Automated Analysis of Written Assessments in STEM: Methodological Issues

Paper set presented at the 2016 Annual International Conference of the National Association for Research in Science Teaching (NARST)

Baltimore, MD, April 14-17, 2016.

Papers included in this set:

*The Development of Constructed Response Astronomy Assessment Items*
Matthew Steele, John Merrill, Kevin Haudek and Mark Urban-Lurain

*Applying Automated Analysis to Develop a Cost-Effective Measure of Science Teacher Pedagogical Content Knowledge*
Molly Stuhlsatz, Chris Wilson, Zoe Buck Bracey, Mark Urban-Lurain, John Merrill and Kevin Haudek

*Automated Analysis Provides Insights into the Challenges to Students’ Understanding of the Processes Underlying the Flow of Genetic Information*
Rosa Moscarella, Kevin Haudek, Jennifer Knight, Alexandria Mazur, Karen Pelletreau, Luanna Prevost, Michelle Smith, Matthew Steele, Mark Urban-Lurain and John Merrill

*Predicting the Accuracy of Computer Scoring of Text: Probabilistic, Multi-Model, and Semantic Similarity Approaches*
Minsu Ha and Ross Nehm

*Discussant Comments*
Jonathan Osborne
Automated Analysis of Written Assessments in STEM: Methodological Issues

Background

Constructed response (CR) assessments, for which students use their own language to demonstrate knowledge, are widely viewed as providing greater insight into student thinking than closed form (e.g., multiple-choice) assessments. In the past, constraints on the time required to evaluate student responses made CR assessments significantly more challenging to execute and evaluate for large number of responses than multiple-choice assessments. Presently, advances in both technology and research into assessment methodology make it feasible to apply these techniques in instructional settings with the potential to have substantial impacts on curriculum design, teaching, and learning (Magliano & Graesser, 2012). Constructed-response questions can potentially reveal not only students’ understanding of specific concepts, but about their ability to explain and build arguments about these concepts.

A recent report by the National Research Council encourages the use of authentic scientific practices in science teaching and learning. Such teaching and learning should focus on these authentic practices consisting of the activities in which scientists engage, such as constructing scientific explanations and arguments about natural phenomena (National Research Council, 2012). As such, we collect student explanatory writing about scientific phenomena as Observations (see Figure 1; NRC Assessment Triangle from Pellegrino, Chudowsky, & Glaser, 2001).

Students learning science often have a mixture of “scientific” and “nonscientific” ideas that reveal heterogeneous thinking, or a mixed model of both scientific and nonscientific ideas (Cognition, Figure 1). Beginning students often learn discrete knowledge pieces that are unconnected, and this emerges when they are unable to build coherent or valid scientific explanations or arguments. Revealing these mixed models of student thinking is important formative feedback for both instructors and students that can lead to positive impacts on student learning, for example, leading to targeted instructor-student conversations (Gerard & Linn, 2016). Developing measures of the composition, structure, and stability of student thinking about core scientific ideas and practices is a challenge which may be difficult to accomplish via multiple-choice assessments alone. These constraints of closed-form assessments can be mitigated by CR assessments that capture student’s explanatory models. Furthermore, learners should
have opportunities to practice constructing explanations and arguments and receive timely feedback, in
order to monitor their own learning (National Research Council, 2012). Additionally, educators need
means to assess whether and how students progress in these desired practices. In order to do so,
researchers and educators need to develop clear criteria for evaluating student performance in these tasks.

A key constraint to the widespread use of CR assessment are evaluation constraints: 1) there must
be set evaluation criteria; 2) CR items often take longer to evaluate; 3) scoring may take multiple raters;
and 4) there may be concerns over objectivity of grading (see for example Liu, Lee, & Linn, 2011). For
some scientific practices and domains, well-developed scoring rubrics exist (e.g. scientific explanation,
argumentation and evolution (Liu, Lee, Hofstatter & Linn, 2008; McNeill, 2011; Nehm et al, 2010). For
other scientific domains with less existing research into student understanding and/or ability, creating,
revising and applying a scoring rubric is essential. One approach to this task is to use grounded theory to
create an emergent coding scheme (Saldana, 2009). However, such an approach takes expertise,
considerable time and large numbers of responses. In any case, once an evaluation rubric has been
chosen or developed, expert hand-scoring of explanations is still required. There are numerous
disadvantages to this hand-scoring approach including time and resource constraints, delayed feedback to
students, and disagreements between raters (Nehm & Schonfeld, 2008).

On the other hand, recent progress in computerized tools and scoring models allow deep and
rapid data exploration, evaluation and feedback in both research and instructional settings (Interpretation,
Figure 1) (Gerard & Linn, 2016; Moharreri, Ha, & Nehm, 2014). Such computerized tools facilitate the
identification of student ideas in their written responses, allow iterative development of emergent coding
schemes, and facilitate the creation of statistical scoring models to reduce the burden of hand-scoring.
The ultimate goal of developing these automated assessment methods is to develop accurate and reliable
scoring models that are able to score students’ scientific explanations at accuracies equal to two human
expert scorers. These tools also reveal the “pieces” of knowledge students use and allow the exploration
of connections between ideas in students’ mental models.

Methodological Overview

In this section, we provide an overview of our approach to developing, validating and
implementing assessments as background. The entire process is captured by the Question Development
Cycle (QDC, Figure 2). In general, we use linguistic methods to extract features from students’ writing,
and then use those linguistic features as variables in statistical models that predict human raters’ scores of
the student’s writing.

In the first stage of the QDC, we design new questions (Figure 2, top right box in figure) to
measure student thinking about important disciplinary constructs. Data collection (Figure 2, middle right
of circle) is typically done by administering the questions using an online system in which respondents can enter their responses. **Lexical resource development** is done using lexical analysis software to extract key terms and scientific concepts from the students’ writing. These terms and concepts are used as variables for **exploratory analysis**, which may aid in **rubric** development (Haudek, Moscarella, Weston, Merrill, & Urban-Lurain, 2015). For projects that have well-defined scoring rubrics and use machine learning algorithms, the exploratory and rubric development stages are largely bypassed. We use both analytic and holistic rubrics, which are designed to capture specific content/ideas or overall quality respectively, for **human coding** of student responses. During **confirmatory analysis**, the lexical resources are used as independent variables in statistical classification techniques to predict expert human coding of student responses. The entire process is iterative, with feedback from the various stages informing the refinement of other components. The final product of the QDC is a **predictive model** that can be used to completely automate the scoring of a new set of student responses. These models are used to predict how experts would score the responses at interrater reliability levels (computer-expert) similar to levels achieved between two well-trained expert scorers. Both during and at the end of the QDC, rich insight into student thinking and explanatory models can be revealed.

**Paper set overview**

In this paper set, we highlight four different projects using or investigating automated analysis of constructed responses. Each paper focuses on a different stage of the QDC and contributes different advancements to methodology or computer scoring evaluation. The goal of each project is to develop a reliable, automated scoring model to evaluate large numbers of written responses.

In “The Development of Constructed Response Astronomy Assessment Items”, Steele et al, describe the process of adopting questions from previously developed concept inventories (in specific, the Light and Spectroscopy Concept Inventory) for use in an undergraduate astronomy course. The paper reports on some early stages of exploratory analysis to reveal common ideas in student responses.
Stuhlsatz et al. report on the development of a scoring rubric for measuring teacher pedagogical content knowledge in “Applying Automated Analysis to Develop a Cost-Effective Measure of Science Teacher Pedagogical Content Knowledge”. The paper reports on the foundations of a pedagogical content knowledge framework which was used to develop the scoring rubric and the results of an initial human coding of responses using this rubric.

Moscarella et al. describe using lexical analysis and statistical scoring models to reveal student issues with the genetic process of transcription in “Automated Analysis Provides Insights into the Challenges to Students’ Understanding of the Processes Underlying the Flow of Genetic Information”. This paper highlights how patterns and issues can be identified using large numbers of computer scored responses by a confirmatory model.

In “Predicting the Accuracy of Computer Scoring of Text: Probabilistic, Multi-Model, and Semantic Similarity Approaches”, Ha and Nehm describe novel approaches to identifying responses scored with low confidence by an automated predictive scoring model in the science domain of evolution. Identifying such responses may lead to improvement in scoring models or provide a means for researchers and/or instructors to examine only “high-quality” data predictions.

**Author Note**

This work is supported by the National Science Foundation (DGE 1438739 and DUE grants: 1323162 and 1347740). 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.
References


The Development of Constructed Response Astronomy Assessment Items

Matthew M Steele, John Merrill, Kevin Haudek, and Mark Urban-Lurain

1 Introduction

Concept inventories comprised of multiple choice questions are widely used in physics and astronomy to measure students pre-instruction knowledge and post instruction gains (Hestenes, Wells, & Swackhamer, 1992; Maloney, OKuma, Hieggelke, & Van Heuvelen, 2001, and others). Concept inventories are valuable because they address big ideas in the discipline with which students often struggle. The multiple choice construction of these instruments, however imposes structural limitations on the number and nature of misconceptions and mixed conceptual models which may be presented as item distractors (Smith & Tanner, 2010). Constructed response questions provide one means of addressing these limitations.

In the effort to develop constructed response questions in the field of astronomy, we have adapted items from well-known concept inventories. These items are comprised of two, three-question sets. By adapting items from existing concept inventories the resulting items are targeted as a supplement to the original instrument, providing an additional perspective with a deeper view on student thinking of the relevant items.

In this work we discuss the selection and adaption multiple choice items as part of the new question design of the AACR Question Development Cycle (appropriate reference), the testing of the resulting items, and a summary of patterns in student thinking that appeared in the body of collected responses.

2 Study Design and Procedure

For this work we focus on a set of questions drawn from the Light and Spectroscopy Concept Inventory (Bardar, Prather, Brecher, & Slater, 2007, hereafter LSCI). The selected items probe student thinking on connections between emission and absorption line features and the physical properties of their source objects. A classical test theory analysis by Schlingman, Prather, Wallace, Rudolph, and Brissenden (2012) revealed a number of items on the LSCI which displayed high post-instruction difficulty values and low post-instruction discrimination
values. For these items students had difficulty answering the question after instruction and their success in doing did not correlate strongly with their score on the entire LSCI. For this study we have selected three high difficulty/low discrimination for adaptation to constructed response items, in order to investigate why students struggle with concepts targeted in the corresponding items.

2.1 Item selection and Adaptation

The items selected for adaptation are LSCI items 2, 17, and 21. The revisions to the text of these items when adapting from multiple choice to constructed response format were left as minimal as possible. Item 2 was adapted by simply removing the provided answer and distractors. Item 17 was reworded to remove the “which of the following format language. For item 21 a prompt for the “what kind of object” was replaced with query for “what physical properties”. The original multiple choice answers provided object descriptions as permutations of the temperature (hot/cool) and density (dense/diffuse). With this construction simply removing the provided answers could allow the students to answer with a category of object (star/nebula/galaxy) rather than focus on the intended object properties. The constructed response version of the question was rephrased to explicitly focus on the source object’s physical properties.

The exact formats of both the original LSCI items and the revised constructed response items may be found in appendix A. Hereafter in this work the constructed response items will be referred to by names that describe the key concept underlying the item, rather than the LSCI item numbers. The constructed response item derived from LSCI item 2 is referred to as the “object color and absorption features” item. The item produced from LSCI item 17 is known as the “emission line wavelength comparison” item. Finally question adapted from LSCI item 21 is called the “emission line sources” item. The items are discussed in the order which they were administered for this study, rather than the LSCI arrangement.

2.2 Item Administration

To test the converted constructed response items, they were administered to in two non-major university astronomy courses. On section was small (N~50) on-line summer session, the other a medium enrollment (N~150) traditional classroom course during the regular academic year. In the online section questions were given both pre- and post-instruction in order to solicit the greatest diversity of student responses. The items given only post-instruction in the tradition section. Both sections saw the items as part of regular weekly homework assignments. In total the response set for the three items contain 43 pre-instruction and 110 – 113 post-instruction responses. The resulting responses are analyzed for their textual content in the following section.
3 Results and Analysis

This section describes the exploratory rubric development procedure and the analysis of student responses of the three items. The rubric development and scoring processes is discussed in subsection 3.1. Each the results of scored student responses to each constructed response in considered individually in the subsequent subsections.

3.1 Rubric Development and Scoring

Student responses to constructed response items can display a wide range of ideas representative of both focused content understanding an loosely collected conceptual associations. The development of exploratory rubric were assembled with the goal of capturing as much student thinking on the items content as possible without excluding or preferring different levels of sophistication. As such the bins in the rubrics are intentionally fine grained, with the goal of coding for the presence of small coherent ideas in the student response. Bins in the rubric are non-mutually exclusive so a single student response may be coded in multiple bins. Under this scheme complex thought will be indicated by the presence of rubric bins which maybe connected to form a chain or reasoning. For example a complete well-constructed response may be coded for a rubric bin that indicates a comparison, another bin that indicates a reason for the comparison. A less coherent student response may then be coded as a collection of bins which correspond with an astrophysical object’s physical properties.

The first step in constructing a exploratory rubric is examine a body of student responses for each constructed response item. The first response in the item set is examined and a rubric bin is defined for each explicit coherent concept it contains. The threshold for coherence is that reader developing the rubric is able to identify and describe the concept with a brief “handful of words” definition. To be viable the concept must be readily identifiable in isolation from the remainder the the response text. Since the rubrics used in study will ultimately be used by a automated machine learning system, ambiguous or implied concepts will not prompt a bin be added to the rubric, since the automated scoring system does not have the ability to interpret implied information. The set of responses is progressed through adding bins necessary as new concepts are encountered. If during the initial assembly of the rubric it becomes apparent that a concept is being used in fundamentally different ways indifferent responses the initial bin maybe redefined to a narrower concept and additional bins added for the new concept use case.

Once the first pass through the response set has been completed the initial rubric is examined to check for bins sufficiently similar that they cannot be reliably distinguished using individual student responses as test cases. Any such similar bins are consolidated. Then the rubric bins are sorted into broad categories such as “comparisons”, “reasoning”, or “physical properties” dependent on the content and/or use of the member bins. The specific type and functions of these categories will vary from item to item depending on the content and structure
of the question. It should be noted that these categories are not explicitly coded for in student responses, they are merely used as a logistic aid for the coding and analysis procedures. The exploratory rubrics can be large, so having a way to organize the bins can speed the scoring process and provide an initial framework for the analysis. The rubrics developed for the three constructed response items considered in this work are presented in appendix B.

The completed rubric is then used to score the student response set which was employed in its development. As in the rubric construction phase a concept must be coherent, identifiable in isolation, and not require inference beyond what the text makes explicit to be scored for a particular rubric bin. The results of the scoring of the test set are given along with exploratory analysis is sections 3.2, 3.3, and 3.3.

