Construction of Rubrics to Evaluate Content in Students’ Scientific Explanation Using Computerized Text Analysis
Kevin C. Haudek, Rosa A. Moscarella, Michele Weston, John Merrill and Mark Urban-Lurain
Michigan State University

Paper presented at the annual meeting of the National Association for Research in Science Teaching (NARST), Chicago, IL, April 11-14, 2015.

Author Note
This work is supported by the National Science Foundation (DUE grants: 1438739, 1323162, 1347740, 0736952 and 1022653). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.
More information about the Automated Analysis of Constructed Response research project at: www.msu.edu/~aacr.

Correspondence may be addressed to Kevin C. Haudek (haudekke@msu.edu) at Department of Biochemistry and Molecular Biology, Michigan State University, 603 Wilson Rd., East Lansing, MI, 48824.
Abstract

A major challenge to using constructed response items in high enrollment undergraduate STEM courses is the ability to evaluate the scientific content of student explanations, which can be aided using evaluation rubrics. However, traditional rubric development is highly qualitative and is often an iterative process and requires subject expertise and the ability to identify emerging themes from student writing. Here, we report on leveraging the results of lexical and statistical analysis of students’ writing to develop a rubric to evaluate scientific content contained in student responses to two questions used in undergraduate biology courses. Lexical analysis was used to identify and categorize relevant content in student responses. These lexical categories were used as variables in K-means clustering, which helped identify emergent themes or patterns across student responses. Each cluster was defined by categories that included relevant disciplinary content. These clusters were used as the basis of an initial rubric. Several raters applied this rubric to a subset of student responses and after revisions to the rubric and scoring iterations, achieved varying levels of inter-rater reliability (from 0.4 to 0.8 Cohen’s kappa) on different rubric criteria. We believe this methodology may be broadly useful for reducing the effort necessary for rubric creation.

Keywords: lexical analysis, rubric development, text analysis
Constructed response, or open response, assessment items require students to construct their own answers to a given question prompt. Many reports note that such constructed response writing assignments are a more authentic task and can provide greater insight into student thinking than similar multiple choice items (Birenbaum & Tatsouka, 1987; Bridgeman, 1992; Moore, 1994). There exists, however, an evaluation barrier to the use of such assessments, especially in large enrollment, undergraduate STEM courses. Two factors may contribute to this evaluation barrier: 1) the creation of an objective scoring rubric with which to evaluate the responses and 2) the amount of time and/or resources necessary to reliably score a large number of responses.

A rubric is a guide for “qualitative rating of authentic student work” (e.g. student writing; Jonsson & Svingby, 2007; Popham, 1997). Generally, rubrics may be classified as either analytic or holistic (Allen & Tanner, 2006; Moskal, 2000). Holistic rubrics try to evaluate the overall quality of the entire work and are usually based on general descriptions of performance or qualities. Analytic rubrics are often at a smaller “grain-size”, and try to analyze the presence or absence of some specific feature within a work. Multiple analytic rubrics may be used together on a single work to evaluate several specific features of interest. These scores may be combined to form a final, overall score (Moskal, 2000). Creation of a rubric to evaluate student writing requires subject expertise and/or qualitative analysis to identify emergent ideas. Rubrics can be formulated to detect scientific content knowledge (both correct and incorrect), as well as many different constructs related to performance and scientific practices (Allen & Tanner, 2006; Montgomery, 2002; Moskal & Leydens, 2000). Rubrics are often refined over multiple iterations of expert application of the rubric and discussion of revisions (Allen & Tanner, 2006). Therefore, rubric creation can be a challenging and time-consuming task, especially when faculty
lack instructional experience in a given course or level or where there is little to no previous educational research about student learning within a specific content area. In such cases, using grounded theory to develop an initial coding scheme allows the exploration of concepts used by students in their writing.

Using a grounded theory approach to create an emergent coding scheme which identifies student conceptions in undergraduate STEM learning is an iterative process. A foundational idea in grounded theory is that a coding scheme should emerge out of observation, in this case, actual student writing (Trochim & Donnelly, 2007). In this way a wide variety of codes may be produced, from very expert-like concepts and ideas to numerous alternative or irrelevant conceptions. These codes can later be condensed, merged or otherwise revised during iterative rounds of coding or between different coders (Trochim & Donnelly, 2007).

