Challenges in Developing Computerized Scoring Models for Principle-Based Reasoning in a Physiology Context

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

1.0 Introduction

Principled-based reasoning (PBR) is a central practice which experts use to define relationships that explain scientific phenomena (Dunbar, 2000; NRC, 2000), or build a robust mental model of the mechanisms underlying a system (Knapp and D'Avanzo, 2010; Modell, 2000). Foundational ideas, or core concepts, are found in all science disciplines, but principles include those ideas which underlie and connect disciplines. Principles apply to multiple contexts and content areas to guide organization and learning across content areas (Parker et al, 2012). Teaching with principles reduces the common student approach of learning as memorization of processes and leads students to a deeper understanding of complex interactions (Knapp and D'Avanzo, 2010; Parker et al, 2012). Examples of principles include those such as conservation of matter and energy, hierarchical nature of biological systems, and *flux* and may be represented by an equation or relationship. Hartley et al (2011) suggests that students' misconceptions about biological processes are related to their failure to understand fundamental principles. Instructors assume students understand and are able to use PBR even when they do not; thus, assessments could provide more insight into student thinking of principles (Hartley et al., 2011).

Recent calls for reform in undergraduate biology education have focused on key ideas in the discipline, which have broad applicability across specific discipline contexts (AAAS, 2011). In the sub-discipline of physiology, seven fundamental ideas have been identified as important in undergraduate education. One of the most applicable to physiology, and more broadly as a principle, is *flux* (Michael et al., 2017). *Flux* describes the passive flow of substances and heat down gradients and can be described as proportional to the gradient over the resistance. In differing contexts, *Flux* can appear as Ohm's Law, Fick's Law of Diffusion, or Poiseuille's Law. Many students use these equations, but it is difficult to know if students understand the underlying principle that crosses disciplines. A Learning Progression (LP) framework is one way we can categorize complex PBR about different student ideas about *flux*.

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 highest level, or upper anchor, represents the target for expert-level reasoning (e.g., graduating seniors). The lowest level, or lower anchor, represents novice students who are entering their introductory courses (e.g., entering freshmen). Intermediate levels of the LP describe multiple, possible pathways and strategies for students to progress to the upper level. A LP categorizes patterns across large groups of students (Smith et al., 2006); although, not every student will progress through each level to reach the upper level. To understand undergraduate student reasoning with and about *flux*, we developed constructed response (CR) questions in a physiology context aligned with a developing Learning Progression (LP).

Assessment of complex constructs such as PBR, provide rich insight into student reasoning and explanatory models. 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 – particularly in large classrooms – and elicit students' higher order thinking (Allen and Tanner, 2006; Jonsson and Svingby, 2007; Montgomery, 2002). The most common assessment format for a large enrollment, introductory STEM (Science, Education, Engineering, Mathematics) higher education course is multiple-choice (fixed response). Multiple-choice questions are fast, simple, and easy to evaluate; however, these assessments inform the instructor very little, if at all, about the heterogeneity of students' thinking. More complex ' scientific practices such as argumentation, explanation, and integration of core ideas may be difficult to measure in a multiple-choice assessment (Allen and Tanner, 2006; Nehm, Ha, and Mayfield, 2012). Thus, developing constructed response assessments in a classroom context can provide helpful information to help instructors to make educational decisions for student learning; but, CR answers can be difficult to interpret and refine the feedback for instructor and students. While CR items are more time consuming to evaluate than multiple choice items, especially at large scales, recent efforts have advanced computerized-scoring approaches which identify key disciplinary concepts and attempt to predict expert classifications as part of evaluation (Liu et al., 2014; Urban-Lurain et al., 2015).

Automated analysis using computerized-scoring models can be used to aid in the evaluation of scientific concepts in students' written responses in large, undergraduate STEM courses (Haudek et al., 2015; Urban-Lurain et al., 2015). One approach uses machine learning (ML) algorithms, which have the ability to predict classifications of CR at performance levels that have inter-rater reliability (IRR) measures equal to those between two calibrated content experts (Nehm, Ha, & Mayfield, 2012). An expert-coded data set can be used as a training set to build a predictive scoring model; once the model is sufficiently trained or meets acceptable performance benchmarks, it can then be used to predict codes for new data. Developing the scoring model using a training set of data, requires an iterative process including development, use, and refinement of an expert validated scoring rubric and expert scores (or codes) (Nehm et al., 2010). Then using the predictive models, instructors can submit data collected from a developed CR item and generate near "real-time" formative results of their students' thinking about the targeted construct. Thus, these models applied to new data sets can be used to generalize student thinking or ability to inform instruction and learning in the classroom. However, the process of creating a finalized, predictive model is iterative and tedious, and faces challenges that could impede model development. For example, Liu et al. (2014) revealed challenges for automated scoring of science explanations that included small sample sizes, pronoun resolution (or multiple uses of "it" referring to different objects), computer difficulty in identifying incomplete or inaccurate explanations, and difficulty in the computer distinguishing non-normative ideas from normative. Challenges for a computerized scoring model are dependent on variables, such as the practice being assessed, coding rubric, and sample of collected data, so not all challenges in model development can be anticipated. However, lessons learned from these challenges may inform item and model development during successive iterative rounds or prospective new constructs and approaches.