The rubrics developed in this work are exploratory in nature. The goal in their construction is both to simultaneously evaluate the constructed response question and for a preliminary picture of student thinking around the specific system the items address. While they are designed to be identify fine grained concepts and minimize the affects of coder interpretation and inference they have not been validated with in depth student interview data nor tested for reliability with multiple coders. As such the results they produce should taken as informative of the range of concepts contained within students responses, but not a robust probe of student understanding of the items content.

3.2 Emission Line Sources

When a bound electron transitions from high energy excitation state one with lower energy state, the atom must release energy equivalent to the difference in the two excitation states. If the atom exists in relative isolation, the most common means of releasing this energy is in the form of a photon. Since the energy levels available to an electron in a atom are quantized, each transition between two given level will result in the production of a photon with identical energy. To the observer a photons energy is measurable through its wavelength. A collection of photons produced by a specific excitation state transition from an element will all have the same wavelength. When viewed through device like a prism or diffraction grating, a collection of photons with identical wavelength appears as a bright line of a specific color. This type of electromagnetic radiation is there for known as emission line radiation, and the object that produces it is an emission line source. In order for an astrophysical systems to produce emission line radiation two conditions must be met; 1) the object must be hot enough for electrons in the objects atoms to exist in an initial high energy state, 2) the object must be diffuse enough such that photon emission can serve as a major mechanism for de-excitation and not be dominated by atom-atom collisional de-excitation. The emission line source item, both in its original LSCI and revised constructed response version is targeted at prompting the students to identify these two conditions.

The rubric for the emission line source constructed response item, adopted from LSCI item 21, is comprised of 15 bins (see table 2). These bins may
roughly be divided into two categories; rubric bins which address the nature and interactions of the astrophysical system, and bins that discuss the physical properties of the source. For illustration the “object type” bin which notes a discussion of whether the emission is produced by a nebula, star, planet or other class of object falls in the astrophysical system category, while the “density” bin covering mentions of the density of the source media is of the physical property type.

The original LSCI multiple choice answers are covered by rubric bin 7 “temperature” and bin 9 “density”. It should be noted that neither the “temperature” bin, or the “density” distinguishes the quantity of the relevant measure in the rubric scheme. A response that discusses ”high temperature” and one that describes “low temperature” will each be coded only for the “temperature” bin. Therefore it is impossible for the exploratory rubric presented here to exactly reproduce the original LSCI item answers. The objective of the exploratory rubric is to test if students identify “temperature” and “density” as the key physical properties which determine if a source will produce emission lines.

The results of the test set of student responses is summarized in figure 1. The reference ID numbers for each bin in the figure are provided in table 2. The upper portion of the figure displays the frequency of rubric bin occurrence in

Figure 1: Frequency of responses for the emission line sources exploratory rubric. Pre-instruction responses are displayed in the upper portion with the post-instruction frequencies in the lower portion. The numeric rubric bin IDs can be found along with descriptions in table 2.
the pre-instruction response set, while the lower portion gives the rubric bin frequencies for the post-instruction responses.

The two rubric bins which correspond to the original LSCI answers, “temperature” and “density” have frequencies of 43% (14% pre, 54% post) and 12% (5% pre, 15% post) respectively. Only 8% (0% pre, 12% post) of student responses were scored for both the temperature and density rubric bins. After temperature the second most frequently occurring bin was “chemical composition” which occurred in 36% (9% pre, 46% post) of student responses.

Responses from the pre-instruction set appeared to be disproportionately focused on concepts grouped in the “astrophysical system” category of bins. Specifically the “light/matter” (30% pre, 20% post), “properties of light” (23% pre, 15% post), and “energy content” (30% pre, 4% post), were over represented.

### 3.3 Object Color and Absorption Features

Hot dense objects like stars produce light through thermal radiation. The energy released in the core of the star gradually migrates to a layer of the star known as the photosphere through a combination of thermal conduction, convection, and photon scattering processes. By the time the energy leaves the
photosphere in the form of photons, thermodynamic processes have shaped their energy density function to a characteristic distribution known as a thermal, or blackbody, distribution. The emissions that correspond with this distribution are known alternatively as blackbody radiation, thermal, or continuum emission. (The term continuum is used to distinguish the continuous emission over a wide range of wavelengths in opposition to the discrete specific emission of emission line radiation.) The observed both the peak wavelength and luminosity of any object whose emissions display a thermal spectrum is set by the temperature of the location from which the photons originate. In the case of a star this is the photosphere. A number of processes may add additional features to the spectrum of a star, for our purposes the most important is absorption by gases in the star’s atmosphere. The absorption process is the inverse to emission line mechanism described in the previous section; bound electrons in low energy state absorb photons that have the specific discrete energies necessary to excite the electron to a higher energy state in its atom. This process selectively removes specific wavelengths from the stars spectrum, as photons with the corresponding energies are absorbed by the stars atmosphere and hence to do not make it to an outside observer. Since these removed wavelength appear as dark lines when
the stars light is dispersed into a spectrum they are known as absorption lines. To an observer the color of a star is set by the wavelengths of light at which the star is most luminous. These dominant wavelengths are set by the stars thermal continuum emission. The absorption line processes are a secondary effect which reprocess a portion of the stars emissions, but do not directly set peak emission wavelengths.

In the object color and absorption line features item students are tasked with recognizing that thermal continuum emission is the primary driver determining the color of a star. The spectra the students are presenting with in the question are images of dispersal spectra which the students will be familiar with, but which do not contain sufficient information to determine the peak wavelengths of the stars light, and therefore its color. In both the LSCI version and the constructed response adaptation the students attention is focused on “the dark absorption line spectra”, with intention the student recognize the the stars color may not be identified by absorption line features alone.

The 17 bins that make up the rubric for the “object color and absorption features” item based on LSCI item 2, may divided into three categories. The first of these categories is a color comparison of the two stars considered in the item. An example bin from this category is “x is blue, z is red”. Responses in this bin compare the two stars and determine star x is bluer than star z. The second category of rubric bins is reasoning bins. Reasoning bins describe the rationale for identifying the color comparison. The final category of bins is “other information” indicating the response discusses concepts not directly relevant to the central color comparison.

In this exploratory rubric the multiple choice answers of the original LSCI item map directly to rubric bins 1, 2, 4 & 5. As seen in the scoring summary in figure 2, rubric bin 1, “x is blue, z is red”, is by far the most frequently occurring bin being scored in 64% (58% pre, 66% post) of the test response set. The opposite comparison in bin 2, “x is red, z is blue” occurs in 18% (12% pre, 20% post) of responses. Especially noteworthy is that “not enough information” bin appears in less that one percent of all student responses.

Figure 3 displays the frequency of reasoning category rubric bins among responses that were also coded for “x is blue, z is red”. The combined pre/post frequency of each of the three reasoning bins is approximately equal (“Color produce by Doppler shift” 10%, “Light absorbed equal light observed” 11%, “Lines signify emission” 10%). Responses from the post instruction set preferentially describe the stars color as being a product of Doppler shift (4% pre, 12% post), while the pre instruction responses favor an explanation involve the color of light absorbed by an object being equivalent to the color observed (16% pre, 9% post). In neither group, however, do a majority of the students provide a reason for selecting “x is blue, z is red” (28% pre, 32%) post.

### 3.4 Emission Line Wavelength Comparisons

The specific quantized energy levels available to a bound electron in an atom are set by the composition of the atoms nucleus (the number of protons, and to a
lesser extent the number of neutrons) and the number of bound electrons. Since each element has a distinct number of protons, it will also have a unique set of quantized energy levels available to its electrons for any given ionization state. It follows that a unique set of energy levels will produce a unique set of differences between levels, and consequently will produce a unique set of photon wavelengths when electrons transition between levels. To the observer an ionization state of a given element will produce a characteristic emission line spectra which can be used to identify the presence of elements in distant astrophysical sources. It is important to note, though, that the absolute wavelength of an emission line may be modified by a Doppler shift of the emission. When there is relative motion between an emission source and the observer the observed wavelength of a photon will be shifted dependent on the magnitude and direction of the motion. When the motion is decreasing the source-observer distance the observed wavelength is higher than its rest frame wavelength (bluer in color, known as blueshift), when the motion increases the distance the observed wavelength is longer (redder in color, known as redshift). While Doppler shift is able to change absolute wavelength of a photon it does not change the relative wavelengths of photons from the same source. This means that observing a set of emission lines that display wavelength relative to one another it is not only possible to measure

![Figure 4: Frequency of responses for the emission line wavelength comparisons exploratory rubric. Pre-instruction responses are displayed in the upper portion with the post-instruction frequencies in the lower portion. The numeric rubric bin IDs can be found along with descriptions in table 4.](image)
Table 1: Permutations of composition, temperature, and motion rubric bins

<table>
<thead>
<tr>
<th>bin permutation</th>
<th>pre (%)</th>
<th>post (%)</th>
<th>all (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Composition, Temperature, &amp; Motion</td>
<td>0</td>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td>Only Composition &amp; Temperature</td>
<td>7</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>Only Composition &amp; Motion</td>
<td>2</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>Only Temperature &amp; Motion</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Only Composition</td>
<td>9</td>
<td>26</td>
<td>25</td>
</tr>
<tr>
<td>Only Temperature</td>
<td>2</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>Only Motion</td>
<td>0</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Permutation Total</td>
<td>23</td>
<td>87</td>
<td>69</td>
</tr>
</tbody>
</table>

what element produced the collection of lines, but of how the source is moving relative to the observer.

The temperature of a gas comprised of a specific element determines the ionization states of its constituent atoms. Therefore the collection of emission lines and identifiable ions of an element will vary with temperature. At any given temperature, however, multiple ionization states may exist, so a determination of the presence of an ion or even collection of ions, is insufficient to make a measure of temperature. In order to determine the temperature a ratio of ions for a given element must be made. This ion ratio is only measurable through measurement of the luminosity of emission lines, not there wavelength.

Both the LSCI and constructed response versions of the emission line wavelength comparison question ask the student to consider the physical properties of an emission line nebula which are directly measurable from the emission lines wavelengths alone. The LSCI version provides the three parameters considered above (composition, motion, and temperature) while the constructed response version requires students to determine the parameters relevant for consideration themselves.

The exploratory rubric for the "emission line wavelength comparisons" item contains 11 bins which for convenience may be divided into two categories; "LSCI properties" and “Other properties”. The bins in each category describe physical properties of the emission line source. Those assigned to the “LSCI properties” category describe physical properties which were present in the original LSCI multiple choice answers (see appendix A.3). The “other properties” items discuss source object parameters not considered in the LSCI item.

The original LSCI responses were permutations of three properties; composition, temperature, and motion. In the rubric presented here each of these three properties is assigned it’s own bin (bins 1, 2, & 3 respectively). Figure 4 shows that these bins are the most frequently occurring in the post-instruction sample (composition – 69%, temperature – 42%, motion – 39 %) and all display significantly higher representation than in the pre-instruction sample (composition – 19%, temperature – 12%, motion – 5%). The “other property” bin with the highest frequency is that of “color”, occurring in 51% of pre-instruction responses and 22% of the post-instruction set.
Figure 5: Conditional frequency of responses for the emission line wavelength comparisons exploratory rubric. All responses in a given panel are positively scored for the titled rubric bin. Conditional Frequency = \( N(\text{Bin A}|\text{Bin B})/N(\text{Bin B}) \). For which N is the number of responses, “Bin A” is the rubric bin listed on the horizontal axis, and “Bin B” the the panel title rubric bin.

To directly compare the exploratory rubric results with expected results for the original LSCI answers the co-occurrence the composition, temperature, and motion bins must be examined. The results of this exploration are summarized in table 1. Of note is that 87% of the post instructional responses include a permutation of these three rubric bins, while only 23% of pre-instructional response do. It should also be noted that permutations not covered by the the original LSCI answers account for 14% of student responses (9% pre, 16% post).

Another tact of examining the co-occurrence of the composition, temperature, and motion bins is to examine if the mention of one physical property makes it more likely that either of the other two is also mentioned. As such we define the conditional frequency as the number of responses scored for a rubric bin given the scoring of conditional rubric bin normalized to the number of responses scored with the conditional bin (Conditional Frequency = \( N(\text{Bin A}|\text{Bin B})/N(\text{Bin B}) \)). The results of this comparison are given in figure 5. Student responses that are coded for composition are equally likely to mention either temperature or motion, while the majority of students who mention temperature or motion will also mention composition.
4 Discussion and Conclusion

The student responses to the emission line source item present here fail to correspond to the possible answers presented by LSCI multiple choice version. As noted in section 3.2 only 8% of student responses included the two concepts, “temperature” and “density” permutations of which comprised all possible options of the LSCI text. The data considered in the present study is unable to distinguish if this discrepancy may be attributable to the change of language of the constructed response version. Since the constructed response language is more specific, inquiring about “physical properties” rather than “type of object”, it seems difficult to explain how it would prompt a less focused set of student responses.

Independent of the effects of the language change the results of this study do call into question the students ability to interpret the text of the constructed response emission line source item. The diversity of concepts suggest that some portion of the students may have interpreted the item as asking “what can we learn/measure about the physical properties of the object” rather than the intended “what physical conditions are necessary to produce emission lines”. This line of reasoning is further supported by the “composition” rubric bin being the second most frequently occurring bin. While composition is measurable physical property of the emission line source, it does not serve as condition for production of emission lines. (Assuming the composition the students refer to is more specific than normal, non-dark matter.)

The pre-instruction focus on “astrophysical object” rubric bins on the emission line source is likely due to instructional effects. In the courses used in this study students begin the term studying the properties of light and basic light/matter interactions. The most common pre-instructional rubric bins were “light/matter interaction”, “energy content”, and “properties of light”, all three of which are less prevalent in the post-instructional set. As such the over-representation of the “astrophysical object” bins in the pre-instruction set may simple be a case of may students writing what they knew rather than directly responding to question.

In order to test the possibility of students misinterpreting the language of the item to inquire about measurable physical parameters rather than required conditions for emission line production, the question should be modified for future work. A refined question could read: “If the light coming from a distant object produces a bright emission line, what do we know about the physical properties of the distant object which are necessary to produce this emission?” With such a wording it should be possible to investigate whether the question wording or the underlying concepts are the source of the observed student difficulties.

For the object color and absorption features item students do not appear to struggle with the interpretation of the question. In the combined pre/post-instruction sample 84% of students provide a color comparison covered by the option in the original LSCI item. Rather the difficulty lies in providing a correct answer, with less than one percent of the post-instruction students provide a response coded in the “not enough information” bin. It is entirely possible that in
the constructed response format students do not view “not enough information” as an acceptable answer.

If the assumption is made that students view there possible choices in the constructed response as being limited to responses which would be coded as alternatively “star x is blue, star z is red”, “star x is red, star z is blue”, and or “stars same color”, the set of responses is problematic. Given a scenario where the thermal continuum emission of the two stars in negligible for the color comparison, perhaps the stars share identical thermal emission and so the secondary affects of the absorption features are important in color comparison, students still select the incorrect “x is blue, z is red” over “x is red, z is blue” by better than a 3 to 1 margin. The student responses do not provide a clear reason for the overwhelmingly selecting the “x is blue, z is red” comparison. The 32% of post instructional students who describe this comparison in their responses and provide reasoning to support it are split along three possible explanations.