The process of developing an emergent coding scheme generally takes multiple rounds of code development and coding of the data by multiple experts. In short, two or more experts independently analyze the data and produce notes about important features in the data or possibly even an initial coding scheme. Then the researchers meet to discuss these notes/coding schemes. After discussion, the researchers agree on which codes to apply for the research question at hand and apply these codes independently. After this scoring is done, the researchers determine reliability between coders and revise the coding scheme if necessary or try another round of independent scoring. Once acceptable reliability has been reached, the coding scheme can be applied to a larger data set (Saldana, 2009; Stemler, 2001).

The development and application of a coding scheme allows grouping of codes into higher-level “categories” or “themes”(Saldana, 2009). This is important when searching for patterns in the data, as patterns may not emerge until a thorough coding scheme is developed and
reliably applied. Looking for patterns in the data may necessitate re-coding and re-categorizing (Saldana, 2009). During re-coding, either data or codes may be lumped or split, depending on the findings of the previous coding (Saldana, 2009). Identification of the larger “categories” or patterns present in the data allows the creation of a scoring rubric, which generally looks for larger-grain sizes than the coding scheme. By using a ground theory approach, the rubric should capture many of the ideas in the student writing, not only “expert”-like ideas. The result of producing such a rubric is powerful in that it allows for both quantitative and qualitative analyses and provides valuable feedback to both instructor and student (Moskal & Leydens, 2000). However, to reach such a state requires an investment of time and resources by researchers.

**Previous work**

Recent work in the Automated Analysis of Constructed Response research group (AACR; [www.msu.edu/~aacr](http://www.msu.edu/~aacr)) has investigated the utility of computerized techniques to aid in the evaluation of scientific concepts in students’ written responses in undergraduate STEM courses (Ha, Nehm, Urban-Lurain, & Merrill, 2011; Haudek et al., 2011). These efforts seek to overcome the evaluation barrier of using constructed response questions in large enrollment courses. These computerized techniques have the ability to produce predictive scoring models that have inter-rater reliability (IRR) measures equal to those between two calibrated content experts (i.e. computer-human IRR is equal to human-human IRR) (Nehm, Ha, & Mayfield, 2012). Producing such a refined predictive scoring model is dependent on the existence of an expert scored data set that the computer model can use as a “training set” from which to build the scoring model (Nehm et al., 2012). Producing such a scored data set first requires the development, use and refinement of an expert validated scoring rubric (Nehm et al., 2010).
In addition, studies utilizing computerized lexical analysis to investigate and categorize content in students’ short (few sentences) written responses have shown high degrees of association between subject content revealed in student writing and face-to-face interviews (Beggrow, Ha, Nehm, Pearl, & Boone, 2013; Haudek, Prevost, Moscarella, Merrill, & Urban-Lurain, 2012; Weston, Haudek, Merrill, & Urban-Lurain, In Press). Conclusions from such studies suggest that short student responses can be a rich source of data to explore student thinking. Content and ideas revealed by these computerized analytic techniques may then be used as the basis for creating and/or refining scoring rubrics.

For this study, we investigated: How can computerized lexical analysis contribute to the development and refinement of scientific content scoring rubrics?

In this paper, we report on our efforts to use mixed methods (computerized lexical and statistical analysis and human coding) to develop a scientific content scoring rubric for two constructed response questions. We wanted to develop scoring rubrics for two questions of interest (acid/base chemistry and cellular genetics) used in an introductory undergraduate biology course and for which we had large numbers of student responses. The initial coding stage, in which key disciplinary constructs are identified, was facilitated by application of computerized lexical analysis software. The results of this lexical analysis were used in exploratory statistical analyses (e.g. clustering) to find emergent patterns in student responses. These patterns were then used as the foundation for creating a scientific content rubric. These initial rubrics were and continue to be refined over multiple iterations with expert scorers, until conceptual rubric bins were finalized and good inter-rater reliability was produced between raters. Specifics of the questions, procedures and scoring rubrics are elaborated below.
Methods

We performed lexical analysis of student responses to a given constructed response item. The results from this analysis were used in clustering algorithms to identify groups of similar responses. For both lexical and statistical analysis, we used IBM SPSS Modeler (v. 15.0). 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 lexical and statistical analyses. We then use the results of the statistical analysis as the basis for rubric creation.

Items Under Investigation and Data Collection

Acid-base chemistry. We used a previously developed item that was designed to elicit acid/base explanations about a biological problem from science undergraduates (Haudek et al., 2012; item shown in Figure 1). Data were collected as a homework assignment from introductory cell biology courses at three large, public, research universities in the USA. Only responses from students who selected the answer of “Both” and provided a written explanation of their answer are included in this study (n=336).

