Insight into Student Thinking in STEM:
Lessons Learned from Lexical Analysis of Student Writing

Mark Urban-Lurain

Rosa Anna Moscarella

Kevin C. Haudek

Emma Giese

John E. Merrill

Duncan F. Sibley

Michigan State University

Contact: Mark Urban-Lurain

Division of Science and Mathematics Education, 111 N. Kedzie, East Lansing, MI 48824

urban@msu.edu

This material is based upon work supported by the National Science Foundation under awards DUE-0736952 and DUE-0243126. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).

Abstract

Constructed response assessments, in which students have to use their own language to demonstrate their knowledge, can provide good insight into student thinking. We have successfully used computerized lexical analysis in studies of students' conceptual understanding in biology, chemistry and geology. We use a two-stage approach to analyze constructed responses. 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. Based on our analysis of constructed response items, we have: 1) learned how to structure questions so that responses are better suited for lexical analyses; 2) learned effective ways to build and share custom discipline-specific lexical libraries; 3) gained insights into optimal numbers and specificity of categories; and 4) learned about optimizing the granularity of the classification rubrics used to rate student responses. In this paper, we summarize our work to date and describe some of the lessons we have learned in the hope that others can benefit from our work and adopt these techniques.

Keywords: constructed responses, lexical analysis, assessment

Insight into Student Thinking in STEM:

Lessons Learned from Lexical Analysis of Student Writing

Ideally, assessment data should provide information to instructors about their students' thinking so that instructors can design appropriate instructional interventions (Bransford, 2000; Duit, 1995; Larochelle & Bednarz, 1998; 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). As a result, multiple-choice instruments may not detect critical gaps in student understanding or prevailing patterns of thinking that could otherwise be addressed through appropriate instructional intervention.

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). Although widely seen as more reflective of student thinking, they are significantly challenging to execute in large-enrollment courses given existing resource constraints. At Michigan State University, we have successfully used computerized lexical analysis in studies of students' conceptual understanding in biology, chemistry and geology. These methods can be applied across large scales to support the improvement of STEM education by providing data to instructors that can drive instructional decisions. In this paper, we summarize our work to date and describe some of the lessons we have learned in the hope that others can benefit from our work and adopt these techniques.

Overview of Procedure

Deane (2006) broadly categorizes computerized linguistic analyses into three approaches:

1. Linguistic feature-based methods that extract linguistic features (e.g., WordNet, see Fellbaum, 1998; Miller, 1995) and use statistical methods to examine correlations with external variables such as human raters' scores. These methods focus on construct generalizations rather than language structure, and assessments need to be structured to keep features being measured closely aligned with the construct of interest. One advantage of this approach is that the nomothetic span can be engineered separately from the linguistic extraction by statistical modeling, reducing the need for an a priori model of student knowledge.

2. Vector space methods, of which Latent Semantic Analysis (LSA) is best known (Landauer, Foltz, & Laham, 1998; Landauer, Laham, Rehder, & Schreiner, 1997). These methods have been applied to scoring essays based on large corpora (usually thousands) of exemplar essays that are used to train the software. There is no direct extraction of construct-level categories or meaning. Rather, these methods analyze the co-occurrence of words in the training essays and score new essays by comparing the mathematical structure of word co-occurrence to the target essays. These methods are not as effective for short-answers.

3. Linguistic structure analyses that attempt to create student mental model representations by looking at relationships among elements, rather than simple word counts. A major drawback of these methods is the need to carefully engineer the implementation for accuracy based upon a hypothesized student knowledge model.

We take a two-stage, feature-based approach (# 1 above) to analyzing constructed responses as shown in Figure 1. First, we use lexical analysis software to extract key terms and scientific concepts from student writing. We then use these terms and concepts as variables for