Even if the premise is accepted that the current structure of the constructed response version of the object color and absorption feature question is flawed, the most probable conclusion to draw from the set of test responses is that students have significant difficulty with the mechanisms and effects of line absorption processes.

To examine the issues seen here in the object color and absorption features item two modifications to the question are necessary. First the question should be modified with language explicitly allowing the permissibility of a “not enough information response”. A phrasing such as “Is it possible to determine the color of the stars from these spectra? If so, what can you determine about the colors of the two stars?” As a second modification the question should include a prompt for an explanation of reasoning. If the result of student struggles with absorption features presented in the work is robust, more complete descriptions of student reasoning will be necessary to determine the origin of the difficulties.

For post-instructional students the composition, temperature, and motion rubric bins in the exploratory rubric for the “emission line wavelength comparison” item are sufficient to capture the majority of student thought. The added permutations available in the constructed response format boost the response set coverage of 71% for a LSCI-like set of combinations to the 87% in the response set presented here. From these results it is likely that the LSCI format form of the question is largely representative of student thinking on the topic, with the constructed response version provide more complete coverage without the prompting affect potential to multiple choice items.

If the students' interpretation of the question and ability to discuss the relevant concepts for the emission line wavelength comparison item is promising, the frequency of correct responses is not. Only 14% of the post-instruction response correctly included a discussion of composition and motion, but not temperature. This difficulty appears to stem from the fact that both temperature and motion have similar frequencies in the post-instruction sample (~ 40%). Further students who discuss one of these two concepts are equally likely to discuss the other as is the general response population. If this were a multiple choice item such statistics would be suggestive of guessing (and may still be for
constructed response answers). It seems likely however that the students in the test sample were able to identify the key measurable parameters, but struggled with the specifics as of the emission line wavelength comparison.

The pre-instruction responses for the wavelength comparison item on the other hand do not give strong indications of student thinking. The most popular rubric bin is “color”, which may be a result of question administration order and the implied association of color and spectral line features from the color and absorption line features item. Beyond that possible feature the pre-instruction responses appear to be a grab bag of physical properties possibly indicative that student to not have well formed ideas of what may be learn from the comparison of emission line features to laboratory references.

In the next phase of this program the three items will be considered for alteration, re-administered, and predictive rubrics developed for inclusion in the AACR automated scoring system. The “emission line sources” item will be modified to specify a focus on the physical parameters necessary for emission line production, rather than any measurable (or inferable) property. The “object color and absorption features” will be tested with split versions alternately containing and omitting a focusing statement on the thermal continuum. Both versions of this item will also include a prompt for the reasoning supporting the comparison. The “emission line wavelength comparison” item is ready for additional data collection and predictive rubric development in it’s current form.

The three items adapted from the LSCI presented here all display promise to provide additional insight into student thinking on astronomical EM spectrum when presented in an constructed response format. Two of the items have shown significant problems in post-instruction student thinking, and the third suggests an incomplete picture of measurable properties. On the completion of future work all three items should provide a valuable supplement to the LSCI for instructors to assess there students’ learning.

Author Note

This work is supported by the National Science Foundation (DUE 1323162 and 1347740). 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.

References


A LSCI Items and Constructed Response Forms

A.1 Item: Emission Line Sources

 Adopted from LSCI Item 21.

 **Original:** If the light coming from a distant object produces a bright line emission spectrum, what kind of object is it?

 a. Hot and dense.
 b. Cool and dense.
 c. Hot and diffuse.
 d. Cool and diffuse.

 **Constructed Response:** If the light coming from a distant object produces a bright emission line, what do we know about the physical properties of the distant object?

A.2 Item: Object Color and Absorption Features

 ![Star X](image1.png) ![Star Z](image2.png)

 Figure 6: LSCI Item 2 (Bardar et al., 2007)

 Both the original version of LSCI Item 2 and the constructed response adaptation refer to figure 6.

 **Original:** Consider the dark line absorption spectra shown below for Star X and Star Z. What can you determine about the colors of the two stars? Assume that the left end of each spectrum corresponds to shorter wavelengths (blue light) and that the right end of each spectrum corresponds with longer wavelengths (red light).

 a. Star X would appear blue and Star Z would appear red.
 b. Star X would appear red and Star Z would appear blue.
 c. Both stars would appear the same color.
 d. The colors of the stars cannot be determined from this information.
**Constructed Response:** Consider the dark line absorption spectra shown below for Star X and Star Z. What can you determine about the colors of the two stars? Assume that the left end of each spectrum corresponds to shorter wavelengths (blue light) and that the right end of each spectrum corresponds with longer wavelengths (red light).

### A.3 Item: Emission Line Wavelength Comparisons

![Figure 7: LSCI Item 17 (Bardar et al., 2007)](image)

Both the original version of LSCI item 17 and the constructed response adaptation presented here refer to figure 7

**Original:** The bright line emission spectrum shown above is characteristic of the region of the nebula marked in the drawing. By comparing the *positions* of the lines in the spectrum to a known laboratory spectrum on Earth, which of the following properties of the nebula can be *directly* determined?

a. Motion towards or away from Earth only.
b. Temperature only.
c. Chemical composition (type of atoms) only.
d. Motion and chemical composition.
e. Motion, temperature, and chemical composition.

**Constructed Response:** The bright line emission spectrum shown above is characteristic of the region of the nebula marked in the drawing. By comparing the positions of the lines in the spectrum to a known laboratory spectrum on Earth, what properties of the nebula can be directly determined?
### B Constructed Response Exploratory Rubrics

Table 2: LSCI item 21 constructed response exploratory rubric

<table>
<thead>
<tr>
<th>ID</th>
<th>Rubric Bin</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>light/matter interaction</td>
<td>Response includes a description of a process of light/matter interaction; eg. emission, absorption.</td>
</tr>
<tr>
<td>2</td>
<td>object type</td>
<td>Emission lines identify astrophysical object class; eg. stars, nebulae.</td>
</tr>
<tr>
<td>3</td>
<td>properties of light</td>
<td>Emission lines provide information on properties of light; eg. wavelength, luminosity.</td>
</tr>
<tr>
<td>4</td>
<td>state of matter</td>
<td>Emission lines signify a the sources state of matter; eg. gas, solid.</td>
</tr>
<tr>
<td>5</td>
<td>energy content</td>
<td>The emission lines provide information on the amount and/or source of the objects internal energy.</td>
</tr>
<tr>
<td>6</td>
<td>secondary object</td>
<td>The emission lines provide information on secondary sources in the system; eg. cool gas clouds, planets.</td>
</tr>
<tr>
<td>7</td>
<td>temperature</td>
<td>The emission lines provide a measure of the objects temperature.</td>
</tr>
<tr>
<td>8</td>
<td>chemical composition</td>
<td>The emission line allow for a determination of the objects chemical composition.</td>
</tr>
<tr>
<td>9</td>
<td>density</td>
<td>The emission line provide a measure of the source’s density.</td>
</tr>
<tr>
<td>10</td>
<td>distance</td>
<td>The emission lines allow for a measure of the distance between the source and the observer.</td>
</tr>
<tr>
<td>11</td>
<td>mass</td>
<td>The emission lines provide for a determination of the systems mass.</td>
</tr>
<tr>
<td>12</td>
<td>geometry</td>
<td>The emission lines provide information pertaining the the source’s size, shape, etc.</td>
</tr>
<tr>
<td>13</td>
<td>kinematics</td>
<td>The emission lines provide a measure of the objects kinematics.</td>
</tr>
<tr>
<td>14</td>
<td>age</td>
<td>The emission lines allow for a determination of the source’s age.</td>
</tr>
<tr>
<td>15</td>
<td>color</td>
<td>The emission lines provide information on the object’s color.</td>
</tr>
<tr>
<td>ID</td>
<td>Rubric Bin</td>
<td>Description</td>
</tr>
<tr>
<td>----</td>
<td>------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1</td>
<td>x is blue, z is red</td>
<td>Star x has a blue color, star z has a red color.</td>
</tr>
<tr>
<td>2</td>
<td>X is red, z is blue</td>
<td>Star x has a red color, star z has a blue color.</td>
</tr>
<tr>
<td>3</td>
<td>stars different colors</td>
<td>The two stars have different colors, but how they differ is not specified.</td>
</tr>
<tr>
<td>4</td>
<td>stars same color</td>
<td>That stars are the same color.</td>
</tr>
<tr>
<td>5</td>
<td>not enough information</td>
<td>The two given spectra do not provide sufficient information to determine the stars’ color.</td>
</tr>
<tr>
<td>6</td>
<td>color produced by Doppler shift</td>
<td>The color observed is produced by a Doppler shift of the stars light; redshift, blueshift.</td>
</tr>
<tr>
<td>7</td>
<td>light absorbed is not observed</td>
<td>The star’s light appears the colors of light which are not absorbed by its atmosphere.</td>
</tr>
<tr>
<td>8</td>
<td>light absorbed equals light observed</td>
<td>An observer sees the stars as being the same color as the light it absorbs.</td>
</tr>
<tr>
<td>9</td>
<td>lines signify emission</td>
<td>The dark lines in spectra correspond to the wavelengths of light which the star is emitting.</td>
</tr>
<tr>
<td>10</td>
<td>color indicates temperature</td>
<td>The star’s color provides a measure of its temperature.</td>
</tr>
<tr>
<td>11</td>
<td>color indicates luminosity</td>
<td>The star’s color provides a measure of its luminosity.</td>
</tr>
<tr>
<td>12</td>
<td>color indicates object type</td>
<td>The star’s color is determined by it’s object class; eg. blue giant, red dwarf.</td>
</tr>
<tr>
<td>13</td>
<td>color indicates distance</td>
<td>The star’s color provides a measure of its distance from the observer.</td>
</tr>
<tr>
<td>14</td>
<td>color indicates age</td>
<td>The star’s color corresponds to the stars age; eg. blue stars are young, red stars are old.</td>
</tr>
<tr>
<td>15</td>
<td>color indicates lifespan</td>
<td>The stars color provides and indication of its lifespan.</td>
</tr>
<tr>
<td>16</td>
<td>color indicates mass</td>
<td>The stars color can be used to determine its mass.</td>
</tr>
<tr>
<td>17</td>
<td>color produced by thermal emission</td>
<td>The color of the star is produced by thermal radiation, not absorption line features.</td>
</tr>
<tr>
<td>ID</td>
<td>Rubric Bin</td>
<td>Description</td>
</tr>
<tr>
<td>----</td>
<td>-------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1</td>
<td>composition</td>
<td>A comparison of observed wavelength to a known reference allows for the identification of the source objects chemical composition.</td>
</tr>
<tr>
<td>2</td>
<td>temperature</td>
<td>The comparison provides a measurement of temperature.</td>
</tr>
<tr>
<td>3</td>
<td>motion</td>
<td>The comparison provides a measure of the object kinematics.</td>
</tr>
<tr>
<td>4</td>
<td>color</td>
<td>The comparison allows a measure of the objects color.</td>
</tr>
<tr>
<td>5</td>
<td>luminosity</td>
<td>The comparison provides a measure of the source’s luminosity.</td>
</tr>
<tr>
<td>6</td>
<td>distance</td>
<td>The comparison provides a measure of the source’s distance.</td>
</tr>
<tr>
<td>7</td>
<td>density</td>
<td>The comparison provides a measure of the source’s density.</td>
</tr>
<tr>
<td>8</td>
<td>geometry</td>
<td>The comparison provides information on the objects size, shape, etc.</td>
</tr>
<tr>
<td>9</td>
<td>object type</td>
<td>The comparison allows a determination if the object is a star, planet, nebula, etc.</td>
</tr>
<tr>
<td>10</td>
<td>age</td>
<td>The comparison provides a measure of the source’s age.</td>
</tr>
<tr>
<td>11</td>
<td>mass</td>
<td>The comparison provides a measure of the source’s mass</td>
</tr>
</tbody>
</table>
Applying Automated Analysis to Develop a Cost-Effective Measure of Science Teacher Pedagogical Content Knowledge

Molly Stuhlsatz, Chris Wilson, Zoë Buck Bracey, Mark Urban-Lurain, John Merrill, Kevin Haudek

Teacher pedagogical content knowledge (PCK), or the type of teacher knowledge that bridges content knowledge and how to effectively teach the content in classrooms, has been shown to be a significant predictor of both reform-based classroom practice and student achievement in science (Roth et al., 2011). However, despite the importance of this construct, current measures of PCK have limited use, in part due to them being highly time and resource intensive to score and in part because there are no widely accepted PCK frameworks allowing inferences to be made across measures. Although the field has yet to arrive at a consensus model for PCK, such a model is beginning to emerge from efforts such as the PCK Summit, held at BSCS in October 2012, and from a growing body of literature over the last several decades (Berry, Friedrichsen, & Loughran, 2014). This project builds on the prior work of the field to develop a measurement instrument for assessing teacher PCK in science through a video analysis task and automated computer scoring.

Our primary research question for this project is

- How can lexical analysis and machine learning techniques be applied to developing an efficient, valid, and reliable measure of teacher pedagogical content knowledge?

With the following secondary research questions:
- Can we develop automated computer scoring models of teachers’ written responses that closely correlate with expert human coding?
- What feedback can we provide from the automated computer scoring that will facilitate quantitative research and evaluation, professional development and teacher education, and teacher self-evaluation?

This paper focuses on the development of a framework for the construct that is being used to develop a measure of teacher PCK.

Theoretical Framework

Defining PCK

In October 2012, science education researchers and leaders from around the world came together in Colorado Springs, CO, for a summit on the development and measurement of pedagogical content knowledge. We have taken our broad definition of PCK from the summit as follows: “Personal PCK is the knowledge of, reasoning behind, and planning for teaching in a particular topic in a particular way for a particular purpose to particular students for enhanced student outcomes” (Gess-Newsome, 2015, p. 36). From this work emerged a complex model of teacher professional model and skill in
which PCK is embedded as one of several dynamic elements. An important lesson to take from this model is the distinction between PCK and PCK&S, which stands for pedagogical content knowledge and skill (Gess-Newsome, 2015). While PCK is a knowledge base, PCK&S is a knowledge base in action in the classroom. It is important to note that in the development of the framework described here we intend to measure PCK only, which does not require the teacher to be able to apply her or his knowledge—this is why we do not need any classroom observation. As Gess-Newsome (2015) note in their description of the PCK Summit model, there is a tension between what teachers know and what they can actually do, and this distinction is important. Using this framework we are measuring only what teachers know, not what they can actually do. In addition to constructing a synthesis model of PCK,

1) PCK exists on a continuum from weak to strong;
2) PCK can be strengthened through teaching experience, professional development, or other;
3) teaching experience does not necessarily result in increased PCK;
4) teachers with strong PCK are better able to improve student learning;
5) PCK can be found in two forms: knowledge and enactment;
6) enactment of PCK is more difficult to assess and may not lend itself to normative judgments;
7) the explicit knowledge form of PCK may be easier to assess.

