Coding Student Responses: An Iterative and Construct-Driven Approach that Uses Multiple Sources of Data

Tina Chiu, Linda Morell, and Mark Wilson

University of California, Berkeley

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A100692 to University of California, Berkeley. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract

We describe a three-feature approach to coding student responses to middle school science test items. The three features are (1) a construct-driven foundation, (2) using multiple sources to inform development, and (3) performing iterative cycles of revising and fine-tuning to ensure high-quality and usable scoring guides. We propose this method for creating meaningful and reliable scoring guides that are finite and exhaustive, ordered, context-specific, research-based, and have well-defined categories. First, we describe our approach and then provide an empirical example situated in the science content domain of the structure of matter.

Introduction

We begin by discussing the *outcome space* (Marton, 1981), an assessment term in its most general form. We use this term in a broad sense to apply to any set of qualitatively described categories for assigning codes to student responses to assessment tasks. An outcome space can take different forms. For example, a test-question scoring guide is a common format for an outcome space to take. The important assessment features of an outcome space are that the evaluative criteria established are clear for distinguishing qualitatively different responses and that those differences are mirrored as levels within a theoretically hierarchical structure.

This paper describes an approach to coding student responses, which is based on a (1) construct-driven approach that uses (2) multiple sources of data (3) iteratively. This procedure provides insightful information on the meaningful, valid, and reliable scoring guides and illuminates the importance of this often-routine task. This step is essential for providing evidence to support the theory behind an assessment. Examples are provided to help illustrate the procedure.

Theoretical Framework

As a part of the movement for "performance" and "authentic" assessments in the late 1980s and 1990s (e.g. Masters, Adams, & Wilson, 1990; Wiggins, 1990) there has been an emphasis in the area of measurement on item design and coding student responses (Brookhart, 1999; Mertler, 2001; Moskel, 2000; Moskal & Leydens, 2000; Nitko, 1989; Wilson, 1989; Wilson, 2005). These types of assessments often require the qualitative judgment of raters, which can introduce additional errors, such as inconsistent or unreliable ratings. Raters can be anyone—often teachers or outside persons trained specifically to code a specific assessment. Developing an outcome space is one way to minimize such errors.

In the sections below, we first provide a background review of what we mean by the outcome space. Next, we describe additional considerations needed for developing the outcome space in a science context. Finally, we conclude this section with a description of the features that we consider to be essential in developing a meaningful, useful, and reliable outcome space.

Outcome Space

The term *outcome space* (Marton, 1981) is used here, but more common terms include *scoring guide* or *rubric*. We apply this term in a broad sense to any set of qualitatively described categories for coding responses to items. Literature on designing and developing the outcome space define different varieties (Allen & Tanner, 2006; Mertler, 2001; Moskal, 2000; Popham, 1997), such as whether responses should be scored analytically or holistically. In the analytical outcome space, the criteria of interest are each evaluated on a separate scale, whereas in the holistic case, the criteria are evaluated together onto one scale, producing one overall, aggregated score. While the outcome space can take on these different formats, the important feature is that the evaluative criteria established are clear for distinguishing qualitatively different responses.

Jonnson and Svingby (2007) conducted a meta-analysis of 75 studies to investigate the benefits, if any, of using outcome space for performance assessments.

Specifically, they focused on whether it could help enhance the reliability and validity of judging complex competencies, and promote student learning and/or improve the quality of teaching. They concluded that the outcome space could contribute to both inter- and intra-rater reliability if it is analytic, that is specific to a topic, and includes both exemplars and rater training. The outcome space could also help promote learning and teaching because it explicitly lays out the criteria and expectations for both students and teachers. This finding emphasizes the importance of carefully constructing the assessment according to a theory.

Methods to developing, analyzing, and improving the outcome space vary (see Mertler, 2001; Moskal & Leydens, 2000; Parke, 2001; Parke, 2002; Wilson, 2005 for examples). The approach we describe below shares qualities with many approaches. We highlight three main features: (1) construct-driven foundation, (2) multiple sources of information, and (3) the iterative cycle. Our procedure attempts to maximize the reliability and validity for developing useful, informative assessments through the focus on creating a strong outcome space.

We frame our work within the Bear Assessment System (BAS, see Wilson, 2005; Wilson & Sloane, 2000). BAS consists of four building blocks: construct, item design, outcome space, and the selection of a measurement model for analysis. Figure 1 provides a visual representation of BAS.

