluation of the essay. After successful completion of a calibration exercise, each student is assigned a rater score; those well-aligned to the rubric receive a higher rater score. Calibrated peer reviewers then evaluate and score other students' writing. The CPR system has since been expanded into numerous other disciplines such as biology (Clase, Grundlach, & Pelaez, 2010; Gerdeman, Russell, & Worden, 2007), engineering (Furman & Robinson, 2003) and statistics (Enders, Jenkins, & Hoverman, 2010), all with positive outcomes (Russell, 2004). Because the CPR program uses peer review for rating each submission in a web-based system, it overcomes many of the resource constraints for employing student writing, regardless of course size (Russell, 2004).

The Automated Analysis of Constructed Response (AACR) research group is investigating another possibility to overcome the evaluation challenge of writing assessments: the use of computerized analysis of text in large-enrollment, introductory STEM courses (Haudek et al., 2011). We take a two-stage approach to analyzing student writing. First, we use *lexical analysis* to extract key terms and concepts from student writing. We then use these terms and concepts as variables for *statistical classification* techniques to predict expert ratings of student responses to validate the lexical analysis. One advantage to this approach is that developed lexical resources should be applicable or easily extendable within a given discipline. However, one challenge in building statistical classification techniques is that it requires a large

number of scored responses, from which statistical models can be trained and tested. So far, we have successfully used computerized lexical analysis in studies of students' conceptual understanding in various STEM disciplines (Prevost, Haudek, Merrill, & Urban-Lurain, 2012; Weston et al., 2012) and/or predict subject expert ratings of short responses (Haudek, et al., 2012; Nehm & Haertig, 2012).

In this paper, we attempt to answer the question: do these previously developed automated analysis techniques transfer to predict trained peer scoring of longer essays? This report expands on previous work on computerized analysis by evaluating the lexical analysis of a highly-structured, scientific writing assignment, generally answered with long, complex essay responses. We also take advantage of the large number of peer-rated essays from CPR to build a statistical scoring model. In addition, we investigate whether including additional information about the text (i.e. writing-quality and limited semantic information) may improve statistical classification functions.

Methods

Data collection and CPR

Data used in this study was collected via the CPR system from an assignment about concepts in buffer chemistry. Students were directed to visit several buffer tutorial websites, then write about the carbonic acid/bicarbonate buffering system in blood. In addition to the general writing topic and learning goals of the assignment, students were given several specific writing prompts to direct their writing (e.g. *What is a buffer and how does it work?* and *What is acidosis?)* and a rubric that would be used to evaluate their writing. Nearly 400 responses were collected online from two different sections of second semester general chemistry from a large, public university in the U.S. These two sections shared a common instructor and were presented an identical writing assignment, as well as an identical calibration assignment and evaluation rubric (see below). All responses to the writing assignment were combined into a single data set, in order to build a more generalizable statistical model.

In the CPR system, each submitted response is evaluated anonymously by three peer reviewers (Russell, 2004). Each peer has been calibrated as a rater by a calibration assignment, in which the student evaluates training essays and their ratings are compared to levels set by the course instructor. Raters that show a high level of agreement with the set levels are considered well calibrated and given a larger reviewer competency index (RCI). The RCI determines how much weight is given to the reviewer's score of the submitted response. The overall score for a submitted response is calculated by a weighted average algorithm using all three peer scores of the text (from 0-10) and the reviewers' RCIs. In addition to assigning an overall score to the submission, the reviewers are also asked to answer a series of rubric questions which evaluate the content and quality of the submission. For each rubric question, each reviewer must answer Yes or No, and then is given an opportunity to explain. The rubric used for this assignment contained twelve evaluation criteria: one for the overall quality of the essay, rated on a scale of zero to ten (e.g. *Rate this text.),* nine question relating to content (e.g. *Is the removal of bicarbonate ion through the kidneys discussed?)* and two questions relating to writing quality (e.g. *Does the essay have a descriptive topic sentence* and *Is the text free of mechanical errors and/or any unformatted formula(s)/equations(s)?*) In total, we collected 388 submissions for lexical analysis using the CPR system. However, 23 submissions were dropped from statistical analysis because of incomplete rater information (i.e. either a submission was not evaluated by three peers or a reviewer had incomplete RCI data).

