Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 24 Number 7, August 2019 ISSN 1531-7714

Deconstruction of Holistic Rubrics into Analytic Rubrics for Large-Scale Assessments of Students' Reasoning of Complex Science Concepts

Lauren N. Jescovitch[1], Emily E. Scott[2], Jack A. Cerchiara[2], Jennifer H. Doherty[2], Mary Pat Wenderoth[2], John E. Merrill[3], Mark Urban-Lurain[1], & Kevin C. Haudek[1]

[1] CREATE for STEM Institute at Michigan State University, [2] Department of Biology at University of Washington, [3] Department of Microbiology & Molecular Genetics at Michigan State University

Constructed responses can be used to assess the complexity of student thinking and can be evaluated using rubrics. The two most typical rubric types used are holistic and analytic. Holistic rubrics may be difficult to use with expert-level reasoning that has additive or overlapping language. In an attempt to unpack complexity in holistic rubrics at a large scale, we have developed a systematic approach called deconstruction. We define deconstruction as the process of converting a holistic rubric into defining individual conceptual components that can be used for analytic rubric development and application. These individual components can then be recombined into the holistic score which keeps true to the holistic rubric purpose, while maximizing the benefits and minimizing the shortcomings of each rubric type. This paper outlines the deconstruction process and presents a case study that shows defined concept definitions for a hierarchical holistic rubric developed for an undergraduate physiology-content reasoning context. These methods can be used as one way for assessment developers to unpack complex student reasoning, which may ultimately improve reliability and validation of assessments that are targeted at uncovering large-scale complex scientific reasoning.

Constructed response (CR) assessment items, which require students to answer a question in their own words, allow for a more in-depth analysis of students' content understanding and can elicit students' higher order thinking than fixed response items (e.g., multiple choice; Allen and Tanner, 2006; Jonsson and Svingby, 2007). However, CR answers can be difficult to interpret and time consuming to provide feedback – particularly in a large-scale effort. To investigate authentic student thinking and performance, researchers are often exploring which CR coding techniques are most efficient and appropriate (Hunter et al., 1996). Rubrics are typically used to evaluate CR assessments (Haudek et al, 2015; Moskal, 2000), and can enhance the reliability of large-scale coding efforts (Jonsson and Svingby, 2007).

In our effort to create machine learning models for CR assessments, our models (based on holistic rubrics) were underperforming. To address this concern, we decided to explore deconstructing the holistic rubrics into analytic rubrics. This article proposes a deconstruction method that maximizes the positive attributes of both holistic and analytic rubric development, while minimizing their drawbacks, for evaluating complex, scientific undergraduate student thinking. Deconstruction will not only aid in our group's for successful computerized scoring models for student

reasoning, but may also be applicable for other assessment research.

Rubric Development

Researchers often develop codes from emergent patterns found in the data. Assessment developers then need to interpret the data by not only condensing raw data into key concepts, but also arranging those concepts into a logical, systematic explanatory scheme in an attempt to capture and categorize complex thinking (Corbin and Strauss, 2008). The number of codes depend on the nature of the data, which coding method you select for analysis, and how detailed you want or need to be in evaluation (Saldana, 2009). These codes can then be clustered based on rubric type (Moskal, 2000). Rubrics can be used to validate and make reliable assessment of complex student performance and promote learning (Jonsson and Svingby, 2007; Panadero and Jonsson, 2013).

Validity and reliability are critical aspects of assessment development, and thus should be evaluated during development and alignment of CR items and rubrics. Validity refers to 'the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests (p. 11; AERA, APA & NCME, 2014).' Validation is an ongoing process that continues from the beginning of assessment design through development and implementation. There are two forms of validity evidence, empirical and procedural, that span four sources of validity evidence: test content, internal structure, response processes, and relations to other measures (AERA, APA & NCME, 2014). For this report, we focus on test content and response process validity, as we develop, align and implement rubrics for student writing in response to science items. Test content evidence contains frameworks and the relative importance of specific aspects of item content, and alignment studies. Concerns of construct are either underrepresentation or tainted constructs that need to be addressed. Response processes are defined cognitive skills and rigor, clear directions, and analysis of process data, such as how raters apply scoring criteria. However, response processes can be mismatched between actual and intended cognitive processes that an item elicits or test-taking strategies.

Reliability is the ability for scorers to consistently score a given response in a similar way (AERA, APA & NCME, 2014; Humphry and Heldsinger, 2014). Intra-rater and inter-rater reliability (IRR) are ways to measure scorer reliability. Scorer reliability increases as more restrictions or clarifications are made on scoring criteria. Other features that also improve reliability include: establishing descriptions and rules of scoring criteria in advance, discussion of differences in interpretation and negotiation process, and appropriateness of assessed population. It is recommended to report each code, sub-code, and combination of codes for reliability and precision (AERA, APA & NCME, 2014). Well-designed rubrics are crucial to produce valid and reliable assessment results (Moskal and Leydens, 2000).

Rubrics are a critical part of reliable assessment of complex student performance and promote interpretation of student learning (Panadero and Jonsson, 2013). Assessment developers interpret emergent patterns in data by condensing raw data into key concepts and arranging those concepts into a logical, systematic explanatory scheme, in an attempt to capture and categorize complex thinking (Corbin and Strauss, 2008). The number of codes depend on the nature of the data, coding method, and the level of detail targeted in evaluation (Saldana, 2009). These codes are then clustered into a rubric (Moskal, 2000). The two most common, rigorous rubric types used are holistic and analytic (Figure 1; Allen and Tanner, 2006; Jonsson and Svingby 2007) which we describe in more detail below.