Automated analysis of constructed responses can assess student thinking but relies on a set of human coded data based on one or more evaluation rubrics. LPs by their nature, attempt to classify student performances into a single level. LPs may be operationalized by creating an aligned holistic coding rubric for CR assessment items; these rubrics assign responses to a single, mutually exclusive bin based on the observed performance. Another approach to coding is analytic. We define an analytic rubric as evaluating material by the use of multiple, independent coding bins, which are not mutually exclusive. Although grain-size for rubric bins can differ based on research question and approach, generally analytic bins tend to be used to identify a singular idea or concept. Complexity may be inferred from the combination of these rubrics. The purpose of this study is to build an efficient computer scoring model using information from a holistic rubric that captures a complex construct (PBR) in a physiology context. In other preliminary work attempting to build computerized scoring models, we found that an analytic coding approach reduces coding complexity, which may improve the process of model development. Thus, in this study, we continue to investigate if we can align these rubric approaches to develop an analytic rubric derived from a holistic rubric, with the goal of developing a successful computerized scoring model to place student reasoning on a *flux* LP.

As stated above, expert answers in STEM are generally complex in nature and typically include multiple, interacting core concepts and principles (NRC, 2000). Thus, it is not surprising that holistic coding to identify high levels of sophisticated reasoning may contain complex, additive or overlapping language in rubric descriptions. In addition, distinguishing between holistic levels may require deep understanding of the content area and use nuanced, discipline specific ideas. One way to find the distinguishing features of expert-level reasoning is through the process of deconstruction. We define deconstruction as the process of breaking apart the levels derived from a holistic rubric into defining individual, conceptual components which can be used for analytical rubric development and application. These individual components may then be recombined into a single holistic score – which keeps true to the purpose and intention of the holistic rubric and which represents the unique complexity of the components.

This deconstruction process is inherently difficult because of the requirement for a thorough familiarity of both the range of possibilities and the elements comprising an expert answer in student thinking (Allen and Tanner, 2006). There are multiple attributes to why rubric developers may want to deconstruct a rubric. The deconstruction process can help to: 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, or those that aren't important and can be removed; 4) display data in multiple ways in order to communicate feedback for researchers, students, or instructors; 5) solidify organization of concepts into holistic schema where originally these might be vague; 6) allow for expansion of both, quantitative and qualitative, interpretation of results. The deconstruction process can leverage the benefits of both holistic and analytic rubrics, so both of sets of results can be used together to interpret and present 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. The process or resulting rubric: 1) might 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; 4) may require a prohibitive amount of extra time and effort, which is expensive, even though a high number of bins could provide insight into student conceptual difficulties and provide a framework to design interventions; 5) is not often preferable if only the separate dimensions of the scores are summarized in the end without communicating more fine-grained findings of these dimensions (Waltman, Kahn, and Koency, 1998). Thus, it is important to keep the richness of the coding process dependent on the audience and/or end-user of the resulting information. Moskal (2000) states that you can build holistic values based on analytical criteria and warned about weighting specific criteria over others, which might not be what was intended in the original coding scheme. Another warning from Stemler (2001) highlights flaws in content analysis, or systematic technique to compress words of texts into categories based on rules of coding, if there are faulty definitions of criteria and/or non-mutually exclusive and exhaustive categories.

This study will focus on the challenges associated with developing high performing computerized scoring models to identify and classify the use of principled based reasoning in a physiology context from a holistic rubric and the relative trade-offs attempting to represent a holistic rubric as a set of analytic components.

2.0 Study Design

A series of CR items were developed to assess one progress variable within a *flux* learning progression (LP) framework that captures PBR (Doherty et al., 2019). Items were developed as part of an iterative routine between LP development, assessment item and rubric alignment to LP, human coding, and development of computerized scoring models. The *flux* assessment question used as the example for this report is named "EION" (Figure 1). EION assesses undergraduate students reasoning about *flux* using both concentration gradients and electrical gradients.

The EION assessment states:

> *"The figure shows a cell with the following labeled: potassium (K+) ion concentrations, membrane potential (mV), K+ channel. In this situation there is net movement of K+ ions out of the cell (as indicated by arrow). What can we change to cause net movement of K+ into the cell? Identify as many ways as you can. Explain how each causes K+ to move into the cell."*

EION was administered to undergraduate students taking physiology and biology courses at two community colleges and eight universities in the USA. Responses were scored independently by two experts using a five level (and further divided into nine sublevels based on LP indicators) holistic rubric aligned with the *flux* LP, then subjected to confirmatory analysis (CA), as part of the process to develop a computerized predictive scoring model (Urban-Lurain et al, 2015). The developed holistic rubric, sublevel indicators and exemplar student responses for each sublevel are presented in Table 1.