Software and Lexical Analysis

For both lexical and statistical analysis, we used commercially available software, IBM SPSS Modeler (v. 14.2; SPSS, 2011). This software allows the creation of a data *stream*, wherein a single data source can be analyzed by a variety of classification functions simultaneously, including both text analysis and statistical models. Briefly, the text analysis software function classifies written text into categories based on the appearance of specific *terms* (words and phrases) it detects using *libraries* to extract all recognizable terms. Although the software contains default libraries of common terms, they do not recognize most of the technical lexicon of scientific disciplines and therefore custom libraries are created and revised. In this project, we have used and expanded on lexical resources created in previous work (for more detailed discussion of the analytic techniques see Haudek, et al., 2012). *Categories*, which represent a single homogenous concept, may contain multiple terms. These categories can be created and/or revised using linguistic algorithms contained in the software or defined by the user. The output of text analysis used for statistical classification is a series of binary variables, whether each category is present or absent in a given response. Each response is usually placed into multiple categories.

Word proximity program

In order to add additional information about the response, we developed a computer program that identifies text records that contain two words within a given "window" size (the program is available for download after registering for a free user account at: www.msu.edu/~aacr). The user defines both the words the program searches for and the window size constraint. For example, the user may enter the words "pH" and "increase" with a window size of two. This indicates the words "pH' and "increase" must be adjacent words, in either order. Using the same words with a window size of three indicates that there can be no more than one word between "pH" and "increase", again in any order. The output of the word proximity program used for statistical classification is a series of binary variables, whether each word pair is present or absent in the defined window space within each response.

Results of several trials using words that varied only in inflection or singular/plural were combined into a single variable for statistical classification. For example, the results of "pH" and "increases" at window size of two were combined with the results of "pH" and "increased" at window size of two. For this study, we included three word pairs (including word variants as described above) at window sizes of two and three based on the writing prompts and expected text: increase and pH, decrease and pH, shift and equilibrium.

Statistical Classification Techniques

We used the categories from lexical analysis, results from the word proximity program and rater answers to the writing-quality evaluation rubric questions as binary variables in step-wise forward regressions, with the CPR weighted average score as the dependent, target variable. For both stepwise analyses, we used an entry F-value of 3.84 and removal F-value of 2.71 and singularity tolerance of 0.0001.

Results

Lexical analysis

Student responses to the assignment (n=388) were used as data for text analysis. Responses were generally in the range from one to several paragraphs in length (number of characters: mean = 1684, standard deviation = 91, range = 1348 to 2061). During text analysis, there were a total of 119 fine-grained lexical categories created to contain only a single

homogenous concept. These categories were created via three different methods: by linguistic algorithms used by the software, by the researchers based on the writing prompt, or by concepts emergent in student writing. There are four categories used by all submitted responses (*buffer, bicarbonate forms, carbon dioxide* and *hydrogen ion*), which are part of the assignment prompt. Other concepts included in the prompt were also used frequently, such as *acidosis, equilibrium* and *exercise*. Other categories represented target concepts students were expected to address in their writing in order to correctly explain the topic, but did not appear in the question prompt (*conjugate* and *hemoglobin*). Similarly, some students included writing about the learning goals of the assignment, although not directly prompted to do so (e.g. *LeChatelier's principle*). Some emergent categories were used relatively infrequently (e.g. *ATP, protein, temperature*), but still represent discrete, unique concepts present in student writing. Each response was classified into an average of 43 categories (range = 31 to 55), indicating that students used a large number of different ideas in their writing.

Word proximity

The responses to the CPR assignment were much longer (i.e. several paragraphs) and more structured (i.e. several guiding writing prompts) than our previous studies of computerized lexical analysis, in which responses were typically one to a few sentences (for examples see Ha, Nehm, Urban-Lurain, & Merrill, 2011; Haudek, et al., 2012). Because of these differences in the responses, we investigated whether adding some semantic information (i.e. word proximity data) could aid in building statistical models. The rationale being that if two concepts (e.g. "change" and "equilibrium") appeared in the same response, we could not assume that the student meant to link these two ideas together, as they may have occurred in separate sentences or paragraphs. This problem is addressed by using the word proximity program, in which the user defines the space (or window) within which two words must appear. We utilized the word proximity program to evaluate the text for three common word pairs at small window sizes and incorporated these results with the results of the lexical analysis.