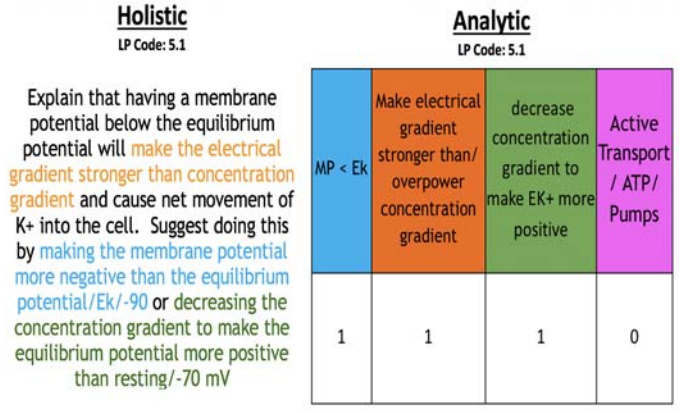

Figure 1. Comparison of holistic and analytic rubrics outlined by a 5.1 learning progression code.

Holistic

Holistic rubrics are generally used for judgement of broader quality of student thinking (Moskal, 2000), can be thought of as awarding a global score, such as a letter grade or rating number (Hunter et al., 1996), and regarded as suitable for evaluating open-ended and higher-order skills (Hunter et al., 1996; Singer and LeMahieu, 2011). Holistic rubrics have rubric bins that

are mutually exclusive; thus, each response can only have one score. A bin is defined as an organizational pattern that clusters responses to a common concept in the rubric. This code can be different scales of measurement, for example, nominal in classifying concepts as normative or non-normative, or ordinal for levels in a learning progression. Holistic coding is often used for large-scale assessments because they are assumed to be a fast and accurate tool for qualitative ratings (Jonsson and Svingby, 2007).

If the patterns described in a holistic rubric are ill-defined, or too generalized, it may be difficult to achieve high reliability (Jonsson and Svingby, 2007). Generalized feedback typically has long descriptors with extended examples as scoring criteria, but may be less diagnostic in determining which piece is missing in student thinking (Hunter et al., 1996).

Analytic

Analytic rubrics are defined as evaluating responses on multiple dimensions by using multiple bins that are not mutually exclusive and are typically binary in coding. Each analytic rubric bin is designed to represent a single concept or attribute. A common way of applying analytic rubrics would be to score for the absence or presence (e.g. 0 or 1) of some attribute. Coders can be more discriminate in a fine-grained way across concepts (Humphry and Heldsinger, 2014; Jonsson and Svingby, 2007). Multiple concepts or attributes can be present within the same response; thus, the response could be present in multiple analytic bins – similar to a checklist (Moskal, 2000).

Research supports the possibility of combining analytical bins into holistic codes (Hunter et al., 1996; Singer and LeMahieu, 2011). But this may not be preferable if the separate dimensions are only summarized (Waltman, Kahn, and Koency, 1998). If bins are too narrow, researchers can deem the bin not acceptable because the bin may lose the essence or 'spirit' of the critical concept being considered. Analytic bins may become too narrow in focus on conceptual tendencies that restrict the coders range of choice, and may include a considerable zone of variation in rank-ordering code outcomes (Hunter et al., 1996).

In our work, we explore the trade-offs between holistic and analytic rubrics in order to enhance the creation and evaluation of large-scale CR assessments. In this paper we discuss how we may maintain both validity and reliability in development (IRR in internal structure) while disentangling the proper organization of expert-like reasoning components and complexity in scientific student writing. Specifically, we will present a deconstruction framework and provide a detailed example from our research.

Deconstruction

We decided to use the term deconstruction to emphasize the analytic nature of the task. Overall, our process resembles a disassembly of something complex into finer-grained components or pieces. Here, we define deconstruction as the process of breaking apart levels and criteria contained in a holistic rubric into individual, conceptual components which can be used for analytic rubric development and application. These individual components may then be recombined into a single holistic score – which keeps true to the purpose, and represents the unique complexity and diagnostic purposes – of the holistic score. Overall, this is a top-down approach where the target constructs are the holistic codes; thus, the whole configuration determines the character of the parts instead of vice-versa.

Advantages and Disadvantages to the Deconstruction Approach

The deconstruction process is inherently difficult because of the requirement for thorough familiarity of both the range of possibilities and the elements comprising an expert answer in student thinking (Allen and Tanner, 2006). However, according to Allen and Tanner (2006), this process can: 1) make coding more reliable; 2) clarify vagueness in coding criteria and variable interpretations in bins and/or specific concepts; 3) find important, distinguishing features that experts want to capture in writing; 4) display data in multiple ways; 5) solidify organization of concepts into holistic schema; and 6) allow for expansion of both, quantitative and qualitative, interpretation of results.

While there are a lot of positive features about deconstruction, there are also some concerns with this process to generate an analytic coding scheme, it: 1) may not capture the breadth of student reasoning; 2) may lose some of the original concepts; 3) may oversimplify a concept or have a loss of complexity; or 4) may require a prohibitive amount of extra time and effort, which is expensive (Waltman, Kahn, and Koency, 1998). Thus, it is important to keep the richness of the codes dependent on the end-users of the resulting information.

Examples from other studies

Other, large-scale research initiatives have alluded to using deconstruction to enhance rubric reliability and acknowledged the challenges of employing such an approach. Liu et al. (2014) applied methods that they referred to as 'transformation of holistic rubrics to concept-based analytic rubrics' for improvement of automated analysis of CR items, but only briefly described this process, and individual components were not directly derived from the holistic rubric.

Haudek et al. (2015) decided to create an analytic rubric to clarify criteria based on the major sources of disagreement among coders to reduce ambiguity and/or subjectivity; although, they also only briefly described this process. Using the resulting analytic rubrics provided a mechanism to uncover more details of the heterogeneity of student thinking.

Urban-Lurain and Weinshank (1999) developed a performance-based rubric for a large-scale undergraduate course. They defined the finest possible granularity of a criterion so multiple graders could quickly and consistently evaluate that criterion over a large number of student responses – an undergraduate class of 1700 students. The conceptual integrity of the original rubric was maintained, but the concepts were defined first, then assessments and rubrics were created concurrently.