Table 1. Holistic rubric of the example assessment item 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 pottasium ions and this would change the membrane potenetial.*

2.1 Building Computer Scoring Models

The question development cycle (QDC) is the methodological process which captures our group's process for developing, validating, and implementing assessment items for automated analysis. Items are referred to as all the components (e.g., rubric, coding, confirmatory analysis) that makes up the application process for EION. In the first step of the QDC, we create a *New Question Design* to measure student thinking about important disciplinary constructs (e.g., EION). We then administered this question to students via online course management systems (e.g., Desire2Learn, Blackboard, Canvas, Moodle) for *Data Collection*. *Exploratory Analysis* uses lexical analysis software(s) or expert pattern finding to extract key terms and scientific concepts from the students' writing that is aligned to the co-development of the *flux* LP. These terms and concepts aid in *Rubric Development that* is also aligned to the *flux* LP. *Human Coding* of student responses used a holistic rubric. The human coding is then used as a training set for *Confirmatory Analysis* as dependent variables in statistical classification techniques to predict expert human coding of student responses. The final product of the QDC is the *Predictive Model* that can be used to completely automate the scoring of a new set of student responses, predicting how experts would score the responses. However, if a predictive model is not successful, *Question Revision* or previous components of the QDC can be changed to enhance a model's performance. Figure 2 shows the order in which the QDC process occurs. This project will focus on the challenges encountered during the Confirmatory Analysis phase of the QDC in attempting to build scoring models to identify PBR about *flux* in student written responses.

2.1.1 Confirmatory Analysis

The main objective of Confirmatory Analysis (CA) is to train a computer model that validates the previous QDC processes. CA uses the expert holistic scoring codes to train a predictive scoring model, and also measures performance of this scoring model. CA is a machine learning system that uses a set of predetermined algorithms to return a value given a set of inputs (student responses). The "learning" refers to how the machine "learns" how to optimally mimic the hand-scoring of experts. In order to learn, CA uses lexical resources, or extracted features from text, as dependent variables.

The Automated Analysis of Constructed Response (AACR; https://create4stem.msu.edu/project/aacr) computerized scoring system treats the task of assigning scores to student writing as a machine learning text classification problem (Aggarwal & Zhai, 2012). During CA, the computerized scoring system generates LP predictions on a given, human coded training set. To generate these predictions, we use an ensemble of 8 individual machine learning algorithms (Jurka et al. 2012). Our ensemble is a group of algorithms that vote independently on the categorization of a particular document, but whose individual votes are combined to make a final categorization. Using an ensemble of algorithms increases the robustness, or time and complexity of data patterns, for better classification accuracy over large sets of data than the use of fewer classification algorithms (Collingwood and Wilkerson, 2012). Each individual algorithm is trained using the EION expert-scored student responses, resulting in computer LP predictions of the human LP codes.

2.1.2 Evaluating model performance

We used Cohen's Kappa, a confusion matrix, and various other statistical measures to evaluate the performance of the computer scoring model. Cohen's Kappa and the expert-computer confusion matrix are used as initial insight into the efficacy of the ensemble and will be our focus in our developing model. Cohen's Kappa is a statistical measure commonly used to quantify human-computer IRR (Cohen, 1960). Cohen's Kappa determines whether the degree of agreement between two raters is higher than what would be expected by chance. The power of Cohen's Kappa is that it compares the absolute agreement relative to the agreement expected by chance. Cohen's Kappa ranges from -1.0 to 1.0. In general, when Cohen's Kappa is greater than 0.8, the ensemble is said to have excellent agreement with the expert scores. If Cohen's Kappa is between 0.6 and 0.8, the ensemble is said to have moderate agreement with expert scores. Values of Cohen's Kappa lower than 0.6 generally indicate poor agreement between the ensemble and expert. We use these values as a benchmark for determining if the computer model is successful.

A confusion matrix evaluates model performance and shows the reference (expert scores) by columns and the prediction (computer predicted scores) by rows. Ideally, 100% agreement between human and computer codes would be found on the diagonal, where all responses scored 1 were present in 1X1, and so on with 2X2, 3X3 and 4X4. However, responses which are predicted to have a different code by the computer than that assigned by experts are called discrepancies. Analysis of discrepancies gives information as to what feature(s) the computer may be clueing into that was different than the experts. Further investigations into discrepancies for model improvement might include collection of more data (low response rates), pattern analysis of discrepancies, etc.

3.0 Results and Discussion

There are a variety of challenges that arise when attempting to improve computer scoring model performance (e.g., Liu et al., 2014; Liu et al., 2016). These challenges are not standardized because model performance depends on the construct being assessed, complex interactions of the steps in developing a computerized predictive model and usually involve iterative rounds for model improvement. Below are challenge-trends that we have faced while optimizing the EION computer scoring model, which exemplifies a complex construct embedded in a discipline context.

The outcome of the first attempt at developing a machine scoring model are in the confusion matrix in Table 2. The confusion matrix shows the reference (expert scores) in columns by the machine learning prediction (computer predicted scores) in rows. Ideally, 100% agreement would be a diagonal; however, we can see that the computer had considerable misprediction in levels 3-5, which correspond to increasingly sophisticated PBR by students. Overall, this first model had an overall accuracy of 72% and a Cohen's kappa of 0.578.

Table 2. *Round 1: EION confusion matrix (Cohen's Kappa=0.578). Reference is the expert score; Prediction is the computer predicted score for a given response.*

	Reference								
Prediction	1.1	1.2	2.1	2.2	3.1	4.1	4.2	5.1	Sum
1.1	112	16	6	8	4	0	2	1	149
1.2	6	58	7	12	1	0	0	0	86
2.1	1	4	37	4	3	0	0	0	49
2.2	1	8	2	80	30	10	10	6	147
3.1	0	1	0	16	78	10	18	6	129
4.1	0	0	0	1	2	3	1	2	9
4.2	0	0	1	3	14	3	81	31	133
5.1	0	0	0	1	4	12	11	49	77
Sum	120	87	53	125	136	38	123	95	779

Iterative rounds of model improvement were conducted in order to improve human-machine agreement. Challenges faced throughout these rounds included: preprocessing replacement of keyword synonyms (addressed between rounds 1 and 2), low frequency of responses (2-3), languages used in student mistakes (3-4), and additive or overlapping language in rubric definitions (all).