Building a statistical model using text analysis variables

The categories generated during text analysis, the character count of each response and the results of the word proximity program were used as independent variables in a linear regression to create a statistical scoring function to predict the weighted average score of the response. Student responses to the assignment that were rated by three peer reviewers were assigned a weighted average score, (n=365, range 2.2 to 9.5, mean = 6.2, standard deviation = 1.6) based on the scores given by each peer reviewer

The stepwise regression selected 18 variables in building the model (model 1; Table 1). The resulting model had R=0.682 and $R^2 = 0.465$. One of the variables chosen in the model was character count, which has a positive beta weight, meaning that responses with more characters (i.e. longer) had a higher score. The model also selected one word pair: *increase pH*. This variable represented several phrases that indicated an increasing pH (see Methods). This variable had a large, positive beta weight, indicating that it was an important factor in gaining a high overall score. In addition to character count and *increase pH*, sixteen lexical categories were chosen, with some having positive beta-weights and others negative. The fine-grained nature of the lexical categories can be noted by examining the difference in weight for the categories *strong* (when used generically) and *strong acid* (a specific, detailed phrase). Many of the significant and positive predictors (e.g. *kidney, glucose break down, ratio*) selected by the regression model represent important ideas in the scoring rubric used to evaluate this assignment (see Discussion).

Improving the statistical model with additional rubric criteria

The evaluation rubric used by students to evaluate the text contained two writing-quality criteria (see Methods). These criteria are difficult to match using lexical analysis, because the evaluation criteria rely on ordered and structured sentences or ideas, which lexical analysis cannot detect. To examine whether the prediction model could be improved by adding these writing-quality scores to the model, we performed another step-wise regression. For this model, each response was given a "weighted" score for each of the two writing-quality criteria. If all three raters agreed that a criterion was either met or absent, the text was assigned a one or zero, respectively. If there was disagreement between raters on a criterion, the average "weighted" score was calculated. This "weighted" rubric score relies on the criteria score each rater assigned to the text and each raters' RCI, so that highly calibrated raters had a bigger influence on the calculated score for a given criteria.

To test whether the statistical model could be improved, we included the weighted rubric score for the two writing quality criteria as independent variables, along with the lexical categories, character count and word proximity results, as above, in a step-wise linear regression. Again, the dependent variable was the weighted average score of the response.

Table 1. Standardized beta coefficients of variables selected by a step-wise forward regression using only lexical variables (model 1).

Variable name	Beta
acids - general *	-0.139
bases - general *	0.236
character count # *	0.134
constant *	0.160
energy	0.106
environment	-0.107
extracellular *	-0.153
glucose	0.103
glucose break down *	0.131
help *	-0.124
increase pH ^ *	0.268
ions *	-0.152
kidneys *	0.124
LeChatelier's principle	0.086
ratio *	0.267
strong *	0.166
strong acid	-0.094
water *	0.158

Notes: Variables in bold face are unique to this model and are not selected by the model in Table 2. # indicates the number of characters in the submitted text response. ^ indicates a variable from the word proximity program. All variables shown in the table have significance of $p<0.05$; * indicates $p<0.01$.

The stepwise regression selected a total of 24 variables (model 2; Table 2). The resulting model had $R=0.743$ and $R^2 = 0.551$, which is a good improvement over the initial model. Two of the variables chosen in the improved models are the writing-quality rubric criteria (Rubric – no mechanical errors and Rubric – topic sentence). Both variables have a positive, large beta weight, indicating that these criteria are important in predicting the overall score of the text, as expected. That is, responses that do not have mechanical errors and include a topic sentence are more likely to receive a higher score. In addition to the two rubric criteria, both character count and the word pair *increase pH* were again selected as variables with positive beta weights. Of the twenty other variables, thirteen are shared with model 1 (indicated by normal type face in Tables 1 and 2). These variables have similar beta-weights in both models. The seven new variables selected in the improved model (highlighted by bold in Table 2) are all lexical categories: *carbonate ion, exchange, function, lungs, named chemical compounds, phosphate buffer* and *transport*.

The scatter-plot of scores resulting from model 2 (Figure 1), shows that in general, the scoring model over-predicts low score responses and under-predicts high score responses. This may be due to the data set having more examples of middle-scoring responses to train the model, rather than responses that score very high or low. Therefore, the model is better trained at scoring "average" responses.

Table 2. Standardized beta coefficients of variables selected by a step-wise forward regression using rubric criteria and lexical variables (model 2).