A writing assessment study by Hunter et al. (1996) compared holistic versus analytic rubrics, and found that over half of the papers were given an identical rating holistically and analytically on a 5-point scale. They recommend using both holistic coding - for an overall measure of competence - and analytic coding - for feedback to individuals and as a means of reducing error in measurement.

Deconstruction Framework

We developed a deconstruction framework that consists of two exploratory and sequential cycles (Figure 2). Cycle 1 is the holistic rubric development. This cycle starts with data collection of student CR. These data are then analyzed with an emergent coding schema to generate a holistic coding rubric. For example, this cycle can follow the NRC Assessment Triangle (2001), criteria from Mohan et al. (2009) and Anderson (2008), to develop an aligned assessment. Thus, the cycle can have multiple iterations in order to improve validity and reliability of the assessment and rubric until experts are satisfied.

Cycle 2 is the rubric deconstruction and analytical rubric development. This cycle begins with two or more experts independently identifying the individual, or fine-grained, conceptual components contained within the holistic rubric developed from Cycle 1. Results of this analysis are used by experts in Phase 2 to design a scheme that shows how conceptual pieces are put together, such as a matrix that shows Boolean logic, and how these pieces relate to the original, holistic coding schema. Rather than just adding up numbers across the Table 2rubric, Boolean logic allows the developer to finely manipulate the presence or absence of specific concepts that are combined to a single holistic code.

The visual representation (Table 1) for the Boolean logic shows the individual concepts on the first row and the holistic codes in the first column. A '1' indicates that a concept is considered important and *must* be present in order to be given that holistic score. Sometimes, holistic levels may either require multiple concepts (exemplified by '*and*' statement in Boolean logic) or allow any of the multiple concepts to be sufficient for a code, as a 'pick one' (exemplified by '*or*' statement in Boolean logic). The '*and*' Boolean logic goes between '1's marked between columns on the same row. The '*or*' Boolean logic is shown with 1's on different rows with the same shaded background and has the same holistic indicator. For example, both Concept 1 and Concept 2 would be required for a holistic code of 5.1. The presence of any one of the Concepts 3, 4 or 5 would be sufficient for the holistic code of 2.1.

The individual concepts and Boolean logic are then validated by the experts coding a subset of responses,

Table 1. Proposed visual representation for Boolean logic.

Holistic Code	Concept 1	Concept 2	Concept 3	Concept 4	Concept 5	Concept 6
5	1	1				
4	1					
3.2		1			1	
3.1		1	1			
3.1			1	1		
2.2				1	1	
2.2			1		1	
2.1			1			
2.1				1		
2.1					1	
1						1

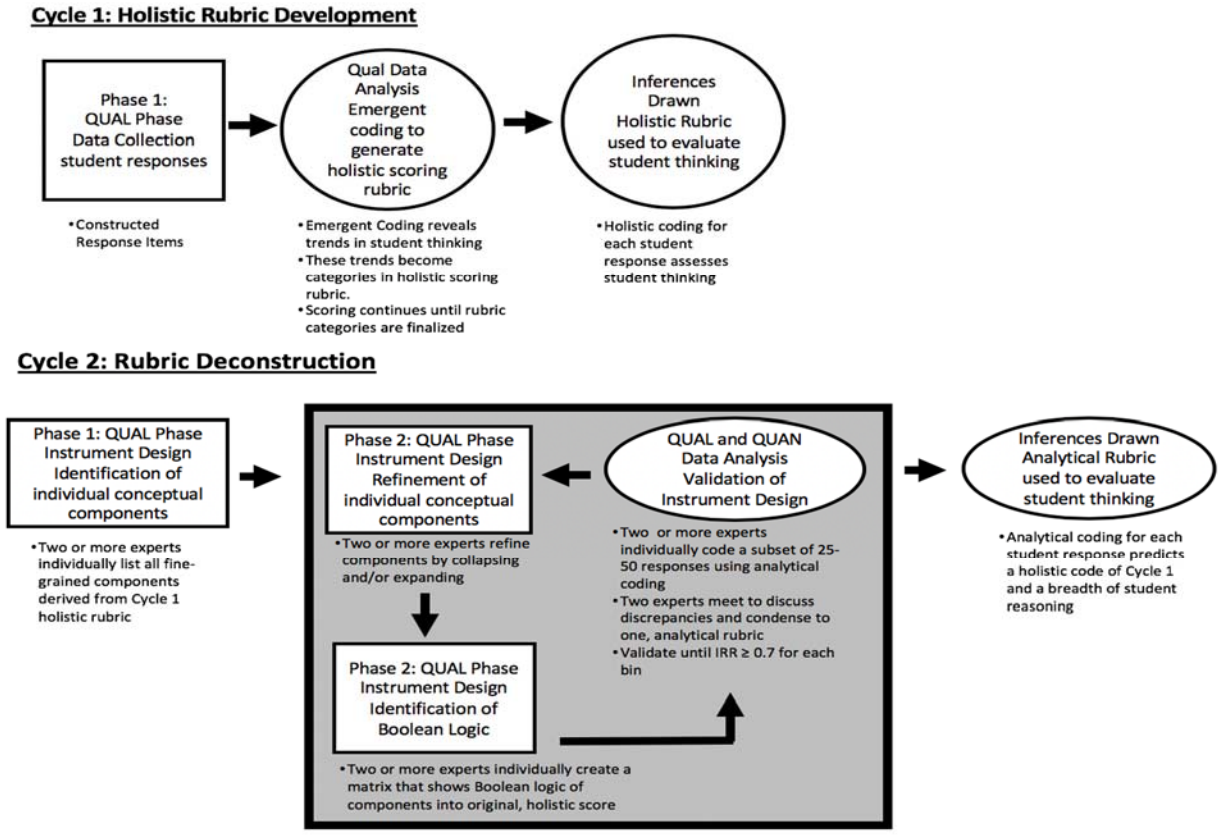

Figure 2. Proposed Deconstruction Framework. Procedural diagram based on Exploratory Sequential Design (Creswell, 2014). Shown is the exploratory sequential design of Cycle 1 that includes the process for holistic rubric development and application. This flows into Cycle 2 and the process of deconstruction and rubric refinement (highlighted box) that is used for analytic rubric development.