3.1 Preprocessing Replacement of Keyword Synonyms

The EION item asks students to reason using concentration and electrical gradients; two key concepts when using this reasoning are equilibrium potential and membrane potential. In a physiological context, these concepts can be expressed in a number of different terms and symbols. This is problematic in that the usage of some symbols and terms is infrequent enough that there are not enough positive examples for the computer to "learn" that these terms/symbols are actually synonyms. Improvement in model performance, as measured by Cohen's kappa, (from 0.578 to 0.669) was made by preprocessing student responses so that all the synonyms for equilibrium potential and membrane potential were standardized to those features. For instance, the phrase *equilibrium potential* was used to replace the following: *E(k+), -90, -91, -92, Ek, Epotassium, E(potassium)*; and *membrane potential* replaced the following: *70, resting potential, MP;* round 2 includes these changes and the resulting confusion matrix is provided in Table 3. The standardization of these features allowed us to essentially tell the computer that they have the same meaning and also limits the diversity for multiple features with the same meaning.

Table 3. *Round 2: EION confusion matrix (Cohen's Kappa=0.669). Reference is the expert score; Prediction is the computer predicted score for a given response.*

	Reference								
Prediction	1.1	1.2	2.1	2.2	3.1	4.1	4.2	5.1	Sum
1.1	127	12	7	10	6	0	2	0	164
1.2	2	36	8	16	3	0	0	0	55
2.1	2	5	49	0	4	0	0	0	60
2.2	0	5	1	43	9	1	4	1	64
3.1	0	1	0	16	137	19	27	4	204
4.1	0	0	0	0	0	5	0	5	10
4.2	0	0	0	3	11	2	110	23	149
5.1	0	0	0	1	1	11	4	54	71
Sum	131	59	65	79	171	38	147	87	777

3.2 Number of Responses

If a feature or a pattern in the text responses is only infrequently exhibited (e.g. <5%) in the data used to train the model, then more than likely the computer model may not use such a feature or pattern as a key variable in the resulting classification algorithm. This leads to lower accuracy and the potential that a given rubric bin could not be accurately predicted. The confusion matrix of round two for EION shows that bin 4.1 only had 38 student responses scored by experts out of the entire dataset (N=777). Of these 38, the computer correctly predicted only 5 responses. Since positive case occurrence was so low and bin 4.1 represents a more sophisticated type of reasoning in the LP, we decided to target more responses which exhibit characteristics of the 4.1 bin by evaluating responses provided by more advanced students (i.e., college seniors). After coding 189 more responses for more upper bin responses, 175 of these responses were added to the data for a dataset of 952 responses. The additional 14 responses were coded with a new emergent code (indicator 3.2) which was added into the analysis during the next round discussed below (section 3.3). A total of 83 responses scored in bin 4.1 by experts. Unfortunately, only 5 of these 83 responses were still predicted to be in 4.1 by the computer; the other responses were predicted into bins 3.1, 4.2 and 5.1. This was most likely due to an overlap in bin definitions between these codes (see section 3.4 below). Thus, overall Kappa degraded from 0.669 to 0.552 (Table 4).

Table 4. *Round 3: EION confusion matrix (Cohen's Kappa=0.552). Reference is the expert score; Prediction is the computer predicted score for a given response.*

	Reference								
Prediction	1.2	2.1	2.2	2.3	3.1	4.1	4.2	5.1	Sum
1.2	31	9	6	0	2	1	4	2	55
2.1	9	49	0	2	5	0	1	0	66
2.2	6	1	46	0	12	5	6	2	78
2.3	10	6	8	129	8	0	3	0	164
3.1	3	0	19	0	147	23	44	15	251
4.1	0	0	1	0	4	5	5	14	29
4.2	1	0	3	0	21	16	126	31	198
5.1	0	0	1	0	9	33	8	60	111
Sum	60	65	84	131	208	83	197	124	952

3.3 Language for Mistakes

Some student responses contain mistakes in physiology content knowledge. The mistakes bin can be described as capturing students who articulate highly sophisticated reasoning, but also include some type of error, usually in content knowledge, that is important to capture in the coding. During our targeting for upper bin responses, we developed an extra indicator and bin for student "mistakes", labeled 3.2. This bin was described as students who were doing level 4.1 or 4.2 reasoning but made some type of error (e.g., "make membrane potential more positive than -91 mv" instead of "make membrane potential more negative than -91 mv"). Only 14 of the additional 189 responses added to the data in round three were scored into this bin as a 3.2 (Table 5).