Variable name	Beta
acids - general *	-0.130
bases - general *	0.202
carbonate ion	0.075
character count # *	0.154
constant *	0.120
exchange	0.082
extracellular *	-0.144
function	-0.082
glucose *	0.132
glucose break down *	0.127
help *	-0.123
increase pH ^ *	0.254
ions *	-0.124
kidneys *	0.109
lungs *	-0.101
named chemical compound	-0.081
phosphate buffer	-0.084
ratio *	0.236
Rubric - no mechanical errors *	0.233
Rubric - topic sentence *	0.123
strong *	0.128
strong acid	-0.097
transport *	-0.113
water *	0.169

Notes: Variables in bold face are unique to this model and are not selected by the model in Table 1. # indicates the number of characters in the submitted text response. ^ indicates a variable from the word proximity program. All variables shown in the table have significance of $p<0.05$; * indicates $p<0.01$.

Discussion

We were able to extend our previous work in lexical analysis to a related topic in chemical buffers. Minimal effort was necessary to extend the libraries of terms and a combined approach to building categories yielded robust bins for categorizing student responses. The long, structured essays used in this analysis contained many different concepts, including those from the question prompt, learning goals and rubric, as well as unique ideas emergent from the students' writing.

Using only the lexical categories and word proximity as independent variables, we were able to generate a statistical scoring model to predict the average weighted score of peer reviewers. This suggests that calibrated peer reviewers are consistent enough so that some of their ratings may be used to build robust scoring models. We note that some variables chosen in the resulting statistical model that have high beta weights (e.g. *glucose break down* and *ratio* (lexical categories) and *increase pH* (word pair result)) selected by the regression represent concepts that are included in the content questions on the evaluation rubric. Using statistical models to select categories with predictive values also lends validation to the lexical analysis and that the peer ratings reflect content that was supposed to be evaluated. Other categories with negative beta weights (e.g. *strong acids* and *acids - general*) may contain terms which are generally incorrect in response to the question prompt (such as discussing strong acids in an assignment about buffers) or too non-specific (using the term "acid" without being specific about related molecules or functions), so that peer reviewers must infer meaning.

Thus far, our research using text analysis has been lexical, whether a response contains a concept or not. There can be no direct relationship between concepts inferred from this, other than the student used (or did not use) two concepts in their writing. In this report, we have begun to explore ways to add additional, limited semantic information to our analysis by including results from a word proximity program as variables in the statistical models. The result that one of these variables was chosen by both models 1 and 2 and had a significant beta-weight supports the idea that such semantic information is one way to further refine lexical analysis results. For this study, we chose three word pairs expected in the essays because of the writing prompt. However, additional word pairs, generated by other expected answers or via emergent from student writing, could be explored and added as variables into future models. This limited semantic information is most likely more useful when evaluating long, complex essays, like those evaluated in this report, as compared to the short (one to two sentences) responses we have previously evaluated.

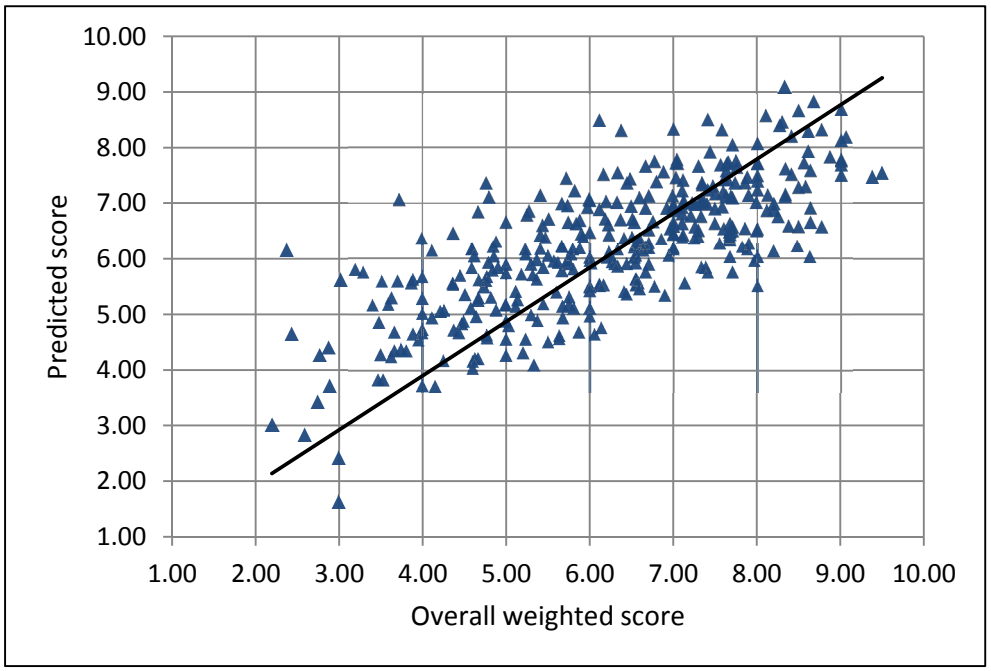

Figure 1. Predicted vs. actual score scatterplot using scoring model 2. The line y = x has been added as a reference.