PCK represents a complex construct that is difficult to define and even more difficult to measure. Current measures of PCK are highly time and resource intensive to score. Our goal with this project was to use automated analysis to develop an assessment instrument that is grounded in the emerging consensus model for PCK, developed over several decades, but efficient enough to be implemented on a large scale. Such an instrument would primarily be used for research purposes, but we also recognize the potential for future applications to be used as a formative assessment tool for teachers interested in improving their practice.

The Assessment Triangle

Effective educational assessments, whether for students or teachers, must carefully coordinate and align three key components: cognition, observation, and interpretation. This coordination is illustrated in the assessment triangle (Figure 1) described in the National Academy of Sciences publication *Knowing What Students Know* (National Research Council [NRC], 2001). The cognition portion of the triangle refers to the models of knowledge, skills, and learning within the domain that is being assessed. The observation component describes the tasks learners are asked to perform to demonstrate their knowledge and/or skills, such as answering assessment items or demonstrating proficiency via performance tasks. The interpretation component includes the tools to make sense of the observations, such as scoring guides, rubrics,
and measurement models. These three elements, both alone and combined, are essential in ensuring that the assessment provides meaningful information about student or teacher understanding. That is, the information that an assessment provides should be the primary focus of all instrument development activities.

This focus on information provided by assessments is also prominent in test validity theory. The modern, unified view of validity emphasizes the consequential basis of test validity in addition to the more commonly addressed empirical basis. While these two constructs are not always easily separated, the main argument is that tests are not inherently valid but instead are only valid for particular uses or decisions (Messick, 1995). That is, the empirical basis of test validity is not sufficient: We cannot merely run statistical tests, surpass commonly cited cutoffs on various indices, and declare a test valid. Instead, while statistical properties are still required, the modern view of validity requires test scores to be informative and to allow those administering the test to make the decisions the test was designed to inform (Wilson, Roth, Taylor, Landes, & Stuhlsatz, 2012). In this project we bring together the recommendations illustrated in the assessment triangle with the unified view of test validity to develop an instrument that will provide meaningful information.

Our ultimate purpose is to refine the interpretation corner of the triangle to develop effective research tools for measuring PCK, with the secondary goal of formative assessment. This paper focuses on the cognition portion of the triangle, which represents our conceptual framework for PCK. Classroom video analysis tasks provide
the observation portion, and human-scored rubrics provide the interpretation. Through the question development cycle (QDC, Figure 2), the next phase in this project will be to improve on each aspect of the assessment triangle through iterative cycles, refining our rubric (cognition), collecting new data (observation), and eventually replacing human scoring in the interpretation portion with effective automated scoring. The QDC describes the cycle of development created by Urban-Lurain et al. (2013) at Michigan State University to produce a predictive model for automated scoring.

The first stage of the QDC is typically to design new questions to measure participant thinking. This is followed by data collection using those questions and lexical resource development using lexical analysis software to extract key terms and scientific concepts from the writing. These terms and concepts are used as variables for exploratory analysis which aids in rubric development. We use both analytic and holistic rubrics for human coding of responses. During confirmatory analysis the lexical resources are used as dependent variables in statistical and machine classification techniques to predict expert human coding of responses. However, in this project we have begun the cycle without going through the exploratory analysis and rubric development stages, with the development of a rubric grounded in existing theory. A large body of work defining PCK has already been done, and we draw on this work to define the construct. Though we began the cycle with a rubric based on the emerging consensus model for PCK, our development process is ongoing, and we will continue to refine the rubric through future iterations of the cycle using data from participants. Thus the cognition corner of the triangle, representing the PCK construct, will become more valid over iterations of the QDC.

Figure 2. Question Development Cycle (QDC) (Urban-Lurain et al., 2013).
Methods
Research Design

As described above, the QDC drives the design of our research, which is proceeding in two phases. In Phase 1 of the project we have been working with a large data set of responses to adapt an existing PCK open-response instrument to align it with the literature on PCK, the findings from the PCK Summit (Berry et al., 2014), and the practices of science outlined in the NGSS. That is, we are defining the construct of the assessment and addressing the cognition corner of the assessment triangle as discussed above (rubric development in the QDC).

The existing PCK instrument asks teachers to watch video clips of science classrooms and demonstrate their PCK in written responses where they describe their observations of pedagogical moves and student thinking. The existing data set provides a corpus of responses that has been scored by human coders using a rubric (human scoring). At this point we have not yet developed lexical resources, so we have refined the rubric based on human scoring and identified new video clips based on our refined rubric (question revision). These clips provide a higher level of alignment with the PCK framework we developed based on the emerging consensus model. Next, we will collect a new set of responses from teachers (data collection).

Following new data collection we will move into Phase 2 of the project and continue through the QDC (albeit not always in a linear fashion), refining and iteratively improving all aspects of the assessment, developing lexical resources, and finally moving into predictive scoring models. To do this we will use lexical categories as independent variables in models that predict the human scoring as the dependent variable. Moving through the QDC will ultimately produce an instrument and predictive scoring models that will provide meaningful information to multiple stakeholders that will inform decisions regarding research, evaluation, professional development, and teacher education.

Cognition: Building a Framework for the PCK Construct

We have broken PCK into six dimensions, described in detail in Table 1. Our six dimensions are rooted in Magnusson’s seminal model of the five dimensions of PCK: knowledge about student thinking, knowledge about instructional strategies, knowledge about assessment, knowledge about curriculum, and orientation toward teaching (Magnusson, Krajcik, & Borko, 1999). However, our interpretation of these dimensions has been informed by several more contemporary efforts, including the conclusions of educational leaders from the PCK Summit, the well-established BSCS STeLLA strategies for science teaching (Roth et al., 2011; Taylor, Roth, Wilson, Stuhlsatz, & Tipton, 2016), the Next Generation Science Standards (NGSS; NGSS Lead States, 2013), and several factors that research has shown make an impact on equitable learning outcomes in science education.
Table 1. PCK Framework

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contextualized analysis of student thinking (CAST)</td>
<td>noticing student thinking re: domain content and/or bringing in knowledge of common patterns of student thinking in the domain</td>
</tr>
<tr>
<td>Contextualized analysis of instructional coherence (CAIC)</td>
<td>noticing how the teacher does/does not maintain a coherent content storyline</td>
</tr>
<tr>
<td>Contextualized analysis of constructivist instructional strategies (CACIS)</td>
<td>noticing how the teacher does/does not invoke teaching strategies that allow students to construct their own understanding individually or in groups</td>
</tr>
<tr>
<td>Contextualized analysis of responsive teaching (CART)</td>
<td>noticing how the teacher does/does not respond to student thinking through formative assessment, “teachable moments,” or allowing inquiry to be student led</td>
</tr>
<tr>
<td>Contextualized analysis of NGSS practices (CANP)</td>
<td>noticing how classroom activity does/does not align with any of the eight NGSS practices (does not need to explicitly mention NGSS)</td>
</tr>
<tr>
<td>Contextualized analysis of scientific discourse and language in the classroom (CASD)</td>
<td>noticing how science/everyday language is/is not used/scaffolded in the classroom</td>
</tr>
</tbody>
</table>

We have left the first category of Magnusson’s model (student thinking) relatively untouched, as it is considered fundamental to the construct. This has been demonstrated by its inclusion in almost every recent framework for PCK found in the science education literature (e.g., Lee, Brown, Luft, & Roehrig, 2007; Krauss, Brunner, Kunter, Baumert, Blum, Neubrand, & Jordan, 2008; Padilla, Ponce-de-León, Rembado, & Garritt, 2008; Schneider & Plasman, 2011; Brown, Friedrichsen, & Abell, 2013; Gess-Newsome, 2015). This aspect of PCK is incorporated into the dimension of our framework called Contextualized Analysis of Student Thinking (CAST). The use of the word *contextualized* in each dimension of our framework is a deliberate attempt to preserve the importance of student and domain context within all aspects of PCK and is explained in more detail in the next section.

The category of instructional strategies is also fundamental in the PCK literature, but because of potential boundary issues we have narrowed the focus to be on instructional strategies that the literature on learning outcomes in science education has shown to be effective over the last several decades—constructivist and social constructivist instructional strategies (Bransford, Brown, & Cocking, 1999). We take constructivism to broadly encompass instruction designed to allow students to build their own knowledge through a combination of strategies including well scaffolded inquiry, questioning that pushes student thinking, and the use of participation structures such as small groups to capitalize on social construction of knowledge. This aspect of PCK is incorporated into the dimension of our framework called Contextualized Analysis of Constructivist Instructional Strategies (CACIS).

We have also narrowed the category of assessment to focus specifically on formative assessment, which is emphasized across the NGSS, the PCK Summit, and the STeLLA strategies and included in the literature on equity as a productive and authentic
way of measuring student learning and improving teaching (Darder, 1991; Wilson & Sloane, 2008; Solano-Flores, 2008; Moschkovich, 2007; Gipps, 1999). Because our respondents are reacting to short clips of classroom video and are not privy to how the teacher uses assessment longitudinally nor to the design of paper assessments, this category has been altered to better reflect how a teacher can act formatively on student thinking in the moment. We call this “responsive teaching.” This aspect of PCK is incorporated into the dimension of our framework called Contextualized Analysis of Responsive Teaching (CART).

We have also reframed orientation toward science teaching to focus on those practices of science emphasized in the NGSS, in a dimension we call analysis of NGSS practices. This dimension looks for how respondents are making observations and suggestions related to scientific practices in the classroom such as constructing explanations, communicating ideas, or engaging in arguments. We believe that there is a strong link between a respondent’s ability to recognize scientific practice as an important aspect of classroom instruction and their orientation toward science teaching as more than a didactic exercise. This aspect of PCK is incorporated into the dimension of our framework called Contextualized Analysis of NGSS Practices (CANP).

We have adjusted the curriculum category to align with what STeLLA research has shown to be fundamental: instructional coherence (Roth et al., 2011). Henze, van Driel, & Verloop (2008) describe the knowledge and beliefs about curriculum to be focused on the purposes of the content in curricular materials—which is essentially staying true to a learning goal that is appropriate for both the topic and the students. In her chapter based on the PCK Summit, “Model of Teacher Professional Knowledge,” Gess-Newsome, (2015) defines curricular knowledge as including “the goals of a curriculum ... the role of a scope and sequence, and the ability to assess a curriculum for coherence” (p. 32). Respondents in this case do not have access to the full curriculum, but what they see in the classroom is an enactment of that curriculum, and a coherent curriculum is reflected in instruction that maintains a coherent science content storyline (Roth et al., 2011). Thus we aim to measure the observations and suggestions that respondents make regarding learning goals and the coherence of classroom activity around those goals. This aspect of PCK is incorporated into the dimension of our framework called Contextualized Analysis of Instructional Coherence (CAIC).

In addition to these dimensions we have added a category on scientific discourse and language. The literature on equity, particularly for students from nondominant cultural and linguistic backgrounds, emphasizes the importance of linguistic supports in the science classroom (Lemke, 1990; Lee & Fradd, 1998; Ash, 2003; Mosqueda, 2010; Lemke, 2001; Shaw, Bunch, & Geaney, 2010; Fradd, Lee, Sutman, & Saxton, 2002). Building off of this, the NGSS (NGSS Lead States, 2013) suggest that when “supported appropriately,” providing “rich opportunities and demands for language learning” (p.50) in the science classroom can be very beneficial for English language learners and other students from nondominant linguistic backgrounds. In addition, anecdotal evidence from the STeLLA study indicates the potential of linguistic scaffolds for improving the
effectiveness of science teaching for all students, regardless of linguistic background
(Taylor et al., 2016). Thus we seek to measure how respondents are making
observations and suggestions related to scaffolding students’ development of scientific
discourses and vocabulary. This aspect of PCK is incorporated into the dimension of our
framework called Contextualized Analysis of Scientific Discourse (CASD).

We use the word *contextualized* to distinguish a simple analysis of content
knowledge or teaching strategies from a true analysis of PCK. Shulman’s (1986) original
conception of PCK was as the *intersection* of content and pedagogy, and thus it is vital
that we do not try to measure these elements individually. Schneider and Plasman
(2011) warn that PCK should not be measured on individual dimensions of subject
matter, pedagogical, and context knowledge. The construct of PCK is a transformation
not an integration of these dimensions (Magnusson et al., 1999), and thus they cannot
be teased out and measured on their own scale (Gess-Newsome & Lederman, 1999).

Schneider and Plasman (2011) frame PCK as knowledge that is complex but less
situated than other types of knowledge that a teacher might draw on in specific
classroom situations, allowing it to be measured across domains and across settings—
but this does not mean that PCK can be decontextualized from science content and/or
student experience. Research has shown that there are no universal heuristics for the
classroom that can be applied across contexts to create a “good teacher,” and a rubric
that claims to measure such a universal, decontextualized construct should be suspect.
According to Magnusson, Krajcik, and Borko, who developed one of the seminal models
for PCK still used by researchers, PCK “is a teacher’s understanding of how to help
students understand *specific subject matter*” that includes knowledge about how that
subject matter can be “organized, represented, and adapted to the *diverse interests and
abilities of learners*, and then presented for instruction” (1999, p. 96, emphasis added).
Thus, picking out teacher observations and/or suggestions about instruction is not
measuring PCK unless attention is given to subject matter and learner experience. In
fact, Brown, Friedrichsen, and Abell (2013) found that even the Magnusson model was
not coherent enough to adequately capture the fluid, integrated nature of PCK and
recommend a more integrated model. The definition arrived upon at the PCK Summit
emphasizes that PCK is “context specific” knowledge—emphasizing that such knowledge
is attached to “the teaching of particular topic in a particular way for a particular
purpose to particular students” (Gess-Newsome, 2015, p. 36).

The word *contextualized* in our rubric is meant to indicate an analysis of teacher
responses that takes into account the importance of how each dimension of PCK is
inextricable from the other dimensions and in particular from understanding content
and understanding students. We are seeking teacher observations and/or suggestions
that are rooted in an understanding of the importance of both domain and of a situated
understanding of student thinking and experience. Thus, a teacher who simply lists
teaching strategies that could be used in a classroom will not score as highly as a
teacher who describes one strategy but considers it deeply and in the appropriate
context.
The inclusion of subject matter and student context in each and every dimension of our framework means that we do not have discrete categories for identifying content knowledge or context knowledge. While we recognize that those dimensions are sometimes included in research (e.g., Rowan, Schilling, Ball, Miller, Atkins-Burnett, & Camburn, 2001), they are not included in the Magnusson model nor in most subsequent models (e.g., Lee et al., 2007; Krauss et al., 2008; Padilla et al., 2008; Schneider & Plasman, 2011; Brown et al., 2013; Gess-Newsome, 2015). Thus, we do not consider them to be adequate measures of PCK without being incorporated into more practical pedagogical dimensions.