Insert Figure 1 about here

The first building block, the construct, includes developing the theory of student understanding based on research. Typically, descriptions are provided for distinguishing between increasing levels in sophistication of understanding in a *construct map*. After describing the construct of interest, items are then designed to target certain levels of understanding as proposed in the construct map. From these items, the outcome space is formed which maps student responses to the construct map. Based on the outcome space, a measurement model is selected to run the analysis. The empirical results are reviewed to see how well it supports the theory, and if necessary, the construct is revised with this new information. Then BAS is repeated until no further revisions need to be made.

Often assessment materials iterate through the qualitative inner loop (Berson, Yao, Cho, & Wilson, 2010) several times before moving on for a whole cycle of development. The qualitative inner loop consists of the construct, items design, and outcome space. This is necessary because often issues arise within these first three steps that need to be addressed before material is ready for larger-scale testing. It is also necessary because in theory and in practice the first three parts of the BAS are tightly connected and reflective of each other. Thus, while this paper will focus on the outcome space, it must be kept in mind that this is only part of the measurement process. In addition, references to other components are inevitable and necessary.

Outcome Space for Science Content

Rubrics and scoring guides are commonplace in education and assessment as a method for situating a student’s understanding within a larger frame from elementary through high school (Moskal, 2000). This is true for science education and assessment, although less prevalent in science than in other areas, for example, writing. Rubrics, when used by teachers, allow standards and scores to be explicit. This gives the student a clear sense of what is expected, how the outcome space is interpreted, and how they might approach a science question.

Researchers have pointed out that the development of science scoring guides can be time consuming (Allen & Tanner, 2006; Morell & Wilson, 2013). However, once an accurate, exhaustive, and high quality rubric has been created, grading and feedback to students can be streamlined and done fairly quickly. In addition, Allen & Tanner (2006) further point out that a high quality rubric is important to have if a course is team-taught or there are multiple people coding student responses. A clear and comprehensive science rubric allows researchers to effectively communicate expectations with coders so that multiple coders apply the same criteria to student responses.

Key Characteristics for an Effective Outcome Space

Establishing a "good" outcome space is a key component in effective measurement. But what exactly makes an outcome space "good"? Wilson (2005), following Masters, Adams, & Wilson (1990), defines the following properties as essential for developing an effective outcome space:

(1) **Well-defined**. There should be a clear description for each coding category and some examples of item responses. This aids in distinguishing between the different categories.

(2) **Exhaustive**. All possible responses could be coded into one, and only one, of the categories. However, there should be a finite number of categories, usually less than ten, to keep the categories distinct yet usable.

(3) **Ordered**. Some categories should be lower on the corresponding construct map than others. The basis for this ordering should reflect both theory and the empirical results.

(4) **Context-specific**. The outcome space should be specific to the construct and the context for the assessment.

(5) **Research-based**. The categories should be theoretically-grounded in research that ties a particular response to a particular level of the construct.

The three essential features that we highlight in our approach ensure that the criteria above are met so that the outcome space materials created have those characteristics.

For instance, the construct-driven foundation ensures that the outcome space is research-based and context specific. Because the construct map is ordered to reflect increasingly sophisticated understanding, it also allows the outcome space to be ordered to reflect these differing levels. Using multiple sources of information helps clarify categories so that they are well-defined. Finally, the iterative cycle helps confirm whether there are enough categories to capture all possible student responses but are also finite so that there are a reasonable number of categories for its users.

While there were many steps for developing an outcome space, we only focus on three. For each feature, concrete examples from a middle school science assessment will be provided to help illustrate this procedure.

Feature 1: Construct-Driven Foundation

The construct itself is the essential core for the assessment and outcome space. Jonnson and Svingby (2007) emphasized that simply having an outcome space does not guarantee validity. Rather, validity should start with a strong theoretical base for the construct of interest. This theoretical base is the pivotal foundation for all items and their corresponding outcome space. Keep in mind that the items were designed and selected to target specific levels in the construct map, as represented in BAS in Figure 1 and discussed above.

An initial outcome space is drafted for every item in the construct to ensure the consistency between the theory, item, and outcome space. The initial outcome space includes the targeted construct map level and a brief description of the “correct” or intended response. Developing the outcome space based on the construct helps meet the “context-specific” and “research-based” properties of an effective outcome space. Because the construct is also ordered, it also ensures that the outcome space will be ordered.