Differences in cell type. We created two constructed response versions of a question based on an item from the Genetic Concept Assessment (GCA; item shown in Figure 2A; Smith, Wood, & Knight, 2008). The two versions differed only in the given cell type (item shown in Figure 2B). In this case, data were also collected through a course management system as a
homework assignment. Data collected came from introductory biology classes at four large, public, research universities in the USA (n=2726). From this entire data set, a subset of 662 responses were used to build the categories necessary for lexical and statistical analysis.

Lexical Analysis

Briefly, the text analysis software function classifies written text into categories based on the appearance of specific terms (words and phrases), which it detects using term libraries. Although the software contains default libraries of terms, these do not recognize most of the lexicon of scientific disciplines and therefore custom libraries are necessary. In this project, we have expanded on lexical resources created in previous work (for a detailed discussion of the analytic techniques see Haudek et al., 2012). Categories, which represent a single homogenous concept, may contain multiple terms or phrases. Categories represent disciplinary relevant ideas and usually represent a single concept. 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 - presence or absence of a category within a given response. The software places each response into zero, one or more categories based on the terms and phrases present in the response.

A. Which of the following human cells contains a gene that specifies eye color?
   A. Cells in the eye.
   B. Cells in the heart.
   C. Gametes (sperm and egg).
   D. Cells in the eye and gametes.*
   E. All of the above.

B. Using your knowledge of genetics, explain how human eye (or brain) and heart cells are different.

Figure 2. A. Original GCA item (Smith et al. 2008). The asterisk (*) marks the most common incorrect answer identified by Smith and Knight (2012). Reproduced with permission of authors. B. Open-ended versions of the original GCA analyzed in this study.
Statistical Analysis

We used K-means clustering as an exploratory statistical technique, to help identify common content “themes” emerging from student responses (for a review of K-means clustering see Jain, 2010). The categories from lexical analysis were used as variables in K-means clustering in order to group similar responses. Each student response is placed into the cluster for which it is closest to that cluster mean (i.e., shortest distance to its centroid), which results in responses in any given cluster being more similar to each other, than to responses in other clusters. The cluster analysis was iterated for values of K=2 to 6, where K is the number of clusters. There are some objections to performing K-means clustering with dichotomous variables and Euclidean distance. However, to circumvent these problems some clustering algorithms represent text documents as vectors, and use encoding values for dichotomous variables or distance metrics that are more appropriate for this binary data type (Cha, 2007; IBM, 2012; Mao & Chu, 2007; Quintela, 2012).

For both items under investigation, one researcher, a content expert in the field, examined the clusters from each iteration (K = 2 to 6) to evaluate for intra-cluster and inter-cluster similarities and whether these clusters were conceptually meaningful for the scientific discipline. These emerging themes were then used to begin to build the rubric bins used for expert scoring.

Rubric Creation and Scoring

For the acid/base chemistry example, random samples were chosen from the entire data set for each scoring round. Each sample was independently scored by two content experts, who met after scoring to discuss scoring discrepancies and modifications necessary to the rubric. After excellent IRR measure (i.e. Cohen's kappa coefficient > 0.8, Landis & Koch, 1977) was established by the two raters on a rubric bin, the raters were considered calibrated for that rubric.
level. Scoring iterations were performed as many times as necessary to achieve excellent IRR on all rubric levels. After the raters were calibrated for all rubric levels, a single rater scored the entire data set.

For the differences in cell type example, the refinement and calibration of rubrics is done by scoring a random sample of 50 different responses (from the entire data set) in each scoring round. For each scoring round, two disciplinary experts scored the responses independently using the agreed upon rubric. The first round of scoring usually helps raters become familiar with both the rubric and students’ answers, and it opens the discussion for refinement of bins based on both the content and learning concerns that an instructor may wish to identify in students’ answers to a particular question.