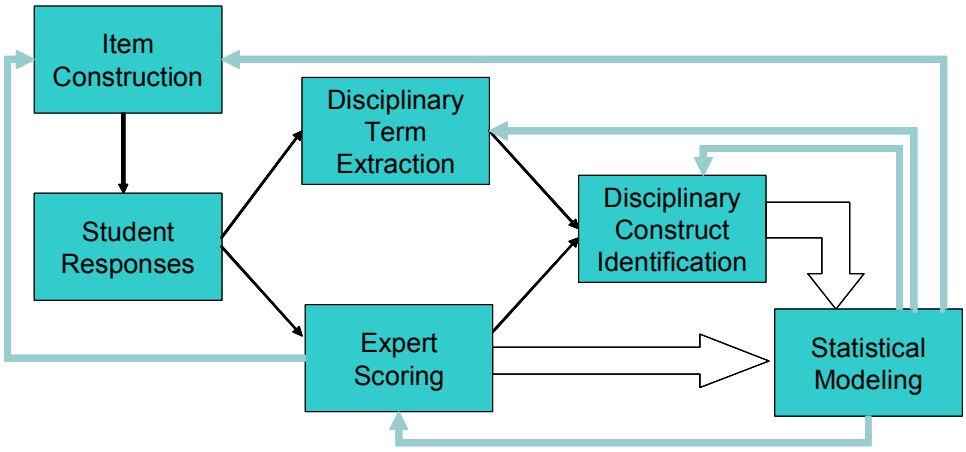

Figure 1: Two-stage, feature-based linguistic analysis framework

statistical classification techniques to predict expert ratings of student responses. This combination of analytic techniques provides insights into student thinking by categorizing written responses based on commonly used terms and phrases.

Computerized Lexical Analysis

SPSS Text Analytics (SPSS, 2009) is lexical analysis software designed to analyze and categorize open-ended responses on surveys. The software classifies written text into categories based on the appearance of specific terms (words and phrases) it detects using libraries to extract all recognizable terms. Graphical tools in the software enable a variety of ways for representing the categorized data that facilitate evaluation of the responses. Figure 2 shows a screen shot of data from an assessment in which students were asked to answer the question: *You have a friend who lost 15 pounds of fat on a diet. Where did the mass go?*

On the right side of the screen, student responses are shown in the middle column. In the right column are the categories into which each response has been classified. The categories for

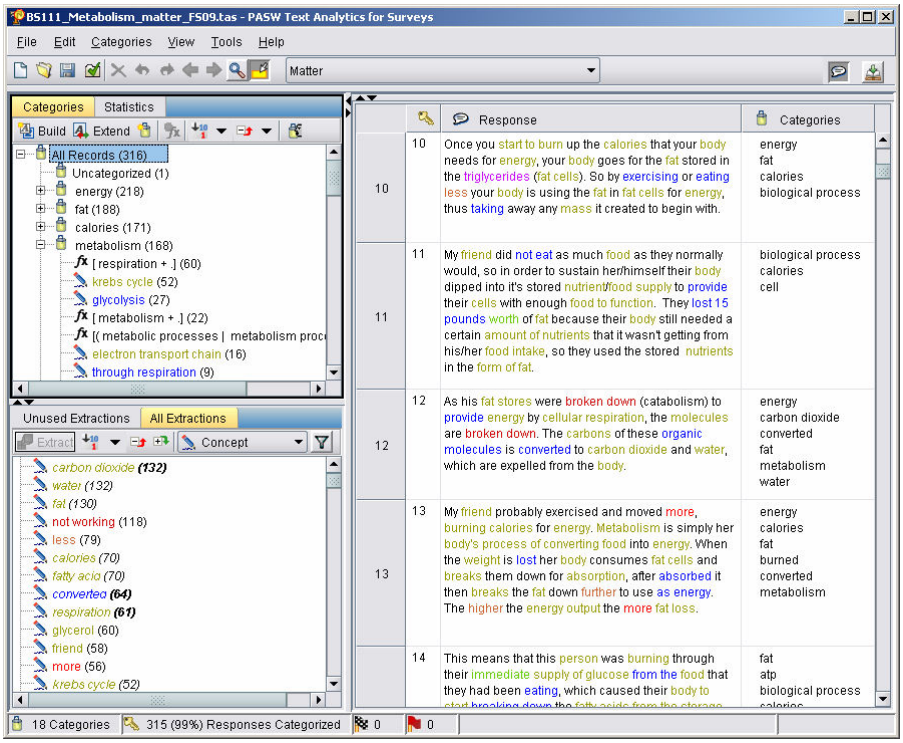

Figure 2: A sample screen from SPSS Text Analytics showing categories, category expansion, term extraction and response categorization. Upper left: list of categories with category "metabolism" expanded to show contained terms, phrases, and functions. Right: list of student responses. Rightmost column shows categories identified by the software for each response.

this question are shown in the upper left panel of the screen. In the lower left panel, the individual terms that the software extracted from the answers are displayed.