**Observation: Teacher Responses to Classroom Video**

The assessment task that comprises the observation corner of the assessment triangle is an innovative video-based lesson analysis task (Figure 3). Teachers watch video clips of science lessons across several content areas. These video clips are carefully selected from authentic classroom video to include a range of student activity and teacher pedagogical moves. Teachers respond to a prompt to make analytical comments about the science content, the teaching, and/or the students. In response to the prompt, teachers are given 20 minutes to watch the clip and then provide a written analysis, with their responses varying between a few sentences and three paragraphs.

Figure 3. Video-based lesson analysis task.

At this point we have not yet developed lexical resources, so we have refined the rubric based on human scoring and identified new, targeted video clips based on our refined rubric. These clips provide a higher level of alignment with the new PCK framework. We hope that by reducing the length of the video clips, focusing on just one or two categories of the rubric, overall teacher response lengths will be reduced. Shorter teacher responses, which we expect to include only these categories, should allow us to identify lexical resources more easily, complete human coding more quickly, and ultimately refine our computer models efficiently.
Interpretation: Human and Automated Analysis of Teacher Responses

We chose to apply a binary scoring scheme to each of the dimensions of the PCK construct in our rubric as a way of increasing inter-rater reliability between human coders and ultimately increasing the likelihood of success with the computer models. We have already completed one round of human coding based on these six dimensions, using teacher responses to the original lesson analysis task. It was clear to us after this first round of scoring that the length of the responses was difficult for both human scoring and machine scoring. Between our two trained human scorers we found that some categories were quite easy to reach agreement, while others were more difficult for the coders to score. Table 2 shows the Cohen’s kappas between the two new coders and the expert coder. Both Coder 1 and Coder 2 were more likely to agree with the expert than with each other, and the most difficult coding category was Contextualized Analysis of NGSS Practices (CANP).

Table 2. Cohen’s kappa for inter-rater reliability.

<table>
<thead>
<tr>
<th></th>
<th>Coder 1 and Expert</th>
<th>Coder 2 and Expert</th>
<th>Coder 1 and Coder 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contextualized analysis of student thinking (CAST)</td>
<td>.78</td>
<td>.67</td>
<td>.52</td>
</tr>
<tr>
<td>Contextualized analysis of instructional coherence (CAIC)</td>
<td>.68</td>
<td>.77</td>
<td>.55</td>
</tr>
<tr>
<td>Contextualized analysis of constructivist instructional strategies (CACIS)</td>
<td>.68</td>
<td>.68</td>
<td>.56</td>
</tr>
<tr>
<td>Contextualized analysis of responsive teaching (CART)</td>
<td>.55</td>
<td>.76</td>
<td>.60</td>
</tr>
<tr>
<td>Contextualized analysis of NGSS practices (CANP)</td>
<td>.31</td>
<td>.37</td>
<td>.38</td>
</tr>
<tr>
<td>Contextualized analysis of scientific discourse and language in the classroom (CASD)</td>
<td>.93</td>
<td>.89</td>
<td>.85</td>
</tr>
</tbody>
</table>

The next step in this project will be to apply human and automated analyses to the new responses gathered through observation—this represents the interpretation corner of the assessment triangle. After we collect a corpus of responses to the new video tasks, we will again go through steps in the QDC to identify lexical resources, do new human scoring, and then use machine learning to see if we can train the computer to score the responses with the same or better reliability as we are able to achieve with human raters. The final product of the QDC is a predictive model that can be used to completely automate the scoring of a new set of responses, predicting how experts would score the responses.

Conclusion

The ultimate goal of this project is to develop effective and efficient tools for measuring PCK that are grounded in theory and informed by data. We recognize the challenges of measuring PCK yet see potential in the development of computer-generated scoring strategies using lexical analysis and machine learning as a way to
economically include the construct in large-scale research in the future. We are developing these tools by cycling through the QDC, thereby iteratively improving every corner of the assessment triangle as it applies to this measure.

It is widely acknowledged that new standards like the NGSS will only have an impact on teaching and learning if there are high quality assessments that are closely aligned with the standards (NRC, 2012). While much attention is currently being placed on challenges associated with student assessment, the measurement of teacher-level variables is of equal importance. In order to help teachers develop understandings of these complex standards frameworks and the abilities to integrate science ideas, practices, and crosscutting concepts, we need to be able to measure these understandings and abilities. We expect that findings from this study will provide ample evidence for the efficacy of scoring complex teacher understandings without the need for highly time intensive and expensive human scorers.

This material is based upon work supported by the National Science Foundation under Grant 1437173. 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


Automated Analysis Provides Insights into Students’ Challenges Understanding the Processes Underlying the Flow of Genetic Information

Rosa A. Moscarella¹, Kevin C. Haudek², Jennifer K. Knight³, Alexandria Mazur¹, Karen N. Pelletreau⁴, Luanna B. Prevost⁵, Michelle K. Smith⁴, Matthew Steele¹, Mark Urban-Lurain¹, John E. Merrill⁶

¹Collaborative Research in Education, Assessment, and Teaching Environments for the fields of Science, Technology, Engineering and Mathematics (CREATE4STEM), Michigan State University. ²Department of Biochemistry and Molecular Biology, Michigan State University. ³Department of Molecular, Cellular and Developmental Biology, University of Colorado Boulder. ⁴School of Biology and Ecology, University of Maine. ⁵Department of Integrative Biology, University of South Florida. ⁶Department of Microbiology and Molecular Genetics, Michigan State University.

Paper presented at the 2016 Annual International Conference of the National Association for Research in Science Teaching (NARST), Baltimore, MD, April 14-17, 2016.

Author Note
This work is supported by the National Science Foundation (DUE grants: 1323162 and 1347740). 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 Rosa A. Moscarella (moscarel@msu.edu) at the CREATE for STEM Institute, Michigan State University, 620 Farm Lane, Room 115, East Lansing, MI, 48824.
Abstract

Understanding genetics is fundamental for biological literacy and is broadly recognized as one of the most difficult disciplines in biology. Although a number of published assessment instruments are available to assess students’ genetics knowledge, it is unclear to what extent these instruments can measure students’ understanding or reasoning. Constructed response (CR) questions, in which students craft responses to questions using their own words, may provide a more authentic assessment of students’ understanding and reasoning. Automated scoring has made it possible to use CR assessments in large enrollment classes. We have created CR assessments in which students are asked to predict the effect of a DNA mutation on the processes of replication, transcription, and translation. We analyzed over 4,000 student responses, collected from introductory and upper-level biology classes across five universities in the U.S., using a computer scoring model, trained with 1,031 human scored responses. Analyses of path maps of students’ responses and written answers revealed students’ alternative conceptions of each process. The results show that students’ understanding of transcription is critical for the overall comprehension of the processes involved in the genetic information flow. Researchers may find our approach useful for evaluating overall students’ performance in multiple-part CR questions. We suggest instructors focus on transcription as a keystone concept to understanding the processes underlying the flow of genetic information.

Keywords: lexical analysis, text analysis, automated scoring, genetics education, central dogma
Introduction

Genetics is a fundamental subject for biology literacy (AAAS, 2011), yet it is one of the most challenging areas for students to learn (Bahar, Johnstone, & Hansell, 1999; Marbach-Ad, 2001; M. K. Smith & Knight, 2012; M. K. Smith, Wood, & Knight, 2008; Wright, Fisk, & Newman, 2014). Students at all levels struggle to understand the nature of genetic information in the cells of an organism and how this information flows, is exchanged and stored (Lewis & Wood-Robinson, 2000; Newman, Catavero, & KateWright, 2012; Wood-Robinson, Lewis, & Leach, 2000; Wright et al., 2014). In particular, the central dogma of molecular biology, stated as “information in nucleic acid can be perpetuated or transferred but the transfer of information into protein is irreversible” (Crick, 1958), is one topic where students have persistent conceptual difficulties (Jensen, Kummer, & Banjoko, 2013; Newman et al., 2012; M. K. Smith & Knight, 2012; Wright et al., 2014).

Genetic information flow is considered a core concept in Vision and Change (AAAS, 2011), and the lack of understanding of genetics can impede the understanding of other core concepts, for example, the mechanisms of evolution (Kalinowski, Leonard, & Andrews, 2010; Klymkowsky, 2010; White, Heidemann, & Smith, 2013). Our efforts to improve students’ understanding of genetics first requires assessing student knowledge.

A growing trend is to assess discipline literacy via multiple-choice concept inventories. These are valuable instruments that can be used to measure student initial knowledge and learning gains. Instruments to assess genetics knowledge include the Genetic Concept Assessment (GCA; M. K. Smith et al., 2008), the Genetic Literacy Assessment Instrument (Bowling et al., 2008), the Biology Concept Inventory (Klymkowsky & Garvin-Doxas, 2008), and the Molecular Biology Capstone Assessment (Couch, Wood, & Knight, 2015). However,
results from these multiple choice tests may not reflect students’ actual knowledge because students have learned test taking strategies and prepare for multiple choice tests differently than they prepare for tests using constructed response questions (Stanger-Hall, 2012). Although multiple true/false questions can reveal heterogeneity of student thinking (Couch et al., 2015), it is unclear to what extent multiple choice tests in general can measure students’ understanding or reasoning (J. I. Smith & Tanner, 2010). In contrast, constructed response (CR) questions may reflect student understanding more accurately than multiple choice questions because they require students to craft explanations using their own words (Bennett & Ward, 1993; Kuechler & Simkin, 2010). CR items are also a more authentic assessment tool that can potentially reveal novel student conceptions not previously considered and therefore unavailable among the options of multiple choice tests (Birenbaum & Tatsouka, 1987). However, an important limitation of CR assessments is the time and resources needed to evaluate student responses, especially in large enrollment classes. This challenge can be mitigated with the use of computerized lexical analysis and scoring algorithms (Ha, Nehm, Urban-Lurain, & Merrill, 2011; Haudek, Prevost, Moscarella, Merrill, & Urban-Lurain, 2012; J. J. Kaplan, K. C. Haudek, M. Ha, N. Rogness, & D. Fisher, 2014; Magliano & Graesser, 2012; Nehm, Ha, & Mayfield, 2012; Park, Haudek, & Urban-Lurain, 2015; L. Prevost, Haudek, Urban-Lurain, Merrill, & Lee, 2012; Urban-Lurain, Prevost, Haudek, Henry, Berry, Merrill, et al., 2013).

In this study we investigated college students’ ability to correctly differentiate the central dogma processes of replication, transcription, and translation when asked to describe the effect of a mutation generating a stop codon. Previous studies have shown that college students believe that stop codons, which signal the end of polypeptide synthesis during translation, stop transcription, the process of producing RNA from information in DNA (M. K. Smith & Knight,
2012; Wright et al., 2014). We used computerized lexical analysis of CR questions to investigate the challenges undergraduate students have in understanding the impact of this mutation on replication, transcription and translation. We analyzed over 4000 written responses from a heterogeneous sample of undergraduate students. Our analytic approach can be adopted by other education researchers to investigate diverse topics related to teaching and learning science through the implementation of open-ended assessments. Instructors may find these results helpful to inform their teaching and help their students succeed mastering genetic concepts in their classes.

**Methods**

We created a three-part CR question based on two items from the GCA (M. K. Smith et al., 2008) about the effect of a mutation, which produces an early stop codon, on the processes of replication, transcription and translation (Fig. 1). Correct answers to the questions in Figure 1 should state that the mutation will not affect either the process of replication or transcription, although the product of both processes will carry the mutation. Translation, though, will be affected because the early stop codon would cause the process to terminate sooner than intended, producing a shorter polypeptide.
Figure 1. CR questions based on two items from the GCA. The original multiple choice questions can be found in Smith and Knight (2012)

The following DNA sequence occurs near the middle of the coding region of a gene.

DNA 5’ A A T G A A T G G* G A G C C T G A A G G A 3’

There is a G to A base change at the position marked with an asterisk. Consequently, a codon normally encoding an amino acid becomes a stop codon.

1. How will this alteration influence DNA replication?
2. How will this alteration influence transcription?
3. How will this alteration influence translation?

Responses were collected as online homework assignments from university students enrolled in introductory biology courses for science majors (BIO, 7 courses, five institutions, N=3230) and upper level genetics courses (GEN, 2 courses, two institutions, N=818) across five universities in the U.S., during the fall 2014 and spring 2015 semesters.

An instructional intervention about this topic/question was developed by some of the instructors responsible for the courses (Pelletreau et al., in preparation). The same intervention was presented to all participating courses. The questions were administered twice during each semester: post instruction about the central dogma but PRE intervention (BIO n=1610; GEN n=430, and POST intervention (BIO= 1620; GEN=388; hereafter refer to as PRE and POST).

Students electronically submitted their responses to each part of the question in separate text boxes in an online Learning Management System. They were explicitly requested not to use any external resources but to answer the questions to the best of their knowledge. Students were awarded participation points for the completion of the assignment, regardless of the correctness of their answers. This study was designated exempt by the University Institutional Review Board (IRB# x10-577).
Students’ written responses to each CR part were analyzed using computer models that predict human expert scoring by combining lexical and statistical analyses (Ha et al., 2011; Haudek et al., 2012; J. J. Kaplan, K. C. Haudek, M. Ha, N. Rogness, & D. G. Fisher, 2014; Nehm, Ha, et al., 2012). For this study, the lexical analysis was performed using IBM SPSS Text Analytics for Surveys v. 4.0.1 (IBM Corp). During lexical analysis, key words and phrases are extracted from the text of each response. The extraction procedure can be refined through the creation of custom libraries to include technical terminology that would not otherwise be recognized by the software’s linguistic algorithms. Then, words/phrases with similar content and that are relevant to the question are grouped into conceptual categories that are defined by researchers with disciplinary expertise. The software uses the definitions of the categories to automatically categorize each student’s response into one or more of the categories (Fig. 2). In this way, each response is now coded into a suite of categories. The set of categories created for the lexical analysis of each question part was used to analyze the data, which provides consistent results across all three questions. After each student’s response is categorized into the various categories, those categories are used as independent variables in statistical analyses. In other words, the analysis of each student’s response is based on the categories associated with their written answers.
Figure 2. Screenshot of IBM SPSS Text Analytics for Surveys v. 4.0.1 (IBM Corp) to illustrate how words are extracted from student’s answers (words in color in right panel) and grouped in categories (upper left panel) that are the variables analyzed in the ensemble scheme. Words extracted, or terms are shown in the lower left panel.