discussing discrepancies and condensing the proposed draft analytic rubrics. Then, experts will evaluate for conceptual meaningful clusters in the context of the assessment question; if some of the codes are infrequent or non-essential as a defining feature in distinguishing student reasoning, then these codes can be removed or condensed into another bin. This cycle can have multiple iterations in order to address and improve validity and reliability of the rubric until experts are satisfied. For example, two coders can meet to discuss the rubric, apply the codes independently to responses, and then meet again to discuss findings, concerns, and address IRR.

Once both Cycles are completed, Boolean logic is used with the analytic codes to calculate a holistic value which is compared to the original code. Discrepancies between these holistic values are evaluated to determine if the rubrics are aligned or if more revision is needed.

Deconstruction Case Study

Here we present an example of the deconstruction process from our research using a formative assessment item about ion flux intended for use in undergraduate physiology courses.

In physiology, seven core concepts have been identified in the discipline (Michael and McFarland, 2011). One of the most applicable to physiology, and more broadly, is *flux* and describes the passive flow of substances and heat down gradients (Michael et al., 2017). We developed a series of CR items to assess one progress variable within a developing *flux* learning progression (LP) framework that captures principle-based reasoning (Doherty et al., 2019). LPs are empirical cognitive frameworks that describe how student thinking about a topic gains sophistication through time (Corcoran et al. 2009). LPs can provide reference points for student progress and levels of achievement. LPs are built using evidence about student reasoning collected by a complex and iterative routine of LP development,

assessment item and rubric development, data collection, human coding and re-alignment of LPs.

The *flux* assessment item used as our example is named 'EION' (Figure 3). EION assesses undergraduate students reasoning about ion *flux* using both concentration and electrical gradients. EION was administered to 1470 undergraduate students taking physiology and biology courses at two community colleges and eight colleges and universities in the USA. Student responses were then analyzed using Cycle 1 in the deconstruction framework (Figure 2). Responses were categorized to develop holistic codes which were aligned with the proposed LP, so that the rubrics are hypothesized to reflect the LP. Because of the iterative process used to develop the LP, there is no 'ground-truth' to base the codes; thus, the rationale is based on expert inferences. Student responses were given only one (of five) holistic codes corresponding to the type of reasoning they provided of the hierarchical coding-scale (Table 2). Nine sublevel codes (e.g., 2.1, 2.2, etc.) were made to improve coding reliability. Sublevels are

Table 2. Holistic rubric of EION.

Level	Indicator	Student Exemplar
5	**5.1)** Explain that having a membrane potential below the equilibrium potential will make the electrical gradient stronger than concentration gradient and cause net movement of K+ into the cell. Suggest doing this by making the membrane potential more negative than the equilibrium potential/Ek/-90 or decreasing the concentration gradient to make the equilibrium potential more positive than resting/-70 mV.	*A more negative membrane potential (less than -91 mV), increased outer concentration, and decreased inner concentration could cause flow into the cell / The first option would allow the electrical forces to dominate the chemical forces and cause movement in. The other two options could cause the concentration gradient to be less extreme decreasing chemical forces (or even flipping them such that they no longer oppose electrical forces)*
4	**4.1)** Suggest increasing the electrical gradient (i.e., making the membrane potential more negative) or decreasing the concentration gradient **in order to** make electrical forces stronger than concentration forces and cause net movement of K+ into the cell.	*1. Make K+ concentration outside bigger than that of the inside 2. Make membrane potential much more negative 1. This will flip the concentration gradient so that the K+ flows inside and the electrical gradient will cause K+ to flow inside 2. This will cause the electrical gradient to be bigger than the concentration gradient, so K+ flows inside*
	4.2) Suggest decreasing the membrane potential below the equilibrium potential (i.e., more negative) to cause net movement of K+ into the cell (often to 'reach equilibrium'). May also suggest reversing the concentration gradient (as in 1.1) but treats concentration and electrical gradients as independent.	*Increase K+ concentration outside the cell, or make the membrane potential more negative. If you increase the K+ concentration outside, the concentration gradient will push the K+ into the cell. If you make the membrane potential more negative, the cell will need to become more positive to reach its equilibrium potential, so K+ will flow into the cell and make it more positive*
3	**3.1)** Suggest increasing the electrical gradient will attract K+ into the cell (e.g., make membrane potential/cell interior more negative, such as -70 mV; make the cell exterior more positive). May also suggest reversing the concentration gradients (similar to 1.1), employing active transport (2.1), AND/OR changing the concentration gradient to make EK+ more positive.	*-decrease the concentration of K+ inside the cell to be below the outside - increase the concentration of K+ outside the cell to be above the inside - decrease the membrane potential of the cell (make it more negative) -add K+ pumps to the membrane / Changing concentration will alter the concentration gradient, therefore shifting the direction of the movement of K+ into the cell - Making the membrane potential more negative will increase attraction of positively charged ions to the inside of the cell*
	3.2) Suggest reasoning with electrical and concentration gradients but makes mistakes (i.e. concentration is stronger than electrical or they both can overpower each other)	*Changing the amount of K+ inside and outside the cell. Because if we make it so that the concentration gradient is stronger than the electric force, the net movement can change and cause the K+ to go into cell.*
2	**2.1)** Suggest using active transport/ATP/pumps to move K+ into the cell against the concentration gradient. May also suggest reversing the concentration gradient as in 1.1 (no mention of electrical ideas) OR opening inward rectifying channels.	*Higher concentration of K+ outside of the cell Lower concentration of K+ inside of the cell Pump K+ into the cell using active transport For the K+ to move into the cell on its own, a concentration gradient is needed in which the concentration of K+ outside the cell is greater than K+ inside the cell. The only way to avoid this is active transport.*
	2.2) Suggest changing the electrical gradient in an unspecified (e.g., 'change' the membrane potential) or incorrect (e.g., make membrane potential more positive) way to move K+ into the cell.	*Change the membrane potential or change the concentration of the K+ ions /if you change the membrane potential it would allow the ions to enter, same with if you change the conc. of ions*
	2.3) Suggest reversing the concentration gradient (e.g., increasing the K+ concentration outside of the cell, decreasing the K+ concentrations inside of the cell) because ions move from high to low concentrations or to reach equilibrium	*Increase the concentration of K+ outside to be greater than the one inside. / Diffusion goes from concentrations of high to low so it would move from the outside to the inside.*
1	**1.1)** Make a general statement about ions moving into or out of the cell, only suggest manipulating channels (incorrectly), explains an irrelevant process (e.g., AP, voltage gated channels), OR make a vague statement about the system.	*A dysfunction of the membrane channel. With a dysfunction, the channel might not permit the regular flow of potassium ions and this would change the membrane potential.*