Learning Progressions in Science Project (LPS)

The Learning Progressions in Science (LPS) project is a multi-year project that investigates how students learn about the Structure of Matter domain. This domain begins with students conceptualizing matter through its macro properties, such as having mass or taking up space, then moving to a more nuanced understanding of the particulate nature of matter. Underlying LPS is a theoretical learning progression, based on previous research (Black & Wilson, 2009; Smith et al., 2006; Smith et al., 2004). The proposed learning progression consists of six constructs, each with its own construct map and is shown in Figure 2. The examples used from this data are from the 2011-2012 data collection cycle. For simplicity, all examples of items and outcome space in this section are selected from the *Particulate Explanations of Physical Changes* (EPC) construct.

Insert Figure 2 about here

An example of the construct map developed for the EPC construct is provided in Figure 3. For LPS, we performed a literature review and worked closely with content experts and middle school science teachers to develop a construct map for each of the construct variables in the Structure of Matter learning progression. An initial outcome space was drafted for every item using the construct map described above as a guide. Because open-ended items can elicit a wide range of responses, explicitly mapping the item to a specific level within the construct map ensured that the item can be mapped back to the original theory and thereby provide information regarding the validity of the theory. Figures 4 and 5 provide an example of an item and its corresponding initial outcome space, respectively. In this open-ended item, students were asked to describe the arrangement of water molecules in water vapor. The reader should note the first column of the outcome space is labeled “Construct Map Level,” which directly maps the level of student responses to the construct map. This ensures the alignment of the outcome space with the theory.

Insert Figures 3, 4, and 5 about here

Feature 2: Multiple Sources of Information

After drafting an initial outcome space, multiple resources were used to help incorporate more information. Figure 6 summarizes the various resources, which includes a comprehensive literature review, content experts, teachers, and students.

Insert Figure 6 about here

The literature review helped all researchers obtain a better understanding of previous work in the area. While this literature review was conducted prior and during development of the construct map and items, it remains a critical resource when designing the outcome space. Likewise, content experts can also provide informative feedback, regarding the strength of the theory and also even the “correctness” of the outcome space.

The literature review also provided information regarding common misconceptions found among students regarding science content. It is often helpful for instructional purposes to include these misconceptions in an outcome space so that practitioners can readdress them specifically with students. Providing a research-based outcome space ensures we draw upon and leverage previous research.

Teachers who teach the target population also provide insightful information regarding how students learn this content and how they may respond to certain items. They also provide feedback regarding the items, its usefulness, and the type of information they would like the outcome space to provide as a tool for teaching. Teachers also help categorize student responses into the response categories and help decide how context-specific responses should be, and relate them to the levels identified on the construct map.

Lastly, students also contributed to the development of the outcome space. This includes their responses via think-aloud interviews and written responses. In these interviews, students verbally express their thoughts as they work through a selected number of items. From these interviews, researchers could see if the item is eliciting the desired thought-processes in students, or in other words, if the student is responding in an expected manner. This step is essential for gathering validity evidence because if the items are appropriate, then they will elicit the intended thought processes.

In addition to the interviews, student written responses are examined through the pilot and field test administrations. Before the field test administration for the project, a pilot administration is conducted, where a small sample of students take the items under test-like conditions. Student responses from this pilot administration are useful as they help provide an idea of how students actually respond to the items in a standard test-taking situation. It also provides an opportunity to identify and fix problems with an item before the field administration where a much larger group of students take the items. Both the interviews and administrations provide “student language” or “student speak” so that the outcome space and items can use more authentic student-centered language.

By incorporating these multiple sources of information, the outcome space becomes well-defined. Resources such as the literature review, content experts, teachers, and students also help provide guidance in ordering these descriptions into levels of more sophisticated understanding.

LPS Continued

For LPS, the draft outcome space was presented to middle school science teachers and the content expert for feedback. Feedback is especially useful for enhancing the descriptions of the different categories. By having these different resources inform the development of the scoring guide, there is clarification on what essential and useful information can be extracted for both informing the learning progression theory and also for informing teaching. It also serves as an accuracy check for the descriptions provided in the outcome space. Information from these sources provides a strong initial foundation for distinguishing student responses into useful, distinct, and informative categories.