Findings

Example 1: Acid-base Chemistry in Biology

A recent study has identified undergraduates’ problems explaining acid-base chemistry within an introductory biology context (Haudek et al., 2012; item shown in Figure 1). This study found that some students who selected a correct answer in Part A could not provide a correct scientific explanation of their selection. These students used previously identified mental models of acid-base chemistry (both correct and incorrect) to explain their choice. We are interested in exploring students reasoning for choosing an incorrect choice, in an attempt to identify common alternative conceptions and barriers to learning this necessary content. Therefore, we chose to evaluate written explanations (i.e. Part B) from students that selected the answer of “C. Both”. Our initial hypothesis was that most students were confusing a hydroxyl functional group with a carboxyl functional group, and therefore would explain that a hydroxyl group would have acidic properties.
**Lexical analysis.** Existing lexical resources developed from a previous project (Haudek et al., 2012) were applied to this data set and revised as necessary. This resulted in the generation of 46 lexical categories in order to capture all relevant content from the selected responses. Each of these categories represents a single, homogenous concepts as it relates to the question (e.g. “base”, “lower pH”, “solution”, etc.) We used the lexical categories as variables in clustering models. We have found that the results from lexical analysis replace much of the initial read-through and preliminary coding steps of traditional qualitative rubric creation.

**Statistical analysis and rubric creation.** We have found that the preliminary results from K-means cluster analyses can be useful for identifying emerging themes from large number of responses by grouping responses with similar ideas together. One researcher, a content expert in the field, examined the clusters from each iteration to evaluate for intra-cluster and inter-cluster similarities and whether these clusters were conceptually meaningful in the context of the question. These emerging themes were then used to begin constructing the rubric used for expert scoring. We have used the results of the statistical exploratory techniques to replace the code grouping (finding patterns) in traditional qualitative rubric creation (Saldana, 2009; Trochim & Donnelly, 2007)

The researcher settled on four clusters: three clusters with seeming content relevance to the question and a fourth cluster of responses containing mostly miscellaneous, infrequent and irrelevant ideas (in our experience, a frequently required scoring bin). The three content relevant clusters demonstrated good intra-class similarity, while the miscellaneous cluster did not. The three content clusters that formed the basis of the initial rubric were named by the researcher as “Both basic”, “One basic, one acidic” and “Balancing”, as these were the apparent emerging
themes from these clusters (see Table 1; rubric titles are italicized throughout the text). The determining text analysis categories had good content alignment with important disciplinary

Table 1. Scoring rubric description and examples for acid-base chemistry responses.

<table>
<thead>
<tr>
<th>Bin title</th>
<th>Categories in clusters&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Description</th>
<th>Example</th>
<th>Kappa&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Present&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Both basic</td>
<td>hydroxyl group, amino group, basic, hydrogen, accept hydrogen, raise pH, lower pH, solution, acidic, donate hydrogen, ionization</td>
<td>Response explicitly states that both compounds are basic or describe basic behavior (e.g. accept protons) for both molecules</td>
<td>Both the amino and hydroxyl compounds will act as a base so they will have the same effect on the pH</td>
<td>1.0</td>
<td>86</td>
</tr>
<tr>
<td>2: One basic, one acid</td>
<td>hydroxyl group, amino group, basic, acidic, acidic and basic, raise pH, lower pH</td>
<td>Response explicitly states that one compound (either hydroxyl or amino) is acidic and that the other compound is basic. Response can also describe the chemical behavior of acids/bases (e.g. donating/accepting protons)</td>
<td>the hydroxyl group would act in an acidic way and lower the pH, where as the amino group would act as a base</td>
<td>0.98</td>
<td>125</td>
</tr>
<tr>
<td>3: Balancing</td>
<td>basic, equal effect</td>
<td>Response contains one of 3 descriptions: 1. both compounds have an effect and response explicitly indicates that effects are about equal and balance or cancel out 2. one compound increases pH and other compound decreases pH equally or by the same amount. 3 &quot;neutral&quot; as a description of the end-product of a reaction with opposite direction of change in pH.</td>
<td>I chose this answer because aminos are acids and hydroxyls are bases. They both will have the same effect because they will cancel each other out.</td>
<td>0.84</td>
<td>15</td>
</tr>
<tr>
<td>4: Charge</td>
<td>N/A&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Response explicitly states that one or both compounds carry a charge or activity results in charged compound in cytoplasm. Non-specific references to ionization are not sufficient; also H+ (and OH-) is excluded from this bin.</td>
<td>Well both molecules OH and NH2 have a negative charge so they would have the same effect on the pH. Both could lose an H+ which would make the cytoplasm more acidic.</td>
<td>1.0</td>
<td>20</td>
</tr>
<tr>
<td>5: Both acidic</td>
<td>N/A&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Response explicitly states that both compounds are acidic or describe acidic behavior (e.g. donate protons) for both molecules</td>
<td></td>
<td>0.92</td>
<td>15</td>
</tr>
</tbody>
</table>