Although the software has several standard libraries of common terms, they do not recognize most of the technical lexicon of scientific disciplines. In our work, we have created several custom lexical libraries containing biological terms, as well as synonyms, abbreviations, and spelling variations and misspellings (Moscarella et al., 2008).

One challenge for students learning cellular biology is that they often fail to account correctly for matter and energy transformations, as in cellular metabolism (Wilson et al., 2006). To examine this, we created constructed response items asking students to trace carbon backwards during cellular respiration (Moscarella et al., 2008). After we created biology libraries and categories, the software was able to correctly categorize over 90% of the student

responses without human intervention. Lexical analysis allows us to categorize all compounds and processes in the responses to find patterns in student thinking.

Such insight into student understanding would be difficult to achieve if all student responses were manually evaluated. Tools such as lexical analysis software facilitate the processing of large numbers of student responses. This is essential for discerning broad patterns in student thinking and testing the robustness of those patterns by comparing student responses across assessments aligned to test similar concepts. Not only does this represent a powerful method for analyzing the outcomes of instructional change, it offers the potential for delivering a critical component of reformed science teaching – near real-time feedback that informs instructors about their students' learning.

Statistical Classification Techniques

While lexical analysis can provide formative feedback to instructors about common student conceptions, it is important to understand the relationships among the concepts identified by the lexical analysis that are indicative of expert understanding in a discipline. Humans evaluate constructed responses using a variety of *rubrics* that are designed to assess these relationships. We are combining the results of our lexical analyses with statistical classification techniques to predict how expert human raters would apply rubrics to evaluate student writing.

For example, in our research on acid/base chemistry of biological functional groups (Haudek et al., 2009), students were asked to *give an explanation of a strong acid* and then asked to *give an explanation of a weak acid*. A subset of student responses was scored using a 3-bin rubric by two experts (inter-rater reliability: intraclass correlation = 0.95). Discriminant analysis identified 14 of the lexical analysis categories that were the most useful for predicting expert ratings of the student responses. These categories were well aligned with the expert rubric,

providing good evidence of content validity. The function classified 83.8% of all cases correctly ($p < .001$); we would expect to classify 33% of the cases correctly. When inter-rater reliability is calculated between the experts and the computer predictions, the intra-class correlation is 0.882 ($p < .001$).

Our results also highlight the challenges of expert rating of student writing. It is often a lengthy process to develop the rubrics and train raters to ensure acceptable inter-rater reliability. It is also time-consuming for raters to evaluate large numbers of student responses and maintain consistency over time.

Improving The Utility And Efficiency Of Computerized Lexical Analysis

Based on our analysis of constructed response items, we have: 1) learned how to structure questions so that responses are better suited for lexical analyses; 2) learned effective ways to build custom discipline-specific lexical libraries; 3) gained insights about optimizing numbers and specificity of categories; and 4) learned about optimizing the granularity of the classification rubrics used to rate student responses. Finally, we have learned that in any *de novo* set of data analyzed by existing custom libraries and categories, typically 10% or less of student responses are uncategorized by the lexical analysis software. However, our experience has shown that most uncategorized responses are irrelevant to the question posed and therefore add little to our understanding of the class performance.

Question Structure

In several areas of study, we are now in the third and fourth iteration of response gathering for particular questions. This iterative process of using student responses to refine the question has allowed us to gather data better suited to lexical analysis. One key lesson is that multiple response boxes (in which students can enter text) are very useful for questions with

multiple components. For example, one original prompt investigating student understanding of general chemistry asked students to "Explain the difference between a weak and strong acid." Students were presented with only a single text box in which to respond. Our analysis of responses to this question quickly found that the majority of students were defining both strong and weak acids, instead of keeping their answer focused on the difference between them. This complicated the lexical analysis as the software cannot determine whether the extracted chemistry terms were being applied for a description of strong or weak acids or both. For example, here are two student responses:

"When mixed into a solution, weak acids will completely dissociate, where as when strong acids are mixed into a solution, they will not be affected."