We created a three bin holistic scoring rubric for each question that was used to score a set of responses (N = 1031) as completely correct, irrelevant/partially correct, or incorrect (L. B. Prevost, Smith, & Knight, in revision).
Table 1. Examples of students’ responses for each rubric definition (explanation in italics)

<table>
<thead>
<tr>
<th>Rubric Level</th>
<th>Replication</th>
<th>Transcription</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Correct</strong></td>
<td>It will not influence DNA replication. <em>Response suggests that replication is unaffected by the point mutation.</em></td>
<td>There will be no change in transcription but the newly transcribed mRNA sequence will include the mutation. <em>Response states the point mutation will not cause transcription to end early.</em></td>
<td>The amino acid sequence will be shorter than expected. <em>Response suggests that translation will end early at the new stop codon.</em></td>
</tr>
<tr>
<td><strong>Incomplete/ Irrelevant</strong></td>
<td>This will cause a mutation. <em>Incomplete: The response only states that a mutation will occur.</em></td>
<td>It could cause the DNA to be fixed or could cause mutations in the cell that could be bad. <em>Irrelevant: This response does not refer to transcription.</em></td>
<td>You would end up getting a different kind of protein. <em>Incomplete: It is unclear how the protein will be different.</em></td>
</tr>
<tr>
<td><strong>Incorrect</strong></td>
<td>DNA replication would stop. <em>Response suggests replication will stop once it reaches the point mutation.</em></td>
<td>The mRNA will stop at this point and read no further, so the mRNA will be significantly shorter. <em>Response suggests that transcription will end early.</em></td>
<td>Translation will proceed as normal. <em>Response suggests that the stop codon will not affect the process of translation.</em></td>
</tr>
</tbody>
</table>

The statistical analysis consists of an ensemble of machine learning algorithms developed in R that is used to predict the scoring of student responses (dependent variable) based upon the categories assigned during lexical analysis (independent variables) (Jurka, Collingwood, Boydstun, Grossman, & van Atteveldt, 2012). The ensemble consists of nine classification algorithms, each of which is trained on the set of human scored student responses. When a new student response is fed to the system, each individual classification algorithm returns a set of membership probabilities for each rubric level. The rubric level that has the highest aggregate classification probability across the ensemble is assigned as the score for that student response. The interrater reliability between human scores and computer predictions was within the range
usually seen between human raters using well calibrated rubrics (Landis & Koch, 1977; Nehm, Ha, et al., 2012).

We used two approaches to visualize students’ performance on each question part. First, we created stacked bar plots to show the proportion of answers in each rubric level for each question part PRE and POST. Second, we created path maps that trace all student responses through each question part. To do so, the responses of each student were paired by question part and rubric level, and the frequency of each question/rubric level pair was determined. These maps allow us to visualize the general pattern of students’ performance by looking at the frequencies of responses in all combinations of correct, irrelevant/incomplete, and incorrect for the three parts of the questions.

Additionally, we performed a qualitative analysis of student’s written responses, following an inductive approach (Bryman, 2012), to characterize alternative conceptions students have for each of the process encompassed in the central dogma.

**Findings**

Students in both BIO and GEN showed similar trends in the proportion of answers scored correct, incomplete/irrelevant or incorrect, although GEN had larger percentages of correct answers, for both PRE and POST (Fig. 3). In both courses, we found that the question that had the largest percentage of correct answers for the PRE data collection was translation, while transcription had the lowest. POST data showed improvement in all three question parts for both courses. It is noteworthy that for the transcription data, though the proportion of correct answers increased, it still had the largest proportion of incorrect answers in both courses.
Moreover, for the BIO data the percentage of incorrect answers remained almost the same as PRE, showing little gain from instruction.

![Stacked bar plots showing the percentages of answers in each rubric level, PRE and POST, for both BIO and GEN](image)

**Figure 3.** Stacked bar plots showing the percentages of answers in each rubric level, PRE and POST, for both BIO and GEN

We use “path maps” to represent the relationships among students’ responses to each question part (i.e. correct, incomplete/irrelevant, incorrect; Figs. 4 and 5). The size and color of the circles represents the percentage of the students’ responses from the total responses that falls in that category. The color and thickness of the arrows represent the percentage of responses from the total that were classified in the other categories.
We can focus on any one of the question parts (Fig. 4 focuses on transcription) and explore the patterns for the other question parts. Alternatively, we can focus on a rubric level (Fig. 5 focuses on correct responses) and investigate the patterns for the different question parts. For instance, the top panel of Figure 4 is focused on student’s correct responses for transcription, i.e., 100% of those responses. From those students in BIO, 66% responded correctly to replication and 65% to translation. The percentage of students that responded either incomplete/irrelevant or incorrect to those question parts was <20%. In GEN the pattern is more pronounced: over 90% of the students that responded correctly to transcription also responded correctly to replication and translation. In BIO, from all the students whose transcription response was scored as incomplete/irrelevant, similar percentages responded to replication and translation either correct, incomplete/irrelevant or incorrect (left middle panel Fig. 4). Interestingly, a large percentage of students whose transcription answer was scored incorrect were scored correct for translation (BIO= 56%, GEN= 76%, bottom panel Fig. 4)
Figure 4. Path map showing patterns of students answers for transcription. Here, we show how students answered to both replication and translation when they answered to transcription either correct, incomplete/irrelevant, or incorrect. For instance, the bottom of the figure shows the path students followed when they answered to transcription incorrectly. For BIO, similar percentages of students answered to replication either correctly or incorrectly (~40%), while almost 60% responded correctly to translation.
Figure 5. Path map showing the patterns of students correct answers for each process. The top of the figure shows the path students followed when they answered to replication correctly. For GEN, 84% of students who answered to replication correctly also answered correctly to transcription, and 74% responded correctly to translation. Less than 20% had responses either incomplete/irrelevant or incorrect in the other two processes.

Representing the classifications of student responses using a path map reveals that students’ understanding of transcription seems to be key for the full comprehension of the flow of genetic information (Fig. 4). As it can be seen in the path maps, a larger percentage of students with a correct answer for transcription also responded correctly to replication and translation than those who have transcription either incomplete/irrelevant or incorrect responses, in both courses (Fig. 4). For the BIO data set, >65% of students who responded correctly to transcription also responded correctly to replication and translation; while of students who responded correctly to replication, 64% responded correctly to transcription and 46% to translation; from those who responded correctly to translation, only 40% responded correctly to
replication and 55% to transcription. For the GEN data set the results are more homogeneous, but still a larger percentage of correct responses is observed in the group of students who responded correctly to transcription.

Qualitative analysis of written answers from the POST data focused on the trends observed in the path maps with >20% of students answers following a particular path. In each example, the responses for replication, transcription and translation were given by the same student. This analysis shows that students have several common alternative conceptions about each process. Some of these alternative conceptions are shown below (Rep=replication, Trc=transcription, Tsl=translation; co=correct, ii=incomplete/irrelevant, in=incorrect):

1. Some students think that a shorter protein at the end of translation is due to the fact that the mRNA was shortened during transcription because of the stop codon: Rep-co: “The alteration will not affect DNA replication. DNA replication does not involve the use of STOP codons” Trc-in: “The RNA polymerase will stop coding prematurely. The mRNA strand will be short, and it will code for an incomplete protein.” Tsl-co: “The STOP codon will stop translation of the protein. Since this error already occurred previously in transcription, an incomplete protein will result from the already short mRNA strand.” [Note: The rubric focuses on the responses to the particular process. In this case, the answer is still correct despite the wrong premise for the answer]

2. Students still can reach the correct answer for translation (the process will be halt and therefore a shorter protein will be made, likely dysfunctional) by having wrong ideas about the effect of such mutation during both replication and transcription: Rep-in: “When the DNA polymerase gets to the A that was a G it will stop because it is now a stop codon.” Trc-in: “When the DNA polymerase gets to the newly formed stop codon
transcription will stop leaving the rest of the DNA not coded for.” Tsl-co: “After the DNA has been transcribed (or the small part that has) translation will then translate the mRNA into a protein. The protein will not be the correct protein because it was stopped prematurely.”

3. Other students thought a stop codon will stop transcription but did not clearly explain what the effect on translation would be: Rep-co: “It will not affect it.” Trc-in: “Transcription will be stopped prematurely.” Tsl-ii: “It will be translated incorrectly.”

4. There were students that thought that translation will not be affected by this mutation, although the other processes will: Rep-in: “This will tell the DNA to stop before it really is supposed to. This will make it shorter than it is supposed to be.” Trc-in: “The mRNA will be shorter than it is supposed to be and will not contain all of the correct amino acids.” Tsl-in: “Translation will not be affected.”

5. In other instances, students did not make a clear statement that a stop codon would stop translation prematurely, but instead referred to a change in the amino acid sequence: Rep-co: “This alteration will not influence DNA replication.” Trc-co: “This alteration will not influence transcription.” Tsl-ii: “This alteration will influence the amino acid sequence made during translation.”

6. Some other responses suggest that the effect of this mutation on replication is independent of the other two processes: Rep-in: “it will stop the DNA from replicating any further” Trc-co: “it will change the G to an A so when it is being copied to RNA” Tsl-co: “the stop codon stops the translation into amino acids therefore creating a whole different amino acid than what it was supposed to be”
7. Some responses suggested that none of the three processes could take place: Rep-in: “Replication can’t occur without transcription” Trc-in: “It will stop transcription prematurely, making the cell nonfunctional.” Tsl-in: “Translation can’t occur”

8. Some students believe that the short protein is caused by a replication that stop prematurely, producing a shorter DNA that was transcribed into a shorter mRNA and therefore translated into a shorter protein: Rep-in: “The strand will be completed short” Trc-in: “It will only transcribe the first part of the DNA.” Tsl-co: “The RNA strand will be translated off a short DNA strand and will result in a shortened protein.” [See note above]

9. Often, students answered to the effect of the mutation in translation without addressing how replication and transcription would be affected by this mutation: Rep-ii: “If the DNA strand now reads 5’ AATGAATGAGAGCCTGAAGGA3’ Then the complementary mRNA strand will read 3’ UUACUUACUCUCGGACUUCCU 5 The nonsense mutation will result in a nonfunctional protein” Trc-ii: “If this is a nonsense mutation, it will result in a shortened protein that is probably nonfunctional or if it is functional then it is highly impaired. (…)” Tsl-co: If this is a stop codon, then the translated protein will probably not be functional, as it will be much shorter and with less amino acids than intended”

10. Some students’ answers did not address the question; answers were irrelevant or incomplete: Rep-ii: “It can be described as a nonsense mutation which doesn allow the correct amino acids to be made, creating an incomplete protein.” Trc-ii: “It creates an error in the mRNA replication, the mRNA will be make incorrect amino acids.” Tsl-ii:
When the mRNA goes through translation to code for proteins needed it won't have all the correct amino acids to create the necessary proteins.”

11. Some students responded correctly to replication and translation but their answer to transcription was unclear: **Rep-co:** “No effect on the replication process, except that the strand formed from this template strand will have a T where it otherwise would have had a C.” **Trc-ii:** “Could affect boundary between intron/exon, resulting in more/less of the gene being expressed as mRNA than should be.” **Tsl-co:** “mRNA with the new Stop codon will cease translation earlier than otherwise would have. Protein will be shorter than it should be (truncated protein).”

**Discussion**

In this study we used automated scoring to score a three-part constructed response question in several large enrollment biology courses, both introductory and upper division, across five universities in the U.S. We have demonstrated that this methodology makes it feasible to integrate writing in the classroom for formative assessment (Ha et al., 2011; Haudek et al., 2012; Jennifer J. Kaplan et al., 2014; Moscarella et al., 2008; Nehm, Beggrow, Opfer, & Ha, 2012; Nehm, Ha, et al., 2012; L. Prevost et al., 2012; Urban-Lurain et al., 2009; Urban-Lurain, Prevost, Haudek, Henry, Berry, & Merrill, 2013; Weston, Haudek, Prevost, Merrill, & Urban-Lurain, 2014), which accomplishes two objectives: 1) helping students organize their knowledge and 2) providing instructors with richer insight about student thinking in the domain. In this article we elaborate on the second objective, namely using this type of assessment to obtain greater insight into students understanding of a particular subject- in this case, genetic information flow.
Genetics is a complex subject that students struggle to master (e.g., Bahar et al., 1999; Marbach-Ad, 2001). Our results not only corroborate that students have difficulties differentiating among the key molecular processes that underlie the central dogma, both in introductory biology and upper level genetics courses, but also provide greater insight into student reasoning and misconceptions. Through this analysis we were able to identify the process of transcription as a critical concept for the full understanding of the flow of genetic information.

We came to this conclusion about transcription after distinguishing two patterns in the data: First, the fact that transcription held the largest proportion of incorrect responses even after intervention (Fig. 3). Many students believe that a stop codon stops transcription, a persistent misconception previously noticed by Smith and Knight (2012). Second, the pattern that made clear that transcription may be a keystone concept emerged from the path maps. More than 65% of the students who responded correctly to the transcription part of the question in BIO also responded correctly to replication and translation. For GEN, this percentage is above 88% (Fig. 4). This pattern is not seen in students who answered correctly either replication or translation. In those cases, the percentage of students who have also a correct response for the other two processes is much lower (Fig. 5).

An interesting conceptual issue found in our results, not previously reported in the literature, was that some students thought that a stop codon could also stop replication. This was a more prevailing problem in the PRE data. When students were first presented with a problem in which they had to predict the effect of a mutation in replication, two potential problems were revealed: 1) students do not have a clear conception of what codons are and 2) failed to correctly predict the impact of mutations on the processes of replication, transcription and translation.
Students’ struggle with understanding what codons are could help explain why students harbor the idea that stop codons terminate the synthesis of RNA during transcription. While other assessments target this common misconception (e.g., GCA; M. K. Smith et al., 2008), the CR assessments allow us to explore what students think about stop codons and their effect on the different processes involved in the central dogma, offering a broader picture of diverse student conceptions.

The design of appropriate instructional interventions requires knowing the nature of students’ learning difficulties. The path maps make it possible to visualize students’ alternative conceptions. This approach, along with the qualitative analysis of written responses, gave us a unique insight into the overall students’ performance across all the parts of the question. The path maps allow instructors to analyze students’ responses to multiple-part questions and rapidly identify specific concepts that may be affecting students’ global comprehension of a science phenomenon. This information can be used to plan instructional interventions and to inform curriculum development.

Our results suggest that transcription may require special instructional attention. We are currently investigating the details of students’ struggles when learning transcription. We are analyzing interviews of students with different levels of genetics understanding to characterize students’ conceptions, misconceptions and thinking about the processes of replication, transcription and translation. Meanwhile, we recommend that instructors use the results of this study and explicitly address in their classes the distinction between DNA replication and protein synthesis and have students predict and evaluate the effects of mutations at different levels. Students should be able to describe the processes and products of transcription and translation,
know how these processes are initiated and terminated, and explain how the products of transcription are used in translation.

Comparing the results obtained for BIO and GEN exposes the need of finding ways to make genetics more accessible for introductory biology students. McDonnell, Barker & Wieman (2016) compared two groups of students, both taught by the same instructor using the same active learning practices and only differing in the sequence in which concepts and definitions were presented. They found that that the group in which the presentation of concepts preceded the definitions showed a better understanding of the new concepts. An interesting hypothesis for future inquiry is to investigate whether reducing emphasis on definitions in lower level classes will help students to focus more on the deep understanding of the content. This needs to be aligned with teaching practices and assessments that promote higher level thinking (Momsen, Long, Wyse, & Ebert-May, 2010). Considering that introductory biology courses might be the only courses in which some students are exposed to genetic concepts, along with the fact that genetic literacy is gaining increased importance, instructors should consider what students need to learn to be literate in genetics. We join the call of Duncan, Rogat, & Yarden (2009), Elmesky (2013), Shea & Duncan (2013), and Wright et al. (2014) for the development of learning progression schemes to help instructors improve their students’ learning of genetics.