Notes: <sup>a</sup> Categories shown are present in more than 25% of responses in each cluster; <sup>b</sup> n=50; <sup>c</sup> n=336, scored by a single rater; <sup>d</sup> N/A. These rubric bins were created by expert scorers during rubric revisions, not by text analysis.
ideas in each cluster (Table 1). For example, the categories “basic”, “acidic”, “basic and acidic” (where both terms co-occur within a response), “raise pH” and “lower pH” were frequent in the One basic, one acidic cluster; these categories all relate to classification or description of basic/acidic behavior. The outcome of K-means clustering places each response into only one cluster; however, the two raters using the rubric agreed to use these clusters as an analytic rubric instead, in order to better identify mixed ideas, as responses could belong to more than one rubric.

**Rubric application and revision.** Two content-expert raters applied the initial 3-criteria rubric (shown as rows 1-3 in Table 1) to a subset of 50 responses. The goal of the two independent raters was to establish excellent IRR for each rubric level. Although both experts did score responses using the rubric, both experts viewed this initial scoring round largely as an exploratory opportunity, to get accustomed to both the rubric and student responses. After completing the scoring independently, the two raters met to discuss and revise the scoring rubric. Both raters agreed: 1. the three initial scoring criteria were valid and frequent in student responses and should continue to be used; 2. it would be helpful to jointly find examples of actual student answers to better qualify the explicit vs implicit meaning and 3. another analytic criterion, Charge, was necessary to capture emergent ideas and improve the rubric. The additional criterion, Charge, was added to help capture responses that were not scored in the other criteria. Responses in this criterion did not use acid/base ideas as explanations, but instead used an ionic charge as the basis of explanation.

The two raters applied this new rubric to another sample of 50 responses. After this round of scoring, the two experts agreed that a fifth scoring criterion would be useful. This added criterion, Both acidic, was relatively infrequent (~5%) in student responses within the
sample, but was deemed a necessary contrast to the initial, existing criteria (Table 1, rows 1 and 2) that described acid/base behavior of the functional groups. It is likely that due to the infrequency of this code that such a theme did not emerge out of initial statistical classification attempts.

The two raters scored another sample of 50 responses using this revised 5-criteria rubric. They reached excellent Kappa levels for four of the five criteria, with only Balancing being below 0.8. The raters agreed that the five criteria seemed sufficient to capture necessary content. The raters discussed scoring disagreements and narrowed the Balancing criterion by outlining three acceptable descriptions (see Table 1). Each of these descriptions was exemplified by an actual student response. After discussion, a new random sub-set of 50 responses was re-scored by both raters for the Balancing criterion only, with a resulting Kappa of 0.84. After this round was complete, IRR was sufficiently high in all criteria that one rater proceeded to score the entire data set alone (Table 1). Some responses, with little/no relevant content, were not matched to any of the five rubric criteria.

**Example 2: Differences in Cell Type**

Nowadays, the general population is regularly exposed to genetic concepts through TV shows, movies, or debates in the news concerning genetically modified organisms or personalized medicine, to mention just a few; however, this does not necessarily contribute to a better understanding of genetics by the public because information in the media is not always correct or accurate (Bowling, Huether, et al., 2008; Kılıç & Sağlam, 2013; Knippels, Waarlo, & Th Boersma, 2005; Lanie, 2004; Marbach-Ad, 2010; McElhinny, Dougherty, Bowling, & Libarkin, 2014; Muela & Abril, 2013; Petty & Christensen, 2010). Genetics is widely recognized as a fundamental area of Biology where students have the greatest conceptual difficulties
perhaps mostly due to the linguistic challenges in teaching genetics (Thörne, 2012). Students in the classroom, including biology majors, struggle to understand some basic concepts, like the nature of genetic information in the cells of an organism (Smith & Knight, 2012; Wood-Robinson, Lewis, & Leach, 2000; Lewis, Leach, & Wood-Robinson, 2010). Such understanding is fundamental to Biological literacy and is related to at least 3 of the 5 core concepts proposed in Vision and Change: evolution; structure and function; and information flow, exchange, and storage (AAAS, 2009). Both the Genetic Literacy Assessment Instrument (GLAI; Bowling, Acra, et al., 2008) and the Genetic Concept Assessment (GCA; Smith et al., 2008) which are instruments created to assess core concepts in genetics, contain questions that evaluate student’s understanding of the genetic characteristics of the different cells in an organism. A common misconception not only found in the general public but also among students at different levels is the belief that cells in different tissues (e.g., brain and heart) within the same organism have different genes (Lewis et al., 2010; Petty & Christensen, 2010; Smith & Knight, 2012; Wood-Robinson et al., 2000). In fact, Smith and Knight (2012) found that this idea is hard to overcome, even after instruction. In a study using the GCA to investigate persistent conceptual difficulties in undergraduate genetics classes, Smith and Knight found that more than 20% of students chose an option that implies that a particular structure (i.e., eye) has unique genes related to its function or appearance (i.e., genes that determine eye color) (Figure 2).