"A weak acid is one that doesn't dissociate completely and a strong acid dissociates completely in a solution."

Obviously, these two students are describing quite different behaviors for strong and weak acids in solution. However, because both responses use similar extracted terms ("dissociate", "not/doesn't", "completely") the software cannot distinguish between these responses, which can pose a significant problem for evaluating student answers. Also, we found that about 25% of the students included examples of strong and weak acids in their response, which we did not expect because we had not asked them to do so in the original question. In addition, these examples further complicated analysis due to the fact that some students gave an incorrect molecule as an acid example (e.g., H_2SO_4 as a weak acid).

To address these problems, we devised a new response set-up to this question. We now have created four independent text boxes in which students enter their responses. Thus, in this revised question, students are instructed to give, in separate text entry fields, the example of a

strong acid, the explanation of strong acid, the example of weak acid, and the explanation of weak acid. This change in response format has significantly improved categorization and potential for computer rating prediction (Haudek et al., 2009), and illustrates a subtle but significant insight we have attained in question design.

Question wording.

Analysis of student responses has prompted us to re-word questions to make them more amenable to lexical analysis. A critical lesson is that any word in the stem of the question will most likely be repeated by students in their responses. One way of forming appropriate questions is to craft an "ideal" answer that an instructor would like/expect to see and identify the critical terms in the response. One can then devise a question that prompts such a response without using any (or as few as possible) of the target terms. Decisions about what scientific terminology to use and accept in responses should also be made in advance. For example, in one question about free energy, the terms "spontaneous" and "non-spontaneous" appeared in the question stem. However, we decided that the terms "exergonic" or "endergonic" reactions (the chemical term equivalents of spontaneous and non-spontaneous) were meaningful and should be categorized.

Care must be taken when using words in the stem, such as modifiers (adjectives and adverbs), that may cue students in their responses. For instance, one of our questions contained the word "rapidly". We found that many responses contained synonyms of rapid (e.g. "quickly", "fast") but also related terms such as "slow", "rate", "kinetics", etc. This leaves the scorer to make a difficult distinction between terms in the response that reflect the student's knowledge and terms that prompted by a question cue. This becomes even more critical if judging the quality of a student's answer is dependent on how they use these specific terms.

Because repetition of stem words (and related words) in responses can add extra work and confusion to the lexical analysis, we have found that shorter, more direct question stems often produce categories that reflect student responses more accurately. Alternatively, we have also tried the approach of creating a category within the lexical analysis of "Question stem words". Although we thought this would be a reasonable approach, we found that this solution is problematic in that responses that include *only* stem words (which are generally more indicative of a poor- or non-responses) do not show up as *uncategorized* (see the section Number and Specificity of Categories below). Because of this, our current strategy relies mainly on careful wording of the question stem.

Response length.

We have performed lexical analysis on a variety of responses from a range of questions. Expected responses to these questions range from a single word (or molecular formula) to several sentences in length (Urban-Lurain et al., 2009). In general, responses should be of sufficient length (but not too long) to provide enough extracted terms to allow a more accurate categorization. Responses that are too long will include too many unrelated terms. This leads to categorization of too many responses in a given category, which reduces their discrimination for classification purposes. However, responses that are too short may only be categorized in one or two categories. This is problematic if an instructor is interested in the connections students are making between concepts, and thereby categories.

In the simplest questions we have used, students were prompted to supply a molecule (name or formula) in response to a question "provide an example of a strong acid" (Haudek et al., 2009) or a question that asked students to trace carbon in cellular respiration (Moscarella et al., 2008). These single word/molecule responses are easy to classify into categories appropriate for

the question. In most cases, this means a single response resides in a single category. However, it is possible to create a heirarchy of categories in which single responses could belong in multiple categories. In such simple questions, relatively few students (about 1-2%) supplied multiple molecules in a single response box and none supplied written statements explaining their choices, although there were several "I don't know" and blank responses.