References


Smith, J. I., & Tanner, K. (2010). The Problem of Revealing How Students Think: Concept Inventories and Beyond. *CBE-Life Sciences Education, 9*(1), 1-5. doi:10.1187/cbe.09-12-0094


Predicting the Accuracy of Computer Scoring of Text: Probabilistic, Multi-Model, and Semantic Similarity Approaches

Minsu Ha (Kangwon National University, Korea)
Ross Nehm (Stony Brook University--SUNY)

Multiple studies have shown that automated computer scoring systems (ACSS) are able to grade essays and short answers as well as trained human raters (Magliano & Graesser, 2012). However, computer-scored results from current ACSSs continue to display discrepancies with those of human raters (about 5-10%; see Moharreri et al., 2014). Such scoring discrepancies appear to be caused by the lexical ambiguity of words, the use of very uncommon words by respondents, and the discordance of semantic information between the training corpus (the data used to develop the scoring algorithm) and the testing corpus (the data to be scored; see Ha et al., 2011, Nehm et al., 2012).

Although low-confidence predictions (LCPs) in ACSS scores are known to be due to many factors, current machine-learning systems that are used to score open-ended text, such as EvoGrader (www.evograder.org; Authors, 2014) are not capable of informing users of LCPs. In addition, no study to our knowledge has been conducted that explores different methods (as opposed to different models) for identifying LCPs in ACSS. Being able to know more about the confidence of computer scoring predictions would be helpful to users of these systems and allow them to use these new assessment tools more effectively.

Study Aim

Our study explored three methods for identifying LCPs in the scoring of open-ended text: (1) the probability of machine-learning predictions; (2) the discrepancies among scores from multiple different computer scoring models; and (3) the semantic similarity between the training data (used to build the model) and testing data (the new data scored by the computer).

The first method we used to identify LCPs makes use of the probability of the machine prediction as an indicator of prediction confidence. Given that machine-learning methods for scoring open-ended text are based on statistical methods, the probability level should be a proxy for prediction confidence. For this method, we used a logistic regression classifier to generate the prediction probability. For example, if the probabilities of logistic regression classifiers were found to be 0.6 and 0.9, then the computer scoring results would indicate ‘presence of concept’ for both cases. However, the ‘0.6’ probability is clearly less than ‘0.9’ because it is very close to the neutral prediction point (0.5). Thus, the probability of prediction can be calculated as the distance from 0.5. For example, a 0.3 probability indicates the ‘absence of concept’ and the confidence level is 0.2 (distance from 0.5); but 0.1 probability also indicates ‘absence of concept’. However, a probability of 0.1 suggests a more confident result has been achieved.

The second method we used for identifying LCPs is to quantify discordance levels among several different scoring models. Agreement among models could be considered as a high confidence prediction and disagreement could be considered as a LCP. The method we used trains several different scoring models using different machine learning classifiers (e.g., SVM and logistic regression) and controls feature extraction (e.g., n-gram selection). For example, the computer-scoring model of the “variation” concept showed the highest kappa values (i.e., human-computer agreement) using logistic regression classifiers, three n-grams
(uni, bigram, and Boolean), non-scaled features, non-self-feature, and no-joint method. On the other hand, the computer-scoring model of the “variation” concept showing the highest precision value used SVM classifier, two n-grams (unigram, and TFIDF), non-scaled feature, self-feature, and joint method. Thus, the kappa model and precision model were heterogeneous and were more likely to produce discordances. Nevertheless, if both models had agreement, then we can be more confidence about the prediction.

The third method we used for identifying LCPs is quantification of the semantic similarity between the training data (the data used to build the scoring model) and testing data (the data to be scored by the computer). Responses with low semantic similarity are likely to produce LCPs relative to responses with high semantic similarity. In sum, three different approaches could be used to attempt to predict LCPs in automated scoring of text.

**Study Design**

The current computer scoring system (ACSS) was trained using 10,270 human-scored responses to the ACORNS (Assessing COntextual Reasoning about Natural Selection) instrument (see Nehm et al., 2012) collected from 2978 students and evolution experts. The 10,270 dataset is known as the “training” dataset in our analyses. We also collected 3807 students’ responses to the new items from 1229 college students. This 3807 data set is known as the “testing” dataset in our analyses. We used Cohen’s kappa values—commonly used to quantify inter-rater agreement—as a measure of scoring accuracy. Cohen’s kappa coefficients range from 0.0 to 1.0 (Bejar, 1991). For this study, we focus on scoring models for six concepts that occur in the EvoGrader scoring system. To summarize, we used the three methods noted above to quantify LCPs for the six concepts in our ACSS.

We tested the efficacy of three methods to detect LCPs based on two different analysis levels: response level and corpus level (see Ha and Nehm 2016 for details). At the response level, the three methods will detect whether the model can confidently score each response. At the corpus level, ACSS users want to know if the ACSS is well suited to scoring their corpus (set of students’ responses). We used 3807 student responses to 36 different items in our analysis. Our ACSS shows the different level of accuracy across 36 items. In the second level of our analysis, we tested the correlation between ACSS accuracy in each corpus and the number of LCPs in each corpus. If significant and high correlations between the two factors are shown, then these methods can be used to find which corpus is not aligned with the ACSS.

**Findings**

*Probabilistic detection of low-confidence scoring predictions*

In order to use the probability of machine learning prediction as an indicator of LCPs, the probability level (i.e., distance from 0.5) of predictions must be different between non-prediction-error cases (i.e., agreed scoring between computer and human) and prediction-error cases (i.e., disagreements between computer and human). Figure 1 illustrates the means and standard errors of the probability levels of prediction between non-prediction-error cases and prediction-error cases. The probability levels of prediction for non-prediction-errors were over 0.4 and almost reached 0.5 (note that 0.5 is the maximum probability of prediction). On the other hand, the probability levels of prediction for prediction-error ranged between 0.2 and 0.3. Independent sample t-tests were performed and showed significant differences in
effect sizes for the six concepts (Cohen’s $d = 2.55$ [variation], 2.64 [heredity], 2.04[need/goal], 2.98 [use/disuse], 3.23 [adapt/acclimation]), $d = 1.28$ [differential survival/reproduction]).

![Figure 1](image)

**Figure 1.** Averages of the probability level of prediction between non-prediction-error cases and prediction-error cases.

We used four cutoff values to classify LCPs (distances are 0.1[lowest confidence], 0.2, 0.3, 0.4). For example, the ‘0.1 confidence’ means that the computer prediction had a probability between 0.4 and 0.6 (note that 0.5 is the neutral probability of prediction).

Table 1 shows the original kappa values between computer scoring and human scoring and kappa values if we removed LCPs and corrected the LCPs using a human grader. In addition, Table 1 shows the percentages of LCPs at each cutoff value.

It must be noted that ACSS users can use ACSS for two purposes. One is to understand students overall performance in a class (class level). The other is to measure each student’s performance (student level). For the first purpose, the instructor does not need to use LCPs and removes them from the computer scoring results. For the second purpose, instructors need to revise (rescore) the LCPs manually. Although it will require labor, it permits more accurate information about each student.

For example, the ‘adapt/acclimation’ scoring model shows 0.688 kappa value, which is not sufficient to be used in a classroom setting. The 0.1 confidence predictions comprised 1.1% of the sample. When the 1.1% of responses was removed from the dataset, the kappa expectedly increased (to 0.716) and when revised by a human grader the kappa increased to 0.720. Similarly, the 0.4 confidence predictions comprised 9.8% of cases. When the 9.8% LCPs were removed, the kappa increased to 0.896, and when the training data set was revised by human graders this caused the kappa to increase to 0.927. Checking only 9.8% of the sample will save more than 90% of human scoring effort. In addition, it can produce very strong kappa values (0.927). In sum, probabilities may be used to identify LCPs and either eliminate these cases or have a human rater rescore them.
Table 1. Original Kappa value, and kappa values when removing LCPs and correcting LCPs detected by probabilistic method

<table>
<thead>
<tr>
<th>Variatio</th>
<th>Heredit</th>
<th>Differe</th>
<th>Need/</th>
<th>Use/</th>
<th>Adapt/</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>y</td>
<td>ntial</td>
<td>Goal</td>
<td>Disuse</td>
<td>Acclim</td>
</tr>
<tr>
<td>Original Kappa value</td>
<td>0.861</td>
<td>0.900</td>
<td>0.701</td>
<td>0.807</td>
<td>0.625</td>
</tr>
<tr>
<td>0.1 level&lt;sup&gt;a&lt;/sup&gt;</td>
<td>% of LCPs&lt;sup&gt;b&lt;/sup&gt;</td>
<td>2.7%</td>
<td>1.2%</td>
<td>5.4%</td>
<td>2.3%</td>
</tr>
<tr>
<td>Removing LCPs&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.890</td>
<td>0.917</td>
<td>0.739</td>
<td>0.844</td>
<td>0.653</td>
</tr>
<tr>
<td>Correcting LCPs&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.893</td>
<td>0.918</td>
<td>0.753</td>
<td>0.849</td>
<td>0.656</td>
</tr>
<tr>
<td>0.2 level&lt;sup&gt;a&lt;/sup&gt;</td>
<td>% of LCPs&lt;sup&gt;b&lt;/sup&gt;</td>
<td>5.7%</td>
<td>2.8%</td>
<td>11.9%</td>
<td>4.2%</td>
</tr>
<tr>
<td>Removing LCPs&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.912</td>
<td>0.928</td>
<td>0.782</td>
<td>0.861</td>
<td>0.696</td>
</tr>
<tr>
<td>Correcting LCPs&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.917</td>
<td>0.931</td>
<td>0.808</td>
<td>0.871</td>
<td>0.706</td>
</tr>
<tr>
<td>0.3 level&lt;sup&gt;a&lt;/sup&gt;</td>
<td>% of LCPs&lt;sup&gt;b&lt;/sup&gt;</td>
<td>9.2%</td>
<td>4.5%</td>
<td>21.5%</td>
<td>7.0%</td>
</tr>
<tr>
<td>Removing LCPs&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.941</td>
<td>0.941</td>
<td>0.839</td>
<td>0.882</td>
<td>0.716</td>
</tr>
<tr>
<td>Correcting LCPs&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.947</td>
<td>0.946</td>
<td>0.873</td>
<td>0.894</td>
<td>0.738</td>
</tr>
<tr>
<td>0.4 level&lt;sup&gt;a&lt;/sup&gt;</td>
<td>% of LCPs&lt;sup&gt;b&lt;/sup&gt;</td>
<td>13.9%</td>
<td>8.0%</td>
<td>37.1%</td>
<td>14.6%</td>
</tr>
<tr>
<td>Removing LCPs&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.963</td>
<td>0.960</td>
<td>0.886</td>
<td>0.918</td>
<td>0.743</td>
</tr>
<tr>
<td>Correcting LCPs&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.968</td>
<td>0.964</td>
<td>0.928</td>
<td>0.931</td>
<td>0.769</td>
</tr>
</tbody>
</table>

<sup>a</sup>‘less than 0.1’, ‘less than 0.2’ and etc. mean that the ACSS prediction with the probability respectively between 0.4 and 0.6, between 0.3 to 0.7, and so forth (note that 0.5 is the neutral probability of prediction).

<sup>b</sup>‘% of LCPs’ refers to the percentage of the responses with the probability ranged from 0.4 to 0.6, and so forth.

<sup>c</sup>‘Removing’ means that the cases with low confidence (% of cases) were removed the data and new calculated kappa with removed data.

<sup>d</sup>‘Correcting’ means that the cases with low confidence (% of cases) were revised (rescored) by human graders and new calculate kappa with revised data.

We also tested the correlation between ACSS accuracy in each corpus (percentage of error in each corpus) and the numbers of LCPs in each corpus. We found significant and high correlations between percentages of errors in each corpus and the average probability for each corpus. We also tested the correlations between ACSS accuracy and the percentage of LCPs detected by four different probability methods (i.e., 0.1 level to 0.4 level, Table 2). Given the high correlations, the regression equation to predict the potential percentages of prediction errors in corpus can be built. This could enable instructors to figure out whether the ACSS can score their data.

Table 2. The Pearson correlation between the average of confidence, the percentage of cases, and the percentage of errors in 36 corpora (n = 36)

<table>
<thead>
<tr>
<th>Scoring model</th>
<th>Average of confidence</th>
<th>% of cases with less than 0.1 confidence</th>
<th>% of cases with less than 0.2 confidence</th>
<th>% of cases with less than 0.3 confidence</th>
<th>% of cases with less than 0.4 confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variation</td>
<td>0.853&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>-0.781&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>-0.828&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>-0.845&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>-0.850&lt;sup&gt;‡&lt;/sup&gt;</td>
</tr>
<tr>
<td>Heredity</td>
<td>0.561&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>-0.344&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>-0.593&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>-0.521&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>-0.474&lt;sup&gt;‡&lt;/sup&gt;</td>
</tr>
</tbody>
</table>
Differential survival 0.716† -0.367† -0.424‡ -0.347† -0.622‡  
Need/Goal 0.088 -0.139 -0.071 0.001 -0.074  
Use/Disuse 0.785‡ -0.514‡ -0.758‡ -0.799‡ -0.785‡  
Adapt/Acclimation -0.076 -0.429† 0.049 -0.075 0.101  

†p < 0.01, †p < 0.05

Multi-model detection of low-confidence scoring predictions

Our second set of analyses tested the discrepancies among scoring models methods as indicators of LCP and produced promising results. The discrepancy of multiple heterogeneous scoring models was significantly different between non-prediction-errors and prediction-errors using the independent sample t-test. Figure 2 illustrates that non-prediction error cases (NE in Figure 2) show higher agreements between scoring models than prediction-error cases (E in Figure 2). In summary, among 24 tests, we found 13 huge effect sizes (d > 1.45), 1 very large effect size (d >=1.10 and <1.45), 5 large effect sizes (d >=.75 and <1.10), and 4 medium effect sizes (d >=.40 and <.75). In sum, multi-model method can also be the method to detect low-confidence scoring predictions.

Figure 2. Averages of the discrepancy among multiple heterogeneous scoring models between non-prediction-error cases and prediction-error cases (‘Kappa & Precision’ means the agreements between the computer scoring model showing highest kappa values and that showing highest precision value. ‘Kappa &, Precision & Recall’ means that agreements occurred among all three computer scoring models)

Table 3 showed original kappa values between computer scoring and human scoring and kappa values if we removed LCPs (detected by multi-model methods) and corrected LCPs by a human grader. For example, the original kappa of ‘adapt/acclimation’ model was 0.683. However, when 5.3% discrepant data between kappa and precision scoring models were removed, the kappa increased to 0.853 and when the training data set was revised by human graders, the kappa increased to 0.916. Similarly, when the 9.5% LCPs (discrepant data among kappa, precision, and recall models) were removed, the kappa increased to 0.888, and
when scored by the human grader, the kappa increased to 0.940.