Introductory biology courses may be the only reliable source of genetic content to which students will be exposed during their time in college, or even beyond. A major objective of our research group is to identify conceptual difficulties in core concepts in Biology, and the findings by Smith and Knight are used as the basis of our inquiry for this particular question. We wanted
to investigate if this conceptual difficulty could be exposed and if other difficulties could be
distinguished in student writing. Because constructed response questions allow students to craft
explanations that may more accurately reflect their mixed ideas, we created a constructed
response version of the GCA item under investigation.

**Statistical analysis and rubric creation.** Student written responses to both versions of
the constructed response question were used to develop lexical resources, resulting in 23 lexical
categories. Initial exploratory statistical analysis conducted by a content expert produced three
clusters, revealing three major themes in students’ answers: cells are different because they have
1- *different genes*, 2- *different expression*, or 3- *different physiology/function*. The cluster
analysis confirmed the “different DNA” misconception, which is the focus of the distractors in
the original MC question in the GCA. However, constructed responses also revealed an
additional explanation that was not considered by neither the GLAI nor the GCA, i.e. different
physiology/function. In this cluster, students focus on describing the functional differences
between cell types and not on the genetic differences. Based on these three themes found in the
cluster analysis, we developed a 5-bin holistic scoring rubric (Table 2). We decided to add two
additional bins to the rubric in hopes of classifying students that had “mixed models” (i.e.
expressed more than one of the identified themes) and “cannot resolve”, to classify students that
had zero of the identified themes.

**Rubric application and revision:** Two content experts used the rubric to score a subset
of 50 responses. Afterwards, they met to discuss scoring disagreements and refine the rubric. In
the second iteration of the rubric, the *mixed model* and *cannot resolve* bins were merged and a
new round of scoring was done on a new set of 50 answers. IRR was calculated, showing a low
agreement between the two content experts (Table 2, 1st scoring round column). Most of the
disagreements were on the bins *contain different genes and cannot resolve*; raters discussed and revised the criteria for those bins. A second scoring round showed little improvement, if any.

**Table 2. Holistic rubric progression and scoring results for differences in cell type responses.**

<table>
<thead>
<tr>
<th>Original holistic rubric</th>
<th>1st scoring round</th>
<th>2nd scoring round</th>
<th>3rd scoring round</th>
<th>ICCb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same genes, differential expression</td>
<td>Same genes,</td>
<td>Same genes,</td>
<td>Same genes,</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>differential</td>
<td>differential</td>
<td>differential</td>
<td></td>
</tr>
<tr>
<td></td>
<td>expression</td>
<td>expression</td>
<td>expression</td>
<td></td>
</tr>
<tr>
<td>Contain different genes</td>
<td>Contain different</td>
<td>Contain different</td>
<td>Contain different</td>
<td></td>
</tr>
<tr>
<td></td>
<td>genes</td>
<td>genes</td>
<td>genes</td>
<td></td>
</tr>
<tr>
<td>Physiological or functional</td>
<td>Physiological</td>
<td>Physiological</td>
<td>Physiological</td>
<td></td>
</tr>
<tr>
<td></td>
<td>or functional</td>
<td>or functional</td>
<td>or functional</td>
<td></td>
</tr>
<tr>
<td>Mixed Model</td>
<td>Cannot Resolve</td>
<td>Cannot Resolve</td>
<td>Cannot Resolve</td>
<td></td>
</tr>
<tr>
<td>Cannot Resolve</td>
<td>0.51</td>
<td>0.12</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>-0.08</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: aCohen’s kappa; bInterclass correlation for the scoring of three experts