Table 5. *Round 4: EION confusion matrix (Cohen's Kappa=0.683). Reference is the expert score; Prediction is the computer predicted score for a given response.*

	Reference									
Prediction	1.2	2.1	2.2	2.3	3.1	3.2	4.1	4.2	5.1	Sum
1.2	36	8	8	2	1	0	0	0	0	55
2.1	9	51	0	1	5	0	0	0	0	66
2.2	5	1	42	2	9	4	0	4	1	68
2.3	8	5	9	126	5	1	3	1	0	158
3.1	2	0	20	0	166	2	22	28	4	244
3.2	0	0	0	0	0	0	1	0	0	1
4.1	0	0	0	0	1	5	48	1	13	68
4.2	0	0	4	0	19	2	2	158	26	211
5.1	0	0	1	0	2	0	7	5	80	95
Sum	60	65	84	131	208	14	83	197	124	966

These 14 responses were not included in the model for the last round (section 3.2) as they were an emergent code and only coded for during the additional 189 responses, but when these 14 responses were added into the training set which increased to a Kappa of 0.683. So, despite having a very low number of responses, the addition of the "mistakes" bin allowed the language in each bin to become more distinct which, overall, clarified the other upper bins. For instance, before the addition of bin 3.2, only 5 of 83 responses were predicted correctly by the computer model for bin 4.1, but after the addition of bin 3.2, the model correctly predicted 48 of 83 in bin 4.1. This trend also held true for the other upper bins (3-5) in the confusion matrix.

3.4 Additive or Overlapping Language in Rubric Bins

During iterative model development, we noticed that responses with higher levels of sophisticated reasoning (i.e. upper levels of the LP) seem to have some amount of additive or overlapping language. The improvement in model performance from the replacement of synonyms and addition of bin 3.2 to reduce diversity in student response drove us to dig deeper into the classifications of the holistic rubric. Upon further evaluation, bins in the upper levels seemed to have additive ideas or overlapping language in the rubric descriptions that potentially cause the computer difficulty learning patterns of text in the student responses for holistic predictions.

Many of these additive and overlapping issues in language seem to be combinations of specific conceptual components. Some of these 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 into an analytic framework and that such a rubric might improve computer model performance.

For instance, compare the following two codes, as described in the coding rubric. Color has been added to highlight concepts that are shared between these codes.

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.*

Notice how both codes, 4.1 and 5.1, include the concepts: "will make the electrical gradient stronger than concentration gradient", "cause net movement of K+ into the cell", "make the membrane potential more negative", and "decrease the concentration gradient." The difference between these two codes is that 5.1 includes all the concepts of 4.1, but with the addition of the concept "making the membrane potential more negative than the equilibrium potential." This means to distinguish between levels 4.1 and 5.1, the key concept is "making the membrane potential more negative than equilibrium potential". This concept must be consistently identified by the computer model in order to make an accurate prediction between these two levels. Since some of these concepts also appeared in other rubric bins, aligned with other LP levels, the holistic rubric was then deconstructed to determine where and how often this additive and overlapping language was occurring.

3.4.1 Deconstruction

To see the extent of language overlap in the EION holistic rubric, we deconstructed the holistic rubric (a total of 9 bins) for student reasoning into finer-grained, conceptual bins, which represent separate, conceptual or reasoning pieces that are combined in some way to make a holistic bin. These smaller, conceptual components could be determined by presence/absence scoring. In order to be able to recombine these individual components into the holistic score – which keeps true to the purpose and intention of the holistic rubric - we decided to use Boolean logic operators (*AND, OR, MAY HAVE*) to combine the analytic bins to their holistic derivatives. Overall, EION holistic rubric went through three rounds of deconstruction - outlined below.

3.4.1.1 First Deconstruction

The first deconstruction of the EION holistic rubric (Table 6) determined that there were 16 individual, conceptual components within the 9-holistic level rubric. These conceptual components were the fine-grained concepts that could be present in multiple levels of the holistic rubric, but retained the analytical rubric property of not being mutually exclusive. Conceptual pieces 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 that result in the same LP level and these responses could include lower level LP reasoning. For example, these two responses were both coded as 5.1, but the first response reasons about decreasing membrane potential which the second response reasons about decreasing the concentration gradient to make the equilibrium potential more positive:

Code	Rubric Description
5.1	*"If the membrane potential is made more negative than E(K+), positive potassium ions will move back into the cell due to the electrical gradient into the cell overriding the concentration gradient out of the cell. This could be done by actively moving positive ions out of the cell or negative ions diffusing into the cell."*
5.1	*"One way this can be done is changing the membrane potential to something more negative. Another way is keeping the membrane potential at its current -70mV and increasing the outside potassium concentration to let's say 50mM as an example. The nernst equation specifies EK+= 61.5/z * log([K+] outside the cell / [K+] inside the cell) so when the concentration is outside the cell is increased the equilibrium potential increases. In our example with a new outside molarity of 50mV the equilibrium potential is -29mV. In order for the membrane potential to become -29mV from -70mV, K+ ions must move into the cell to increase the positive charges present. Thus, electric forces are stronger than the gradient forces acting on K+."*

Table 6. First deconstruction matrix of the EION assessment example.

Code	MP < EK	Make electrical gradient stronger than concentration 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	Decrease MP	decreasing K+ inside the cell	increasing K+ outside the cell	decrease concentration gradient to make EK+ more positive	Change MP	MP more positive	ions move from high to low concentrations	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																

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. For example:

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."*

3.4.1.2 Second Deconstruction

A second cycle of deconstruction was completed based on discussion between the experts in an effort to find and remove any unnecessary concepts in the analytic rubric. This cycle resulted in the reduction of the rubric from 16 to 8 analytical bins (Table 7). The first change reflected removing the "ions move to reach equilibrium" and "ions move from high to low concentrations" bins, because any holistic level could include these concepts. In other words, both these bins were *MAY HAVE* in Boolean logic in all sublevels of the holistic rubric and therefore these concepts were not defining features of any specific sublevel in the rubric. Thus, these optional ideas can be thought of as "noise" for this item – or additional reasoning students use to support their complex thinking, instead of the distinct LP reasoning we are interested in capturing. The bin "treats electrical and chemical gradients independently" was also removed because of the human difficulty in identifying this concept in student reasoning language. The "mistakes" bin, or where students are reasoning with higher level reasoning but make mistakes, in their understanding, generally in the physiology content, was also vague in rubric definition but very important in where students were placed holistically, so this bin stayed in the analytic rubric as we attempted to find more positive exemplars of student language for clarification.