In our experience, directed explanations generally result in the majority of responses (>95%) consisting of one to two sentences. For example, one question asked students to give an explanation of a strong acid. Answers of this length generally allow an average response to be placed into two to three categories. Obviously, more simple and/or less informative responses are more likely to be placed into a single category.

In recent work, we have been collecting student responses to diagnostic question cluster-type questions. Such question prompts are rather broad, asking students to explain large, organismal phenomena. For example, students were asked to explain what happened to someone's lost weight on a diet (Wilson et al., 2006). A typical student answer was about three sentences (see Figure 2), although a significant number of students wrote complete explanatory paragraphs of longer length, and some wrote only short sentence or phrases. A response of average length was placed into about four categories. The connections between categories allow an instructor to see that students often harbor multiple conceptions of a phenomenon and that choosing a single foil from a multiple choice instrument probably does not reflect the true complexity of student understanding (Ha & Cha, 2009; Nehm & Schonfeld, 2008). Of course in sharing responses between categories, category granularity is also an important factor (see the section Number and Specificity of Categories below).

Building and Sharing Custom Libraries

We have learned two important lessons about building custom libraries. First, it is easier if misspellings have been corrected before analyzing the data, rather than trying to include various misspellings in the libraries. Second, verbs are not extracted by the software by default, so they have to be included in custom libraries. When verbs are added into a library, term extraction is more efficient if their inflections are also included, so that all inflected forms of the words will be extracted. For example, if we are interested in extracting the verb *evolve*, we may decide that will be the term extracted and include as synonyms *evolves*, *evolving*, and *evolved.*

Sharing custom libraries imposes the challenge of keeping them updated, by including all changes that different users may have introduced. To do so, we keep unmodified versions of our custom libraries (i.e., Metabolism libraries) and they are updated as needed. Users of these libraries are asked to not change them. Instead, any modifications necessary for particular projects are saved in a local library. These unique local libraries compared with the appropriate master custom library. Somebody in our group periodically merges changes from the local libraries to the master libraries and the updated versions of the master libraries are made available for all users.

Number and Specificity of Categories

One key task of text analysis is creating *categories*. The software offers two options, one based on frequency of occurrence of *terms* or *types*, and the other based on linguistic analysis using *semantic networks*, *term co-occurrence* and term *inclusion* and *exclusion*. We have found that the combination of the two procedures provides the best result. We begin creating categories guided by an expert answer, which is used to anticipate possible responses that will be given by students. We then refine the categories based on samples of student responses. Part of

this process is deciding on the level of granularity for categories. Categories should be specific enough to identify conceptions at an appropriate level to distinguish among concepts typically held by the student population of interest. Although there is no rule about how many categories should be created, we have tried to keep it fewer than 40. Fine-grained categories can be easily collapsed, if required by further analysis, while broad categories are more difficult to disaggregate. For instance, in our metabolism project (Moscarella et al., 2008), where we asked students to trace matter and energy during cell respiration, we created 25 categories for the component of the question asking how the carbon got into the CO_2 molecule, including relevant processes in cell respiration, such as respiration, Krebs cycle, glycolysis, photosynthesis, etc. These processes can be easily collapsed into a broader *metabolism* category if a reduction of variables is required by the statistical analysis. Finally, we have learned that in any *de novo* set of data analyzed by existing custom libraries and categories, typically 10% or less of student responses are uncategorized by the lexical analysis software. However, our experience has shown that most uncategorized responses do not contain relevant information or are a reiteration of words used in the question stem and therefore add little to our understanding of student thinking.

Based on our statistical classifications of student responses to predict expert ratings, we have learned that we generally create more categories during the lexical analysis phase than are needed to predict expert ratings. For example, in the strong/weak acid question (see the section Statistical Classification Techniques; Haudek et al., 2009), the lexical analysis produced 27 categories, but only twelve were needed to predict expert ratings with 84% accuracy. However, because we do not know when creating the categories which will be needed to predict expert

ratings, it is better to create more categories with finer granularity and use data-reduction technique in the statistical analysis phase of the work.