Actually, the *cannot resolve* bin proved to be too subjective, as raters disagreed on whether an answer could not be resolved indeed or a concept was actually implicit (e.g., gene expression), getting a lower Kappa coefficient in each iteration (0.505 vs. 0.122). Raters met again, discussed the criteria for each bin and refined the rubric, adding the bin *miscellaneous* to accommodate responses with no relevant content. This third round of scoring still found little agreement among the experts in all but one bin (Table 2, 3rd scoring round column). The only bin with an improved score was *different genes*, while the remaining got worse. We all agreed that an analytic rubric would make the scoring task easier because it would reduce ambiguity and/or subjectivity. We created fine-grained bins in an attempt to identify the major sources of disagreement, i.e. whether *same genes* concept was implicit or explicit in the answer (Table 3). The first round of scoring using the new analytic rubric resulted in a better IRR, with two bins getting excellent agreement (*same gene [explicit] and physiology/function*; Kappa > 0.8), two
very good (same genes [implicit] and [stem] cell differentiation; $0.7 < \text{Kappa} < 0.8$), and two that need improvement (different genes and different gene expression; $0.4 < \text{Kappa} < 0.6$). In the last case, the main cause of disagreement was because we did not have clear instructions on how
to score answers that contain both *different gene expression* and *different genes*. Raters met and clarified the criteria for scoring and the rules for co-occurrences (see Table 3).

Raters also agreed that only answers with clear content would be used for building the scoring model, as some answers are poorly written making it difficult to objectively interpret what the student said. At the time this paper was written, this rubric was still under development. We will report the current status at the NARST presentation.

**Discussion**

**Coding and Rubric Development**

We attempted to leverage the output of computerized analytic techniques to build a scientific content scoring rubric. We were successful in that two raters agreed that the rubric identified common ideas in student responses and were able to achieve high levels of IRR using the rubric. In addition, the text analysis categories prominent in the initial clustering represented relevant content to underlying scientific concepts. We were able to replace much of the initial qualitative, exploratory work traditionally done in rubric development with lexical analysis and clustering, thus reducing the time to develop a scoring rubric. The initial lexical analysis replaced many of the early qualitative steps of normal emergent coding, including reading of large number of responses, initial coding and code refinement. Lexical resources, developed for another project were applied with minimal revision to the data set under investigation. This reduced greatly the time in developing and applying a coding scheme. In this way, the lexical categories substituted well as “codes”, in traditional content analyses. We believe a key caveat for this approach is that the lexical categorization should be relatively fine-grained, where each lexical category represents a single homogenous content-specific concept.
The exploratory statistical technique helped find larger patterns in the data, using the lexical categories as variables. Clustering algorithms identified three common themes in the student writing, replacing much of the “pattern finding” or code categorization of traditional qualitative coding. Using these common clusters as themes, allowed the raters to consider infrequent, but still important, ideas during their initial scoring rounds instead of initially concentrating on coding scheme. Raters agreed that the initial cluster-based rubric was improved through typical rubric development practice, such as iterative revision of criteria and selecting examples.

In addition, using computerized text analysis for much of the categorization and coding scheme allowed easy iterations if changes were deemed necessary to the text categories or exploratory clusters (i.e. “re-coding”) (Saldana, 2009). Re-coding was performed through simple manipulations in the software. For example, researchers could easily collapse multiple lexical categories into a single category, or calculate clusters using a subset of variables. Determining the results of such changes took only a matter of a few minutes, after which the re-coding could be adopted or revised again.

**Acid-base Chemistry in Biology**

Results of this research have implications for the discipline-specific content (biology, chemistry) understanding of undergraduates. For the acid/base question, we uncovered little evidence to support our initial hypothesis. Student confusion about hydroxyl and carboxyl groups could be identified in student writing; however, this confusion is exhibited explicitly (i.e. text about carboxyl functional groups in responses) in a much smaller percentage of responses than expected (<2%). However, about 1/3 of the responses were scored as *One basic, one acidic*, which would be the expected scoring result for students that hold both correct ideas about amino
groups and confusion about hydroxyl and carboxyl groups. However, because many students were not explicit about why they believe one molecule acts as an acid, further interviews would be helpful to understand this confusion. In addition, not all students were clear about which molecule was acting as an acid or base, allowing the possibility that some would suggest the amino functional group acts as an acid if prompted during an interview.