After removal of bins, the next change in the second deconstruction cycle attempted to identify the truly distinguishing features of the analytic rubric 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 the holistic code 2.1. These two concepts were combined with an *OR* statement into one analytical bin, because no matter which concept the student used to reason with, they would always be coded as a 2.1. Other bins that were also combined for the same reasoning included the pairs: "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".

Table 7. Second deconstruction matrix of the EION assessment example.

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 the reduction of analytic rubric bins from 16 bins to 8, the deconstruction process also reduced the number of rows, or possible combinations for Boolean operators from 23 to 14, but maintained the original 9-level holistic rubric codes. However, there were still some overlapping and additive concepts such as students can have the concept bin "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 this bin was necessary to code for each of those holistic levels, or was it only necessary for a few.

Coding of 100 student responses (partitioned into the different levels of the original holistic code and then randomly selected) by two coders was done for validation of the second deconstructed rubric. The Cohen's Kappa between two experts across the eight analytical bins ranged from 0.653 – 0.890 with 3 of the 8 bins below 0.7. When the analytical rubric codes were combined with Boolean logic to determine the holistic code, the Cohen's Kappa between coders was 0.683. A comparison was also made of the original holistic codes, or the codes that were used to holistically categorize the student responses before deconstruction, with each expert's calculated holistic code via Boolean logic using their analytic codes. 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. Thus, there were discrepancies between what holistic code was originally given via a holistic rubric and that given by analytical coding and then combined into the holistic code. With this validation effort, a third round of deconstruction began to uncover these discrepancies.

3.4.1.3 Third Deconstruction

The third round of deconstruction added a single bin (Table 8) with the intention of clarifying some of the individual bins that performed poorly during validation. The reason for the additional bin was to create two bins in order to replace the single "mistakes" bin to improve clarification. Thus, the bin "mistakes" was removed and bins that captured these mistakes more clearly and precisely were added: "make membrane potential greater than equilibrium potential" and "make membrane potential positive". After discussion, the overlapping bin of "make membrane potential more negative" outlined above in the first deconstruction (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. The higher holistic levels (4.1, 4.2, and 5.1) could always include lower level reasoning in their answer, or could be coded to have this concept present, but this bin was not essential for students to use in their reasoning to be assigned codes 4.1, 4.2 or 5.1. For example:

Code	**Rubric Description**
4.1	*"Change the concentration gradient by increasing the [K+] outside the cell so that the concentration gradient is less than the electric gradient and in the opposite direction. Decrease the membrane potential so that the electric gradient is greater than the concentration gradient and in the opposite direction."*
3.1	*"Change the membrane potential more negative - increase the K+ ion concentration out of the cell that is higher than the concentration inside the cell Because, the K+ ion is positive charge, which is attracted by negative charge. If we change the membrane potential more negative, the electric forces will increase that attract the K+ ions into the cell. If the concentration is higher than the inside, the K+ ions will move into the cells until reach to the equilibrium."*

Table 8. Third deconstruction matrix of the EION assessment example.

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									

Another change in the rubric included the level of exclusivity of the concepts in codes 2.1, 2.2, and 2.3. Some student responses were difficult to place into only one holistic sub-level code (2.1, 2.2, or 2.3) since these concepts were not mutually exclusive, as the holistic rubric as first suggested, with some students reasoning with all three concepts. These concepts remained as three different analytical bins, but the holistic codes and combinations were changed to reflect that as long as the student has any one of these concepts, the response would be coded as a 2.1.

By distinguishing the bins by concepts and holistic levels more clearly, the rows in the rubrics were reduced from round two of 14 rows to 11 rows. The holistic rubric was also reduced from being a 9-level rubric to a 7-level rubric. Some Boolean logic statements were also reduced by removing some of the additive nature in some combinations, such as removing "decrease membrane potential" *OR* "make cell exterior more positive" *OR* "make cell interior more negative" from codes 4.1, 4.2, and 5.1. While the higher holistic levels still had overlapping or additive concepts, these were a lot more manageable in a more defined rubric matrix of 9 x 11 (Table 8) rather than the original version 16 x 23 (Table 6).

Analytical coding of an additional 50 student responses (also partitioned and then randomly selected) was completed for validation by two coders using the third version of the deconstructed, rubric. Cohen's Kappas between coders ranged from 0.650 - 1.00 with only 1 of the 9 bins below 0.7. The bin which had the lowest Cohen's Kappa, "decrease concentration gradient to make EK+ more positive," did not change in Cohen's Kappa from the second rubric version (0.653) to the third (0.650). While the experts agreed that this concept was important to capture because students use it to support their reasoning in their responses, this concept was very infrequent. For instance, the third rubric validation only had 2 positive cases of this bin in the subset of 50 responses, so a single discrepancy between coders would considerably decrease the Cohen's Kappa. The Cohen's Kappa between the two expert's calculated holistic codes, combined from the analytical bins using the determined Boolean logic, was 0.873. With almost all analytical bins performing well and the holistic codes having a high degree of reliability, the experts agreed that there was no need for another round of deconstruction.