The statistical model 1 was improved by adding the scores resulting from two of the evaluation criteria regarding writing quality from the scoring rubric, thus building another scoring model, model 2. These two added variables were not only selected by model 2, but both had large positive beta-weights, indicating they were important in determining the score assigned to a response. The rubric criteria of *no mechanical errors* had a larger beta weight (i.e. more influence) than the criteria of *topic sentence*, perhaps because peer reviewers have an easier time detecting writing errors (misspelling, unformatted chemical equations, etc.) then deciding what constitutes an adequate topic sentence. Alternatively, peer reviewers could deduct more points for, or be more influenced by, grammatical errors than for/by lack of a topic sentence (also see below).

The calculation of this rubric criteria independent variable was straight-forward where there was agreement between raters. Where there was disagreement between raters, we had to calculate a "weighted" score, which in part was determined by each raters' RCI. This is problematic in that the dependent variable, the overall score, is also influenced by the raters' RCI, therefore some interaction between the independent and dependent variables is expected. In addition to improving the overall prediction model, the addition of the rubric criteria as

variables changed some of the lexical categories chosen by the model. Most of the highly weighted lexical categories from the first model appear in the second model. These consistent variables include concepts related to the content questions of the rubric noted above (e.g. *ratio, glucose break down, kidneys*, etc.) In the future, we will investigate the correlation between the lexical categories that changed between models with each other and with the writing-quality rubric criteria.

One consideration for the statistical models reported here is that the text data we used originated from a single assignment and instructor (albeit from two different sections of the same course). In fact, we have performed preliminary analysis of assignments with similar / identical learning goals and writing prompts, but from different instructors. Including these data degrades the resulting classification functions. Even including data from the exact same assignment but used by a different instructor, degrades the accuracy of the classification function. We believe this is because each instructor sets his/her own rubric and calibration assignment for the peer raters, so that the same writing assignment may be evaluated differently. One goal (and challenge) of automated analysis is to create robust, stable scoring models that are broadly generalizable (see Ha, et al., 2011). However, as we build and explore the limitations of our models, we have chosen to use a smaller, more homogenous data set. Therefore, we must limit our expectations about how well this model could predict a *de novo* set of data if the CPR calibration criteria differ. We will continue to explore ways to balance the creation of the best predictive models with models that are generalizable, so that they can be applied to new sets of data.

Our current model was most accurate predicting weighted scores near the mean, while being less accurate predicting very low or high scores. We note that this problem is similar to that observed with poorly calibrated raters; they tend to over-rate poorly written essays and under-rate well-written essays. One problem for our scoring model may be that the difference between an "average"-scored essay and either a "low" or "high"-scored essay is based on the writing quality and not content. This adds further rationale for attempting to include limited rubric information into our models, so that writing-quality is used as a variable. As well, our hope is that, just like student peer reviewers, our scoring models will improve with additional calibration practice. Using larger data sets to build scoring models may help improve the problem of accurately predicting scores on the extremes by containing more examples of poorly-written (and low scored) and well-written (and high scored) essays.

Finally, we have just begun to explore the relationships between the evaluation rubric criteria and the overall score assigned to the text by each rater. We are interested in determining if certain rubric criteria are "weighted" more heavily by peer raters and whether this changes between raters of different evaluative competency (RCI). Based on our statistical model using the writing quality criteria, we have some evidence that this is indeed the case; certain rubric criteria influence the peer reviewer's overall score of the text more than other criteria.

Overall, we are encouraged by the results of this study combining CPR writing assignments and data with computerized analysis. First, we have shown the extension of lexical resources to new writing assignments. Second, the addition of limited semantic information is helpful in building classification models, at least for longer, structured essays. Having developed a useful proximity program, we now have the capability to refine the word pairs as necessary and include this information in future models. Finally, using some ratings from the evaluation rubric allowed additional refinement of the classification model and validity to the use of evaluation rubrics by peer reviewers.

Acknowledgements

We would like to thank members of the AACR research group for helpful discussions about this project. Analytic resources described in this report, as well as information about our research group and help with text analysis, can be found at: http://www.msu.edu/~aacr.