We were surprised to that nearly $\frac{1}{4}$ of the students were scored as *Both basic*, as we didn’t expect this type of explanation would be so frequent for students that selected this multiple choice option. We now believe this may be an indication of *heterogeneous thinking* in these students: a student knows some correct idea (e.g. amino functional groups can act as a weak base) while still maintaining incorrect ideas about the same or similar topic (e.g. hydroxyl functional groups act as a strong base). This inference seems likely, since it has been reported that students have difficulty distinguishing between molecular and ionic forms of OH (Furio-Mas, Calatayud, & Barcenas, 2007; Haudek et al., 2012). It seems this finding adds another example to the growing literature that closed-form assessments, such as multiple-choice, result in a forced “either-or” item choice: Students are forced to decide between choices that don’t accurately reflect their thinking. In such cases using constructed response items allows students to demonstrate some “correctness” in their written explanation which is hidden by viewing their multiple choice selection alone. Students’ constructed responses allow this *heterogeneous thinking* to emerge and this thinking can be identified through student writing using computerized lexical analysis and scoring techniques (Ha et al., 2011; Haudek et al., 2012). Again, follow-up interviews are necessary to test these hypotheses and to further probe student thinking in these type of explanations.
Differences in Cell Type

For the cell type question, the application of a constructed response item provided information about students’ understanding of this genetic topic that neither of the equivalent MC questions in the GCA nor the GLAI would have revealed. Because neither of the two versions of the AACR question prompts to include gamete cells in the answer, which would explain why none of the responses referred to those cell type. The responses observed in the AACR question are better aligned with the GLAI choices (item #22; Bowling, Acra, et al., 2008); however, when students are asked to think of the genetic differences between two cell types, in an open-ended format, it is revealed that some students think cells are different because they have different functions/physiology, and not that their different functions is due to their genetic characteristics (as it is asked in the GLAI item). Also, although it is known that some students think that different cell types have different genes (Wood-Robinson et al., 2000; Smith & Knight, 2012), there is not information that suggests why students think so and what has made them harbor this wrong idea. The analysis of students’ written answers to these questions would allow instructors better understand what conceptions students have and find ways to address misconceptions during instruction (Prevost, Haudek, Norton-Henry, Berry, & Urban-Lurain, 2013). Although scoring using a holistic rubrics may take less time, analytic rubrics are easier to calibrate for IRR, as it has been already noted (Klein et al., 1998). For a question like the cell type, where students can display different levels of knowledge and writing proficiencies, scoring an answer simply as correct/naïve/wrong is a challenging task. In this case, analytic rubrics not only make it easier for raters to agree on the scoring of the specific content in an answer, but also allow uncovering in greater detail the ideas students have, rather than just scoring their answers right, wrong, or somewhere in between. We believe analytic rubrics provide better insight of what students know
and do not know, so instructors can use this information to address conceptual issues during instruction (Popham, 1997). For instructors that are also interested in the overall performance of their students, these analytic bins can be combined in holistic models based on the co-occurrence of specific bins in the students answers (Moharreri, Ha, & Nehm, 2014; Nehm, Kim, & Sheppard, 2009) which is an additional advantage of using analytic rubrics.

The creation of these scoring models using text analytics is time consuming; but, this investment of time has the possibility of substantially reducing the required time for scoring new data using the same lexical resources (Weston, Parker, & Urban-Lurain, 2013), which would make feasible the administration of constructed response items to assess students understanding of science.

Because this rubric is still under development, we do not have numbers to compare with Smith and Knight findings regarding the persistency of the misconception that different cell types contain unique set of genes; but our results thus far corroborate alternative conceptions found by both Smith et al (2008) and Bowling et al (2008).

**Future Directions**

We will continue to investigate other exploratory statistical techniques. Specifically, we are exploring using hierarchical clustering methods to identify the emergent themes. One advantage of such an approach is the number of clusters is not pre-defined; however, content expertise is still necessary in order to interpret the emergent “theme” from the clades. In some cases this may be difficult if all the lexical categories (i.e. codes) are used as variables, as some of these categories contain novice, incorrect or irrelevant ideas.

We will continue to refine the scoring rubric used in the Differences in cell type to achieve desired levels of IRR. We will then administer this item as pre and post instruction
homework, in order to later interview students that completed both assignments. We are interested in revealing alternative conceptions students may have, as well as understanding the cause of them, e.g. why they think that cells in different tissues have different genes. We also would like to understand why they provide a physiological explanation(s) when they are being asked for the genetic differences between different cell types. This information should be useful to refine the question prompt or for instructors to target potential source(s) of confusion and clarify them promptly during instruction.

**References**