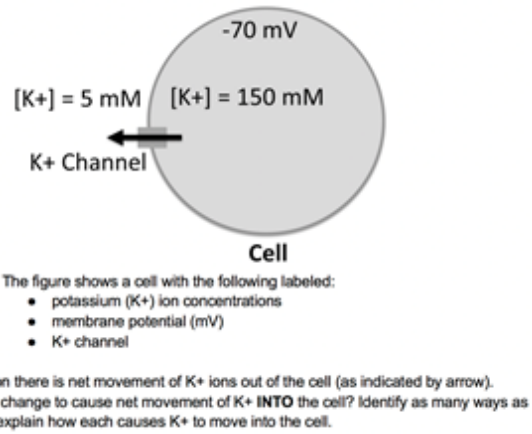

Figure 3. Constructed response item 'EION.'

nominal within one level; thus, indicating the different pathways students can reason within a holistic level.

The holistic rubric bin language had combinations of specific conceptual components that students used in explaining their reasoning across different LP levels. Some conceptual components could be determined by presence/absence dichotomous scoring, instead of multi-level holistic, quality scoring. This suggested that the rubric definitions might be suitable for deconstruction. For instance, we compared the following two codes described by the holistic rubric. Color has been added to highlight concepts shared between these codes (Figure 4).

Code	Rubric Description
5.1	*Explain that having a membrane potential below the equilibrium potential will make the electrical gradient stronger than concentration gradient and cause net movement of K+ into the cell. Suggest doing this by making the membrane potential more negative than the equilibrium potential/Ek/-90 or decreasing the concentration gradient to make the equilibrium potential more positive than resting/-70 mV.*
4.1	*Suggest increasing the electrical gradient (i.e., making the membrane potential more negative) or decreasing the concentration gradient in order to make electrical forces stronger than concentration forces and cause net movement of K+ into the cell.*

Figure 4. Example of student responses using combinations of specific conceptual components that suggested suitability for deconstruction.

To see the extent of language overlap in the holistic rubric, we proceeded to Cycle 2 of the deconstruction framework (Figure 2) outlined below. Each 'round' of deconstruction represents one turn within the gray box of Figure 2.

The Deconstruction of the Holistic Rubric

A principle of the framework that emerged is that there are multiple ways for analytic components to combine to give the same holistic code. Deconstruction of EION (Table 3) resulted in 16 individual, conceptual components within the 9-level holistic rubric. Conceptual pieces derived from the holistic rubric are shown as column headings. Rows represent different LP levels, and each row shows one way the conceptual pieces can be combined to show the presence of that LP level. Having a 1 in a cell represents that a conceptual piece is required for that LP level. For many LP levels (e.g., 5.1), there are multiple combinations of concepts that result in the same LP level and these responses could include lower level LP reasoning. In addition, conceptual pieces can occur at multiple LP levels. Thus, a concept such as 'make membrane potential more negative' can be part of either 3.1, 4.1, 4.2, or 5.1 codes, but is not required for codes 4.1, 4.2, or 5.1, because students could also use other concepts to be coded at the same level. Examples are shown in Figure 5.

Code	Rubric Description
4.1	*"1. Make K+ concentration outside bigger than that of the inside. This will flip the concentration gradient so that the K+ flows inside and the electrical gradient will cause K+ to flow inside. 2. Make membrane potential much more negative. This will cause the electrical gradient to be bigger than the concentration gradient, so K+ flows inside."*
3.1	*"Higher concentration of K+ outside the cell. Make the membrane potential more negative. Molecules move from high concentration to low so K+ would move into the cell. Opposite charges attract so a more negative charge would pull K+ ions back into cell."*

Figure 5. Example of student responses using conceptual pieces that occur at different LP levels.

Table 3. Original deconstruction matrix of the EION assessment example.