The results and materials reported here are based on work supported by the National Science Foundation (Grants DUE 1022653 and DUE 1143642 to M.U-L. and Grants DUE 04-42828, DUE 08-37229 and DUE 08-16660 to A.A.R). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

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

Bransford, J. (Ed.). (2000). *How people learn brain, mind, experience, and school* (Expanded ed.). Washington, D.C.: National Academy Press.

Clase, K. L., Grundlach, E., & Pelaez, N. J. (2010). Calibrated Peer Review for Computer-Assisted Learning for Biological Research Competencies. *Biochem Mol Biol Educ, 38*(5), 290-295.

Enders, F. B., Jenkins, S., & Hoverman, V. (2010). Calibrated Peer Review for Interpreting Linear Regression Parameters: Results from a Graduate Course. *Journal of Statistics Education, 18*(2), 1-27.

Furman, B., & Robinson, W. (2003). *Improving Engineering Report Writing with Calibrated Peer Review.* Paper presented at the 33rd Annual Frontiers in Education Conference, Piscataway, NJ.

Gerdeman, R. D., Russell, A. A., & Worden, K. J. (2007). Web-Based Student Writing and Reviewing Performance in a Large Biology Course. *Journal of College Science Teaching, 36*, 46-53.

Ha, M. S., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying Computerized-Scoring Models of Written Biological Explanations across Courses and Colleges: Prospects and Limitations. *Cbe-Life Sciences Education, 10*(4), 379-393. doi: 10.1187/cbe.11-08-0081

Haudek, K. C., Kaplan, J. J., Knight, J., Long, T., Merrill, J., Munn, A., Nehm, R. H., Smith, M. K., & Urban-Lurain, M. (2011). Harnessing Technology to Improve Formative Assessment of Student Conceptions in STEM: Forging a National Network. *CBE Life Sciences Education, 10*(2), 149-155. doi: 10.1187/cbe.11-03-0019

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J., & Urban-Lurain, M. (2012). What Are They Thinking? Automated Analysis of Student Writing about Acid-Base Chemistry in Introductory Biology. *CBE Life Sci Educ, 11*(3), 283-293. doi: 10.1187/cbe.11-08-0084

Moore, R. (1994). Writing as a Tool for Learning Biology. *Bioscience, 44*(9), 613-617. doi: 10.2307/1312461

Nehm, R. H., & Haertig, H. (2012). Human vs. Computer Diagnosis of Students' Natural Selection Knowledge: Testing the Efficacy of Text Analytic Software. *Journal of Science Education and Technology, 21*(1), 56-73. doi: 10.1007/s10956-011-9282-7

Nehm, R. H., & Schonfeld, I. S. (2008). Measuring Knowledge of Natural Selection: A Comparison of the CINS, an Open-Response Instrument, and an Oral Interview. *Journal of Research in Science Teaching, 45*(10), 1131-1160.

Prevost, L. B., Haudek, K. C., Merrill, J., & Urban-Lurain, M. (2012). *Deciphering Student Ideas on Thermodynamics Using Computerized Lexical Analysis of Student Writing*. Paper presented at the American Society for Engineering Education, San Antonio, TX.

Reynolds, J. A., Thaiss, C., Katkin, W., & Thompson, R. J., Jr. (2012). Writing-to-learn in undergraduate science education: a community-based, conceptually driven approach. *CBE life sciences education, 11*(1), 17-25. doi: 10.1187/cbe.11-08-0064

Rivard, L. P. (1994). A Review of Writing to Learn in Science - Implications for Practice and Research. *Journal of Research in Science Teaching, 31*(9), 969-983.

Russell, A. A. (2004, 2005). *Calibrated Peer Review: A writing and critical-thinking instructional tool.* Paper presented at the Invention and Impact: Building Excellence in Undergraduate Science, Technology, Engineering and Mathematics (STEM) Education, Washington DC.

Russell, A. A., Chapman, O. L., & Wegner, P. A. (1998). Molecular science: Network-deliverable curricula. *Journal of Chemical Education, 75*(5), 578-579.

SPSS. (2011). IBM SPSS Modeler 14.2 (Version 14.2). Chicago, IL: IBM.

Von Glasersfeld, E. (1994). A constructivist approach to teaching. In L. P. Steffe & J. Gale (Eds.), *Constructivism in Education* (pp. 3-15). Hillsdale, NJ: Lawrence Erlbaum Associates.

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