3.4.1.4 Deconstruction Summary

Overall, comparing the holistic rubric to the 3-iteration-deconstructed developed analytical rubric, the analytical rubric clearly defines how students' reason to be given a specific holistic score. The final deconstructed rubric nested from the holistic rubric (Table 2) was represented by a matrix containing 16 integrating concepts, with 23 possible combinations which then placed the response into one of 9 holistic codes. In the last analytical deconstruction (Table 4), 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.

4.0 Conclusion

Our study identified several challenges facing the production of computerized scoring models for written constructed responses including: keyword synonyms, infrequent responses for some bins or including relevant concepts, student mistakes, and overlapping qualifiers in rubric description bins; these issues were identified within the iterative development process of just one item. Keyword synonyms can be preprocessed to replace multiple words with the same meaning, so that the computer can treat these features with the same weight in classification algorithms.

This increases model performance when specific synonyms are infrequent in student responses. If the total number of responses for a bin are very infrequent, or < 5% of the data have that code, then more data may need to be coded, in order to have sufficient examples for the computer to "learn" the patterns for that bin. Students who make mistakes can make mistakes in their reasoning in a variety of ways. The computer looks for common patterns in student language, which makes identifying and distinguishing between a variety of student mistakes that are coded similarly by humans, difficult for the computer. Mistakes should be coded based on a more specific definition in the rubric to improve computer models. However, there is caution to adding a holistic bin which captures mistakes with some amount of valid reasoning, as students can articulate a variety of mistakes but still use some relevant reasoning in their responses. Other items have shown that these mistakes confuse the computer with the correct, holistic bin that has some overlapping features commonly found in the mistakes bin. One approach to dive into the overlapping complexity in mistakes, and fine-grained components, is by deconstruction.

After deconstruction of a holistic rubric, we found that it is possible to keep "all the pieces" of a holistic rubric as part of an in-practice analytical rubric. Deconstruction of a holistic rubric is one way to approach the challenges of the heterogeneity of capturing student reasoning in a content rich assessment. These methods described above could help to improve the quality of large-scale assessments targeted at uncovering scientific reasoning in physiological, and scientific, education contexts. This deconstruction process will continue to be explored and validated by coding student responses derived into an analytical rubric which will hopefully reduce coder cognitive load and potentially improve human-human IRR, which in turn, improves computer modelling performance.

5.0 Acknowledgements

We thank 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 funded through grants from the National Science Foundation (NSF; DUE 1660643). Any opinions, findings, and conclusions or recommendations expressed in this manuscript are those of the author(s) and do not necessarily reflect the views of the NSF.

6.0 References

Aggarwal, C.C., & Zhai, C. (2012) A Survey of Text Classification Algorithms. In: Aggarwal C., Zhai C. (eds) Mining Text Data. Springer, Boston, MA

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

American Association for the Advancement of Science (AAAS). (2011). Vision and change in undergraduate biology education: A call to action. Washington, DC: American Association for the Advancement of Science.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Collingwood, L., & Wilkerson, J. (2012). Tradeoffs in Accuracy and Efficiency in Supervised Learning Methods. Journal of Information Technology & Politics, 9(3), 298-318. doi:10.1080/19331681.2012.669191

Corcoran, T., Mosher, F.A., & Rogat, A.D. (2009.) Learning Progressions in Science: An Evidence- based Approach to Reform (Philadelphia, PA: Consortium for Policy Research in Education).

Dunbar, K. (2000). How scientists think in the real world: Implications for science education. *Journal of Applied Developmental Psychology, 21*, 49-58.

Doherty, J.H., Scott, E.E., Cerchiara, J.A., McFarland, J., & Wenderoth, M.P. (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

Hartley, L. M., Wilke, B. J., Schramm, J. W., D'Avanzo, C., & Anderson, C. W. (2011). College Students' Understanding of the Carbon Cycle: Contrasting Principle-based and Informal Reasoning. *Bioscience, 61*(1), 65-75. doi:10.1525/bio.2011.61.1.12

Haudek, K.C., Moscarella, R.A., 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.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review, 2, 130-144.

Jurka, T.P., Collingwood, L., Boydstun, A.E., Grossman, E., & Van Atteveldt, W. (2012). RTextTools: Automatic Text Classification via Supervised Learning. R package version 1.3.9. http://CRAN.R-project.org/package=RTextTools

Knapp, A. K., & D'Avanzo, C. (2010). Teaching with principles: toward more effective pedagogy in ecology. *Ecosphere, 1*(6), art15. doi:10.1890/es10-00013.1

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated Scoring of Constructed-Response Science Items: Prospects and Obstacles. *Educational Measurement: Issues and Practice, 33*(2), 19-28. doi:10.1111/emip.12028

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching, 53*(2), 215-233. doi:10.1002/tea.21299

Michael, J., Cliff, W., McFarland, J., Modell, H., & Wright, A. (2017). The core concepts of physiology: A new paradigm for teaching physiology. New York, NY: Springer.

Modell, H.I. (2000). How to Help Students Understand Physiology? Emphasize General Models. Adv. Physiol. Educ. *23*, S101-S107.