Holistic Code	MP < EK	Make electrical gradient stronger than concen-tration gradient	Treats electrical and chemical gradients independent	Reason with both electrical and chemical gradients but make mistakes	Make cell interior more negative	Make cell exterior more positive	De-crease MP	De-creasing K+ inside the cell	In-creasing K+ outside the cell	Decrease concen-tration gradient to make EK+ more positive	Change MP	MP more positive	Ions move from high to low concen-trations	Ions move to reach equilibrium	Active transport/ ATP/ Pumps	Open inward rectifying channels
5.1	1	1								1						
5.1	1	1					1									
5.1	1	1				1										
5.1	1	1			1											
4.2	1		1				1									
4.2	1		1			1										
4.2	1		1		1											
4.1		1								1						
4.1		1					1									
4.1		1				1										
4.1		1			1											
3.2				1												
3.1			1				1									
3.1			1			1										
3.1			1		1											
3.1			1							1						
2.3								1								
2.3									1							
2.2											1					
2.2										1						
2.1																1
2.1															1	
1.1																

Round 1

Another principle that emerged is that expert reasoning codes do not have to have lower level reasoning when deconstructing. Therefore, from only using expert discussion for component refinement, the analytic rubric was reduced from 16 to 8 analytical bins (Table 4). The first change reflected removing the 'ions move to reach equilibrium' and 'ions move from high to low concentrations' bins, because all holistic levels could include these concepts. In other words, both these bins were '*may have*' in Boolean logic in all sublevels of the holistic rubric. These ideas are considered additional reasoning students use to support their complex thinking, instead of the distinct reasoning we are interested in capturing for the LP. The bin 'treats electrical and chemical gradients independently' was also removed because of the human difficulty in defining and identifying this concept. The 'mistakes' bin, or where students make mistakes in their content understanding, was also vague in rubric definition but important in where students were placed holistically, so this bin remained as part of the rubric.

After removal of bins, we next attempted to identify truly distinguishing features to reduce columns for more manageable coding by experts. For example, 'open inward rectifying channels' and 'active transport/ATP/Pumps' were originally separate concepts, but were features found only in holistic code 2.1. These concepts were combined with an '*or*' statement into one analytical bin, because no matter which concept the student used, they would always be coded as a 2.1. Other bins that were also combined included: 'change MP' *or* 'make MP more positive'; 'increasing K+ outside cell' *or* 'decreasing K+ inside cell'; and 'cell interior more negative' or 'cell exterior more positive' or 'decrease MP. This combination of concepts allowed coders to key into different language in student reasoning patterns that were related on a conceptual level. Combining similar conceptual pieces seemed to help how coders approached and applied analytic coding. The fewer columns used to code student responses are also ideal to reduce computer model computational resources and time to build machine learning models.

Table 4. First round of refinement for the deconstruction matrix of the EION assessment example.

Holistic Code	MP < Ek	Make electrical gradient stronger than concentration gradient	MISTAKES	Cell interior more negative **OR** cell exterior more positive **OR** decrease MP	Decrease concentration gradient to make EK+ more positive	Increasing K+ outside the cell **OR** decreasing K+ inside cell	"change" MP **OR** make MP more positive	Open inward rectifying channels **OR** active transport/ ATP/Pumps
5.1	1	1			1			
5.1	1	1		1				
4.2	1			1				
4.1		1		1				
4.1		1			1			
3.2		1	1	1				
3.2		1	1		1			
3.2	1		1	1				
3.1				1				
3.1					1			
2.3						1		
2.2							1	
2.1								1
1.1								

Besides reduction of the number of analytic rubric bins from 16 to 8, this iterative process also reduced the number of rows, or possible combinations for Boolean operators from 23 to 14. However, there were still some overlapping concepts such 'decrease membrane potential' *or* 'make cell exterior more positive' or 'make cell interior more negative' in codes 3.1, 4.1, 4.2, or 5.1, so the experts considered if these bins were necessary to code for each level. We used 100 student responses (partitioned into the different levels of the original holistic code and randomly selected) coded by two experts for validation. Cohen's Kappa between two experts across the eight analytic bins ranged from 0.653 – 0.890 with 3 of the 8 bins below 0.7. When analytic codes were combined with Boolean logic to determine the holistic code, the Cohen's Kappa was 0.683 for the two calculated holistic scores. We also compared the original holistic codes, or codes that were used to holistically categorize before rubric deconstruction, with each expert's calculated holistic code via Boolean logic. Expert 1 had a Cohen's Kappa of 0.593 with the original holistic codes and expert 2 had a Cohen's Kappa of 0.638. With this validation effort, we began another round of refinement to uncover these discrepancies.

Round 2

Another principle of the framework that emerged is that clarification of vague bins is essential for improved reliability. This round resulted in the net addition of a single bin to clarify some of the individual bins that performed poorly during validation (Table 5). To improve clarification, the bin 'mistakes' was replaced with two specific reasoning patterns that students often used in specific mistakes for EION: 'make membrane potential greater than equilibrium potential' and 'make membrane potential positive.' The decision of a 'mistake' for these reasonings (i.e., lowering of a level) were based on the *flux* LP framework. Because the LP framework seeks to capture common ways students reason about flux, some of these levels include common errors which are present in lower levels of reasoning. The LP provided a consistent way to address these errors across various flux contexts, but fine-grained coding components were dependent on mistakes elicited by the specific item. Thus, each rubric's deconstruction was referenced to the original LP framework for common mistakes. After discussion, the overlapping bin of 'make membrane potential more negative' outlined above in the original deconstruction matrix (section 3.4.1.1) between 3.1, 4.1, 4.2, and 5.1 was found to be the only distinguishing features for level 3.1. Higher holistic levels

Table 5. Second round of refinement for the deconstruction matrix of the EION assessment example.

Holistic Code	MP < EK	Make electrical gradient stronger than concentration gradient	Make MP more positive **OR** MP> EK	Compares/ contrasts electrical and concentration gradients	Cell interior more negative **OR** cell exterior more positive **OR** decrease MP	Decrease concentration gradient to make EK+ more positive	Increasing K+ outside the cell **OR** decreasing K+ inside cell	"change" MP	Open inward rectifying channels **OR** active transport/ ATP/Pumps
5.1	1	1							
4.2	1								
4.1		1							
3.2		1	1						
3.2				1					
3.1					1				
3.1						1			
2.1							1		
2.1								1	
2.1									1
1.1									

(4.1, 4.2, and 5.1) could always include lower level reasoning in their answer, but this bin was not essential for students to use in their reasoning to be assigned codes 4.1, 4.2 or 5.1.