Table 3. Original Kappa value, and kappa values when removing LCPs and correcting LCPs detected by multi-model method

<table>
<thead>
<tr>
<th></th>
<th>Variatio n</th>
<th>Heredit y</th>
<th>Differe nt survival</th>
<th>Need/ Goal</th>
<th>Use/ Disuse</th>
<th>Adapt/ Acclimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Kappa value</td>
<td>0.861</td>
<td>0.900</td>
<td>0.722</td>
<td>0.807</td>
<td>0.625</td>
<td>0.683</td>
</tr>
<tr>
<td>Kappa &amp; Precision</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% of LCPs&lt;sup&gt;a&lt;/sup&gt;</td>
<td>10.8%</td>
<td>4.9%</td>
<td>5.3%</td>
<td>2.0%</td>
<td>6.5%</td>
<td>5.3%</td>
</tr>
<tr>
<td>Removing LCPs</td>
<td>0.895</td>
<td>0.939</td>
<td>0.746</td>
<td>0.850</td>
<td>0.781</td>
<td>0.853</td>
</tr>
<tr>
<td>Correcting LCPs</td>
<td>0.904</td>
<td>0.947</td>
<td>0.759</td>
<td>0.855</td>
<td>0.928</td>
<td>0.916</td>
</tr>
<tr>
<td>Kappa &amp; Recall</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% of LCPs&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.8%</td>
<td>3.1%</td>
<td>10.2%</td>
<td>2.7%</td>
<td>3.7%</td>
<td>4.8%</td>
</tr>
<tr>
<td>Removing LCPs</td>
<td>0.891</td>
<td>0.927</td>
<td>0.791</td>
<td>0.821</td>
<td>0.649</td>
<td>0.713</td>
</tr>
<tr>
<td>Correcting LCPs</td>
<td>0.894</td>
<td>0.931</td>
<td>0.812</td>
<td>0.826</td>
<td>0.671</td>
<td>0.730</td>
</tr>
<tr>
<td>Precision &amp; Recall</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% of LCPs&lt;sup&gt;a&lt;/sup&gt;</td>
<td>11.1%</td>
<td>5.1%</td>
<td>12.5%</td>
<td>3.4%</td>
<td>10.0%</td>
<td>8.9%</td>
</tr>
<tr>
<td>Removing LCPs</td>
<td>0.901</td>
<td>0.930</td>
<td>0.787</td>
<td>0.851</td>
<td>0.898</td>
<td>0.833</td>
</tr>
<tr>
<td>Correcting LCPs</td>
<td>0.910</td>
<td>0.938</td>
<td>0.813</td>
<td>0.857</td>
<td>0.972</td>
<td>0.902</td>
</tr>
<tr>
<td>Kappa &amp; Precision</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% of LCPs&lt;sup&gt;a&lt;/sup&gt;</td>
<td>12.4%</td>
<td>6.5%</td>
<td>14.0%</td>
<td>4.1%</td>
<td>10.1%</td>
<td>9.5%</td>
</tr>
<tr>
<td>Removing LCPs</td>
<td>0.914</td>
<td>0.951</td>
<td>0.804</td>
<td>0.858</td>
<td>0.908</td>
<td>0.888</td>
</tr>
<tr>
<td>Correcting LCPs</td>
<td>0.923</td>
<td>0.958</td>
<td>0.831</td>
<td>0.865</td>
<td>0.975</td>
<td>0.940</td>
</tr>
</tbody>
</table>

<sup>a</sup>percentage of LCPs in multi-model approach means the disagreements of scoring between two heterogeneous models.

We calculated correlations between ACSS accuracy (percentage of error in each corpus) in each corpus and the number of discrepancies among multiple models (four cases) in 36 corpora (Table 4). We found significant and high correlations between them. Given the high correlations, this method will be used to predict the potential percentage of prediction error in the corpus.

Table 4. The Pearson correlation between discrepancy rate of multiple models (four cases) and the percentage of errors in 36 corpora

<table>
<thead>
<tr>
<th></th>
<th>Kappa &amp; Precision</th>
<th>Kappa &amp; Recall</th>
<th>Precision &amp; Recall</th>
<th>Kappa &amp; Precision &amp; Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variation</td>
<td>0.752&lt;sup&gt;†&lt;/sup&gt;</td>
<td>0.705&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>0.774&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>0.784&lt;sup&gt;‡&lt;/sup&gt;</td>
</tr>
<tr>
<td>Heredity</td>
<td>0.545&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>0.590&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>0.230</td>
<td>0.619&lt;sup&gt;‡&lt;/sup&gt;</td>
</tr>
<tr>
<td>Differential survival</td>
<td>0.114</td>
<td>0.472&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>0.395&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>0.396&lt;sup&gt;‡&lt;/sup&gt;</td>
</tr>
<tr>
<td>Need/Goal</td>
<td>0.485&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>0.165</td>
<td>0.487&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>0.471&lt;sup&gt;‡&lt;/sup&gt;</td>
</tr>
<tr>
<td>Use/Disuse</td>
<td>0.629&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>0.438&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>0.687&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>0.701&lt;sup&gt;‡&lt;/sup&gt;</td>
</tr>
<tr>
<td>Adapt/Acclimation</td>
<td>0.618&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>0.521&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>0.543&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>0.642&lt;sup&gt;‡&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>†</sup>p < 0.01, <sup>‡</sup>p < 0.05
Semantic similarity detection of low-confidence scoring predictions

Lastly, we explored the relationship between semantic similarity (of training data and testing data) and scoring accuracy. We applied a total of 8 different semantic similarity methods based on the feature extraction methods (e.g., masking, TFIDF etc.). In key concept score, we found correlation coefficients ranging from 0.124 to 0.198 (p < 0.01) between semantic similarity and amount of errors in each item (note that high semantic similarity means two corpora are different; thus the correlation coefficient is positive). However, we were not able to find significant and meaningful correlation in for naïve ideas. At the response level, the semantic similarity method is not suitable to be an indicator of LCPs.

Table 5. The Pearson correlation between semantic similarity and amount of errors in each response (n = 3807)

<table>
<thead>
<tr>
<th>Masking data</th>
<th>Method</th>
<th># of errors in key concept scoring</th>
<th># of errors in naïve idea scoring</th>
<th># of errors in whole scoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-mask</td>
<td>Count</td>
<td>0.124*</td>
<td>0.001</td>
<td>0.111*</td>
</tr>
<tr>
<td></td>
<td>TFIDF</td>
<td>0.165*</td>
<td>0.012</td>
<td>0.151*</td>
</tr>
<tr>
<td></td>
<td>Boolean</td>
<td>0.163*</td>
<td>0.020</td>
<td>0.153*</td>
</tr>
<tr>
<td></td>
<td>LSI</td>
<td>0.178*</td>
<td>-0.034†</td>
<td>0.138*</td>
</tr>
<tr>
<td>Mask</td>
<td>Count</td>
<td>0.186*</td>
<td>0.037†</td>
<td>0.178*</td>
</tr>
<tr>
<td></td>
<td>TFIDF</td>
<td>0.198*</td>
<td>0.014</td>
<td>0.179*</td>
</tr>
<tr>
<td></td>
<td>Boolean</td>
<td>0.192*</td>
<td>0.030</td>
<td>0.179*</td>
</tr>
<tr>
<td></td>
<td>LSI</td>
<td>0.194*</td>
<td>-0.006</td>
<td>0.163*</td>
</tr>
</tbody>
</table>

*p < 0.01, †p < 0.05

Table 6 illustrated the correlation coefficient between semantic similarity and total amount of errors in the corpus (36 corpora). The coefficients ranged between 0.686 to 0.919 for key concepts and 0.357 to 0.508 for naïve ideas. Unlike response level, the semantic similarity method in corpus level could be a suitable indicator of LCPs.

Table 6. The Pearson correlation between semantic similarity and amount of errors in each corpus (n = 36)

<table>
<thead>
<tr>
<th>Masking data</th>
<th>Method</th>
<th># of errors in key concept scoring</th>
<th># of errors in naïve idea scoring</th>
<th># of errors in whole scoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-mask</td>
<td>Count</td>
<td>0.686*</td>
<td>0.357†</td>
<td>0.679*</td>
</tr>
<tr>
<td></td>
<td>TFIDF</td>
<td>0.778*</td>
<td>0.427†</td>
<td>0.775*</td>
</tr>
<tr>
<td></td>
<td>Boolean</td>
<td>0.822*</td>
<td>0.393†</td>
<td>0.805*</td>
</tr>
<tr>
<td></td>
<td>LSI</td>
<td>0.807*</td>
<td>0.412†</td>
<td>0.797*</td>
</tr>
<tr>
<td></td>
<td>Count</td>
<td>TFIDF</td>
<td>Boolean</td>
<td>LSI</td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>------</td>
</tr>
<tr>
<td>Mask</td>
<td>0.882†</td>
<td>0.892†</td>
<td>0.882†</td>
<td>0.919†</td>
</tr>
<tr>
<td></td>
<td>0.467†</td>
<td>0.483†</td>
<td>0.459†</td>
<td>0.508†</td>
</tr>
<tr>
<td></td>
<td>0.875†</td>
<td>0.887†</td>
<td>0.873†</td>
<td>0.916†</td>
</tr>
</tbody>
</table>

*p < 0.01, †p < 0.05

**Contributions and General Interest**

Our experiments revealed that the three methods -- probability of machine-learning prediction, the discrepancy of multiple scoring models, and semantic similarity—can be used as indicators of low confidence predictions (LCPs). Using these three methods, future text analysis programs could include a warning or tagging of low confidence predictions. Then, instructors could more effectively use data derived from text analysis programs. For example, our findings indicated that an instructor could improve scoring accuracy (kappa =0.94) when s/he checks a small subset of data (in this particular case 10% of the dataset). Alternatively, instructors could simply remove the LCPs and use the higher-quality data to make instructional decisions. Automated detection of LCPs could serve as a “third intelligence” to evaluate the accuracy of automated scoring of text. The three methods we studied could help to improve the quality of automated analysis of text in science education contexts.

**References**

Ha, M., Nehm, R. H. (2016). Journal of Science Education and Technology


**Author Note**

This work is supported by the National Science Foundation (TUES-1322872). 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.
Discussant Comments

Jonathan Osborne, Graduate School of Education, Stanford University

Why does assessment matter? For three reasons. First, there is assessment of learning with which we are all familiar. Here the driver is very much that we want better (and cheaper) assessment of learning. Second, there is assessment for learning – the formative goal. To be useful for learning, however, the feedback to the teacher needs be swift and timely. Computerized assessment offers that possibility – for once an example where technology might make an educational difference. Third, assessments matter because assessments operationalize constructs – that is that we live in an era that too often the only thing that counts is what can be counted. And the limitations of pencil and paper mean that what we would like to be counted is often limited by the limits of the technologies. The affordances of technology mean that we can ask a much wider and much richer range of questions – particularly those which might demand higher order reasoning skills.

Thus, the work here exploring how to build computerized assessments gets us away from the limits of multiple choice – what you might call beyond multiple choice – is important. That is not to say that all multiple choice is necessarily bad – it is possible to ask quite demanding multiple choice questions – and carefully designed multiple choice questions with good distractors can give a lot of information about the nature of student misunderstandings. All of these papers, however, represent a body of work which is trying to push what it is possible to do with assessment and are to be welcomed.

So to turn to these papers – what do I see as the major themes emerging? In the work of Matthew Steele and his colleagues, I see an attempt to compare the product of standard multiple choice questions to constructed responses and some of the issues that emerge from that. For me – the main one is that when asking a constructed response, the question is much more open to interpretation of what is being asked for. You may not get the responses you want or expect. That means that you either have to tweak the question to communicate its intent or you have to decide on whether what you get is acceptable. It also means that developing good items is a time consuming, intensive and expensive task. In turn, that means we need to think about how the products of such labor can be shared more widely.

The second of these papers by Molly Stuhlsatz and her colleagues is an attempt to do the same thing but assess teachers’ pedagogical content knowledge. This is a brave and ambitious attempt to assess a nebulous construct using a very different approach and is the kind of work that needs to be done – especially as building a better concept of this construct matters if we are to help clarify our understanding. Here the issue with this work is not the coding scheme but the construct that it is trying to assess. For instance, the paper makes an argument that the knowledge is contextualized but, if so, how can we assess it out of the context in which it is used and how can we build a common professional language? And while I agree with the argument that there are no universal characteristics that can define a good teacher, I do think
that there are universal characteristics that can define a professional discourse and that, in this case, we should be seeking to operationalize them through assessment. The other question I am left asking about this work is that there is quite a lot of variation in the reliabilities – what is that a reflection of – the difficulty of operationalizing the construct? In which case, it will be interesting to see if you get better reliabilities with a machine.

Then the third paper is by Moscarella and colleagues. This is looking at the grading of students understanding of the processes underlying genetic information. This addresses an important area as genetics is conceptually difficult – it is abstract and about the intangible and many confusions exist. I would have liked to know more about the figures for reliability. This is a crucial question as, rather like the data on whether self-driving cars have accidents or not is going to make you feel happy about using them, figures on the reliability when compared to a human will make you feel happy about using and trusting the technology. What I do think is interesting about this work is the path maps which show us how understanding one feature is consequential for understanding another – this is really informative and novel providing valuable information for teachers. However, if such information is to be used formatively, it is not enough just to gather it, teachers must consider how to use it.

Finally, there is Ha and Nehm’s paper which is an interesting methodological attempt to study why there is divergence between computers and humans and how we can identify those for which there is Low Confidence Predictions. In essence, how can we make the machine even better than some of the machines are doing. They explore three methods for detecting these response. First, is the concept of a probability distance. Second was different methods of scoring it automatically and then looking for the discordance between the two. And third is a quantification of the semantic similarity between the response and the training data for the machine.

The value of this study for me was not so much the findings – which were interesting but the questions it raised, which were:

- What is an acceptable kappa level for a machine? Do we always have to compare it to a human rating?
- Can we ever rely totally on a machine or will there always be a dependence on human scoring?
- How dependent are these results on the domain? Would the findings be different with a different domain?
- The semantic similarity method does not appear to work but what is it about this approach which does not work?

In all, a stimulating set of papers which I see as working at the leading edge of assessment in science. Hopefully those of us who work on the trailing edge can benefit. If you are not sure what I mean by that comment, then here are examples of leading edge technology, where mistakes were made and from which the rest of us learnt.
• London Underground versus any other form of underground metro system
• British versus European canals
• American NTSC color television versus the European PAL

Thus, there is much to be learnt from these papers about how the field might advance towards the goal of using technology to improve assessment of learning. More important though is the potential it offers to improve assessment for learning.