Montgomery, K. (2002). Authentic tasks and rubrics: Going beyond traditional assessments in college teaching. College Teaching, 50 (1), 34-39. Doi: 10.2307/27559075

Moskal, B.M., & Leydens, J.A. (2000). Scoring rubric development: Validity and reliability. Practical Assessment, Research & Evaluation. 7(10).

Nehm, R.H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. Journal of Science Education and Technology, 21(1): 183-196.

Nehm, R.H., Ha, M., Rector, M., Opfer, J.E., Perrin, L., Ridgway, J., & Mollohan, K. (2010) Scoring Guide for the Open Response Instrument (ORI) and Evolutionary Gain and Loss Test (ACORNS). Technical Report of National Science Foundation REESE Project 0909999.

NRC. (2000). How people learn: Brain, mind, experience, and school: Expanded edition (Washington DC: National Academies Press).

Parker, J.M., Anderson, C.W., Heidemann, M., Merrill, J., Merritt, B., Richmond, G., & Urban-Lurain, M. (2012). Exploring undergraduates' understanding of photosynthesis using diagnostic question clusters. *CBE-Life Sciences Education, 11*, 47–57.

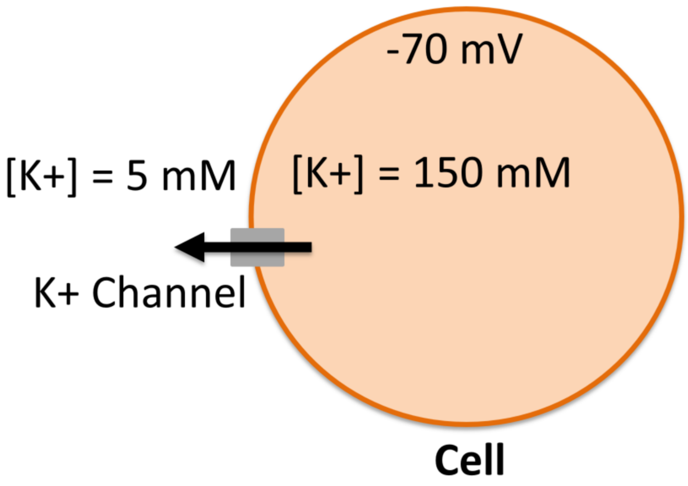
Rice, J., Doherty, J. H., & Anderson, C. W. (2014). Research and Teaching: Principles, First and Foremost: A Tool for Understanding Biological Processes. *Journal of College Science Teaching, 43*(3), 74-82.

Smith, C.L., Wiser, M., Anderson, C.W., & Krajcik, J.S. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and atomic- molecular theory. Meas. Interdiscip. Res. Perspect. *4*.

Stemler, S. 2001. An overview of content analysis. Practical Assessment, Research & Evaluation. 7 (17).

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.

Urban-Lurain, M. , Cooper, M. M., Haudek, K. C., Kaplan, J. J., Knight, J. K., Lemons, P. P., Lira CT, Merrill JE, Nehm RH, Prevost LB, Smith MK, & Sydlik, M. (2015). Expanding a National Network for Automated Analysis of Constructed Response Assessments to Reveal Student Thinking in STEM. *Computers in Education Journal,* 6(1), 65-81.



The figure shows a cell with the following labeled:

- potassium (K+) ion concentrations
- membrane potential (mV)
- K+ channel

In this situation there is net movement of K+ ions out of the cell (as indicated by arrow).
What can we change to cause net movement of K+ **INTO** the cell? Identify as many ways as you can and explain how each causes K+ to move into the cell.

Figure 1. Constructed response item "EION" used for the deconstruction process example.

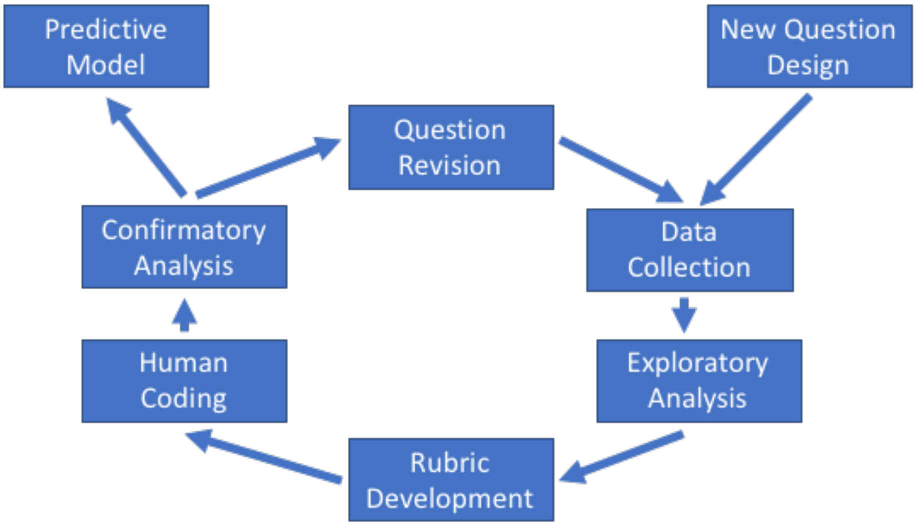


Figure 2. Overview of the Question Development Cycle (QDC) workflow.