Another change included the level of exclusivity of concepts in codes 2.1, 2.2, and 2.3. Some student responses were difficult to place into only one holistic sub-level code because these concepts were not mutually exclusive – as the holistic rubric first suggested. These concepts remained as three different analytical bins, but the holistic codes and combinations were changed to reflect that only one of these concepts were needed to be coded a 2.1. This refinement reduced the matrix from 14 rows to 11 rows. This refinement also allowed us to reduce the holistic rubric from nine to seven levels. Some Boolean logic statements were also simplified by removing some of the additive components. Some higher levels still had overlapping concepts, but were more manageable in a defined rubric matrix of 9 x 11 (Table 5) rather than the original version 16 x 23 (Table 3).

An additional 50 student responses were coded using the analytic bins for validation by two experts during the second round. Cohen's Kappas between coders ranged from 0.650 - 1.00 with only 1 of the 9 bins below 0.7. The lowest Cohen's Kappa bin, 'decrease concentration gradient to make EK+ more positive,' did not change in Cohen's Kappa from the first rubric version (0.653) to the second (0.650). While the experts agreed that this concept was important to capture, this concept was rare. Cohen's Kappa between two experts' calculated for 150 holistic codes, combined from the analytical bins using Boolean logic, was 0.873. With almost all analytical bins performing well and holistic codes having a high degree of reliability, the experts agreed that the rubrics were ready for use to code all the responses.

Summary

The initial deconstructed rubric from the holistic rubric (Table 2) was represented by a matrix containing 16 integrating concepts, with 23 possible combinations to place a student's response into one of 9 holistic codes. In the last round of refinement for the deconstructed rubric (Table 5), the rubric only contains 9 integrating concepts, each of which can be coded independently. The rubric contains 11 possible code combinations, to generate 7 unique holistic codes; however, this can be done automatically via computation after coding is complete. Through iterative rounds of re-evaluation, revising, and validation scoring, the deconstruction process used data-driven by IRR which measured improvement of development progress of a rubric aligned to a LP.

Lessons Learned and Future Directions

Lessons Learned

One challenge for our trained holistic coders was that they found an affordance of holistic rubrics was a 'two-tier' coding process when coding with holistic rubrics which was absent when using analytic rubrics. Specifically, coders reported that when using a holistic rubric, they evaluated student responses both at the indicator sublevel as well as overarching LP level descriptions. This two-tiered approach was particularly useful when coding vague student responses because coders could draw on both patterns to determine a relevant code. When using an analytic rubric, coders said they only relied on the fine-grained approach without the larger context of student reasoning that the holistic level description provided. This made coding responses that did not clearly align with analytic rubric bins challenging to score.

Some have suggested that deconstruction might: 1) not capture the breadth of student reasoning; 2) lead to a loss of original concepts; 3) oversimplify a concept or have a loss of complexity; and 4) not be beneficial, if the separate dimension codes are only summarized in the end. Reflecting on our deconstruction process above, we identified some key lessons learned to address these concerns for future research or application:

Two experts are needed for deconstruction and round refinement. At minimum, two researchers with disciplinary knowledge relevant to the question being investigated should each complete deconstruction in order to determine what emphasis is necessary for concept coding, or criteria/boundaries for coding. If the two experts disagree on the needed emphasis, then a third expert should be available to help make the final decision. An example from EION would be that a content expert would be able to know that a bin 'make MP more negative than equilibrium potential' is one, complete idea and could not be further deconstructed into anything more fine-grained.

Code a subset of student responses for reliability and validity during each round. Reliability and validity are important for both defining the concepts/bins with their coding rules and associated Boolean logic. The refinement process is time consuming, but this careful process of focusing on only the distinguishing features between holistic levels reduces complexity in the Boolean logic, all possible combinations, and increases the clarity of concepts.

Create a feedback loop from Cycle 2 to Cycle 1. Information from the discussions during the deconstruction process and application of the deconstructed rubric should be used to revise and improve the original holistic rubric.

Expert reasoning can include novice reasoning. Expert reasoning includes multiple components, some of which are present in novice answers, but experts also include other concepts not present in novice answers. If coding holistically, these are integrated concepts but if coding analytically, each concept in all the analytic bins need to be accounted for, not just the highest order, or most expert-like, analytic bin. Thus, the analytic bins may provide additional details about the developmental pathways that the LP is attempting to capture.

Code concepts novice to expert, but use Boolean Logic expert to novice. Application of Boolean logic should begin with the expert levels, or higher-order LP levels that require more combinations of bins for elimination, before applying Boolean logic to lower-level or single-bin logic. The more complex the holistic concept being measured, the more complex the Boolean logic coding would be, and extreme care should be taken to make sure that all possible outcomes of coding are placed into the correct holistic score. Also, putting lower level reasoning, or bins that appear more frequently in student responses, first in the coding sheet structure or code book reduces the probability that they will be overlooked – this is advantageous to coders if they are accustomed to holistic coding.

Watch out for low frequency concepts. It is important to recognize low number of positives in a single concept bin – these concepts might be interesting to capture for research, but might not be reliably identified by coders since they would few positive cases or exemplars.

Future Directions

We will perform a cost benefit analysis of key factors such as time, effort, and reliability measures to compare holistic and analytic approaches to determine if the benefits of deconstruction outweigh the costs of time and effort for using either rubric. Our goal is for deconstruction to be used to improve the development of machine-learning applications, in order to allow

more student responses to be qualitatively coded. We will also investigate the combination of holistic and analytic coding approaches to improve computer model predictions of the LP level. Additionally, one could explore the possibility of differential weighting of analytic components within the Boolean logic statements. This has the potential to allow large-scale statistical approaches and generalizability studies to be done on open-ended student responses.