Granularity of Expert Rubrics

Much as we are learning about optimizing the granularity of categories in lexical analysis, our results also highlight the challenges of expert rating of student writing. It is often a lengthy process to develop the rubrics and train raters to ensure acceptable inter-rater reliability. Lexical analysis can help provide additional data for the iterative nature of this process. In addition to raters discussing agreements and disagreements about scoring, using the categories derived by lexical analysis for statistical classification can highlight areas where the rubrics may be vague, or unnecessarily specific. For example, when initially created a scoring rubric for student explanations of the differences between strong and week acids, we used a 4-level rubric for the raters. However, we discovered that the statistical classification was least stable between the middle two categories. Further discussion among the raters revealed that they did not believe that the distinctions between levels 2 and 3 were that clear. By reducing the number of categories, we could improve both the interrater reliability and the statistical classification accuracy. When using these techniques for formative assessment, often a binary (generally correct, generally incorrect) rubric may be sufficient for instructors to gauge the overall understanding of a concept to determine instructional intervention.

Although our research has produced computer-expert inter-rater reliability comparable to expert-expert inter-rater reliability, we do not believe that these techniques should used for high-stakes, summative assessments without human oversight. We see the strength of these techniques as providing formative feedback to the instructor and students about student understanding and revealing student conceptual barriers. We have used these techniques to

analyze writing in online homework. We ask students to answer one or more questions before class. We then analyze the responses and summarize common themes for the instructor before class. Based on these data, the instructor can address misconceptions through any number of instructional activities during the class.

Conclusion

Improving STEM education requires timely and informative assessment of student understanding. While constructed responses can provide good insight into student thinking, they are generally too labor intensive to be used in large enrollment courses. Computerized lexical analysis holds the promise to support using constructed response assessments in large enrollment courses. However, application of these techniques has only recently become feasible with the introduction of software that can be used without the help of linguistics experts. By sharing lessons learned and evolving best practices, we can collectively move research and understanding of student thinking forward.

More information, example questions, and the libraries we have created for SPSS Text Analytics is available at the Automated Analysis of Constructed Response (AACR) research group web site: http://aacr.crcstl.msu.edu.

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.

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

Duit, R. (1995). A constructivist view: a fashionable and fruitful paradigm for science education research and practice. In L. P. Steffe & J. Gale (Eds.), *Constructivism in Education* (pp. 3-15). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Ha, M., & Cha, H. (2009, April 17-21). *Pre-service teachers' synthetic view on Darwinism and Lamarckism.* Paper presented at the National Association for Research in Science Teaching Conference, Anaheim, CA.

Haudek, K., Moscarella, R. A., Urban-Lurain, M., Merrill, J., Sweeder, R., & Richmond, G. (2009, April 17-21). *Using lexical analysis software to understand student knowledge transfer between chemistry and biology.* Paper presented at the National Association of Research in Science Teaching Annual Conference, Garden Grove, CA.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). *How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans.* Paper presented at the Annual meeting of the Cognitive Science Society, Mawhwah, NK.

Larochelle, M., & Bednarz, N. (1998). Constructivism and education: beyond epistemological correctness. In M. Larochelle, N. Bednarz & J. Garrison (Eds.), *Constructivism and education*. Cambridge, U.K.: Cambridge University Press.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39-41.

Moscarella, R. A., Urban-Lurain, M., Merritt, B., Long, T., Richmond, G., Merrill, J., et al. (2008, March 30 - April 2). *Understanding undergraduate students' conceptions in science: Using lexical analysis software to analyze students' constructed responses in biology.* Paper presented at the NARST 2008 Annual International Conference, Baltimore, MD.

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.

SPSS. (2009). *SPSS Text analysis for surveys 3.0 user's guide*. Chicago, IL: SPSS, Inc.

Urban-Lurain, M., Moscarella, R. A., Haudek, K. C., Giese, E., Sibley, D. F., & Merrill, J. E. (2009, October 18-21). *Beyond multiple choice exams: Using computerized lexical analysis to understand students' conceptual reasoning in STEM disciplines* Paper presented at the Frontiers in Education, San Antonio, TX.

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.

Wilson, C., Anderson, C. W., Heidemann, M., Long, T., Merrill, J., Merritt, B., et al. (2006). Assessing students' ability to trace matter in dynamic systems in cell biology. *Cell Biology Education, 5*, 323-331.