Conclusions

This paper describes a systematic approach of rubric deconstruction that started with a holistic rubric that was then deconstructed into a set of analytic rubrics to clarify rubric criteria, ideas and language. Using this process, we believe that it is possible to improve the validity of rubrics while increasing reliability, but with caution. Deconstruction is a long, and tedious task. Throughout the process, those completing and refining the deconstruction should be constantly questioning the purpose of the coding and the defining features for each criterion – whether it is an analytic or holistic approach.

Deconstructing a holistic rubric is one approach to unpack the challenges of the heterogeneity and complexity in student writing while preserving the benefits of both holistic and analytic rubric properties. The deconstruction example used above in formative assessments showed rubric improvement by clarifying and removing concept descriptions to uncover complex physiology-content student reasoning. The methods outlined in this study were targeted at an LP framework used in undergraduate physiology; however, these methods might be applicable to other assessment situations, such as licensing and certification exams, that rely on constructed response or even performance assessments to unpack complexity in student writing. These methods could help to improve large-scale assessment development processes targeted at uncovering complex scientific reasoning across domains.

References

Allen, D., Tanner K. (2006). Rubrics: Tools for Making Learning Goals and Evaluation Criteria Explicit for Both Teachers and Learners. CBE- Life Sciences Education, 5, 197-203. DOI:10.1187/cbe.06-06-0168

AERA, APA & NCME. (2014). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

Anderson, CW. (2008). Conceptual and Empirical Validation of Learning Progressions. Presented at the Meeting on Advancing Research on Adaptive Instruction And Formative Assessment, sponsored by the Center on Continuous Instructional Improvement (CCII). Philadelphia, PA.

Corbin, J, Strauss, A. (2008). Basics of Qualitative Research (3rd ed.): Techniques and Procedures for Developing Grounded Theory. Thousand Oaks, CA: Sage Publications.

Doherty JH, Scott EE, Cerchiara JA, McFarland, Wenderoth MP. (2019). A Learning Progression Characterizing How Students in Biology Understand Ion Movement. Paper presented at the Annual International Meeting of the National Association for Research in Science Teaching (NARST). Baltimore, MD Mar 31-Apr 3

Haudek, KC, Moscarella, RA, Weston M, Merrill J, Urban-Lurain M. (2015). Construction of Rubrics to Evaluate Content in Students' Scientific Explanation Using Computerized Text Analysis. National Association for Research in Science Teaching (NARST), Conference Proceedings.

Humphry, SM, Heldsinger, SA. (2014). Common Structural Design Features of Rubrics May Represent a Threat to Validity. *Educational Researcher*, 43(5), 253-263.

Hunter, DM, Jones, RM, Randhawa, BS. 1996. "The Use of Holistic Versus Analytic Scoring for Large-Scale Assessment of Writing." *The Canadian Journal of Program Evaluation*. 11(2): 61-85.

Jonsson, A, Svingby, G. (2007). The Use of Scoring Rubrics: Reliability, Validity and Educational Consequences. *Educational Research Review*, 2, 130-144.

Liu, OL, Brew, C, Blackmore, J, Gerard, L, Madhok, J, Linn, M. (2014). Automated Scoring of Constructed-response Science Items: Prospects and Obstacles. *Educational Measurement: Issues and Practices*, 33(2), 19-28.

Michael, J, and McFarland, J. (2011). The Core Principles ("Big Ideas") of Physiology: Results of Faculty Surveys. *Advancement in Physiology Education*. 35, 336-341.

Mohan, L., Chen, J., and Anderson, C.W. (2009). Developing a Multi-year Learning Progression for Carbon Cycling in Socio-ecological Systems. *Journal of Research and Science Teaching* 46, 675–698.

Moskal, BM. (2000). Scoring Rubrics: What, When, and How? *Practical Assessment, Research & Evaluation*. 7(3).

Moskal, BM, Leydens, JA. (2000). Scoring Rubric Development: Validity and Reliability. *Practical Assessment, Research & Evaluation*. 7(10). National

Research Council (NRC). (2001). Knowing What Students Know: The Science and Design of Educational Assessment. Washington, DC: The National Academies Press.

Panadero, E, Jonsson, E. (2013). The Use of Scoring Rubrics for Formative Assessment Purposes Revisited: A Review. *Educational Research Review*, 9, 129-144.

Saldana, J. (2009). An Introduction to Codes and Coding: The Coding Manual for Qualitative Researchers. Los Angeles, CA: Sage Publishing.

Singer, NR, LeMahieu, P. (2011). The Effect of Scoring Order on the Independence of Holistic and Analytic Scores. *The Journal of Writing Assessment*. 4(1). http://journalofwritingassessment.org/article.php?article=51

Urban-Lurain, M., & Weinshank, D. J. (1999). 'I Do and I Understand:' Mastery Model Learning for a Large Non-major Course. *Special Interest Group on Computer Science Education*, 30, 150-154.

Waltman, K, Kahn, A, Koency, G. (1998). Alternative approaches to scoring: The effects of using different scoring methods on the validity of scores from a performance assessment. CSE Technical Report 488. Los Angeles.

Acknowledgment:

We thank the members of the Automated Analysis of Constructed Response research group, especially Matthew Steele, for their thoughtful comments regarding challenges we encountered in this project. This work was supported by the National Science Foundation under Grant DUE 1660643 and 1661263.

Citation:

Jescovitch, Lauren N., Scott, Emily E., Cerchiara, Jack A., Doherty, Jennifer H., Wenderoth, Mary Pat, Merrill, John E., Urban-Lurain, Mark, Haudek, Kevin C. (2019). Deconstruction of Holistic Rubrics into Analytic Rubrics for Large-Scale Assessments of Students' Reasoning of Complex Science Concepts. *Practical Assessment, Research & Evaluation*, 24(7). Available online: http://pareonline.net/getvn.asp?v=24&n=7

Corresponding Author

Lauren N. Jescovitch
CREATE for STEM Institute
Michigan State University
East Lansing, MI

email: jescovit [at] msu.edu