From student interviews and the pilot administration, the team identified "exemplar" student responses for the outcome space to help guide raters. "Exemplar" responses are those that are truly representative of the code categories. They are clear examples of the description. Figure 7 provides an updated outcome space after using the information from the various sources. In this figure, there are two notable additions. First, there is the addition of the "Examples" column, where actual student responses are used to illustrate the description. Second, the "Construct Map" column differs from the corresponding "Code" column, or the numbers located in the second column. The code distinguishes between qualitatively different responses, but do not necessarily provide any additional information to the theory. This more fine-grained distinction may be a useful tool for teachers to see the qualitatively different ways students respond to an item. It may also be useful for running future analyses, such as the ordered partition model (Wilson, 1992), where responses cannot be completely ordered, but the distinction between responses are still desired.

Insert Figure 7 about here

Feature 3: Iterative Cycle

Lastly, this procedure is iterative. There are actually two iterative processes, one which helps refine the outcome space and the other which connects all the features described previously back to BAS.

First, the outcome space draft undergoes its own development cycle, called *moderation* sessions. This is an iterative testing ground for the outcome space before use on a larger scale. In this procedure, raters independently code a select number of student responses using the draft outcome space, usually around 10 to 15 per rater, for one item and then compare results. A minimum of three raters is recommended for quality assurance. All disagreements are discussed and changes are incorporated into the outcome space as needed for clarification. Sometimes, additional responses are flagged as representative or problematic and then included as examples to serve as a guide for future similar responses. Another round of moderation begins for the same item and continues until raters reach complete agreement. After complete agreement, the moderation procedure then begins for the next item.

While this iterative method is extremely time-consuming, it is an essential step in ensuring quality ratings, especially if multiple raters plan to evaluate the responses. The goal for this procedure is for the final outcome space to be as well-defined, exhaustive yet inclusive of the range of student responses, ordered, context-specific, and research-based, as possible so that raters would encounter as few problems as possible when coding student responses. Thus, for some items, it may be useful to also include examples of *borderline responses*. Borderline responses are those that appear to belong on the cusp of two categories. By demarcating these responses, it helps provide raters with a clearer definition of the limits of each category.

After completion of the outcome space, the items are then fully scored and analyzed using the selected measurement model. The empirical evidence stemming from the measurement model either provides evidence for or requires adjusting the construct, items, or outcome space. In the latter scenario, the cycle is repeated with adjustments to all the components of BAS as needed, thereby continuing the much larger iterative cycle. The iterative notion is an important feature not just within the outcome space component, but also in the larger measurement process.

LPS Continued

For the LPS project, after revisions were made to the outcome space with feedback from the multiple sources, the team began the *moderation* process. This process allows us to test the usability and clarity of the outcome space. Unclear categories were adjusted so that complete agreement among raters would be maximized.

Figure 8 shows the final outcome space for this item after the moderation procedure. Note the inclusion of a borderline response under the "Examples" column for code 1b. The response, "They are more scattered," initially gave raters difficulties because it could be coded as a "2" if one assumes that the student implied that the molecules in water vapor are more scattered than those in water or ice. However, because it was not explicitly stated, the final decision was to code the response as a "1b".

This helped clarify that an explicit comparison of water vapor molecules to those in other states was needed for a code of “2”.

Insert Figure 8 about here

While this procedure helps identify problematic categories, such as unclear categorical descriptions or difficulties with classifying student responses, it can also help identify problematic items that were not initially identified. For example, there were some items where no agreement could be reached after numerous moderations. Figure 9 provides an example of such a problematic item. This item followed a multiple-choice item and required students to provide an explanation for their multiple-choice selection. Coding student responses was problematic because it became difficult to identify any additional information beyond what was elicited in the multiple-choice item (i.e. many students wrote out their multiple choice selection). Subsequently, the item was removed from the analysis and underwent review by the research team and content expert.   After further investigation the item was removed from the study because it was deemed too flawed and problematic.

Insert Figure 9 about here

After running the measurement model for the item described earlier in Figure 5, it was revised for the new round of data collection. This revision is shown in Figure 10. The item is now converted to a multiple-choice item. All the information from the prior iteration was used to adjust the item. In addition, this new iteration also included the same procedure where information from teachers, content experts, and student responses were incorporated.

Insert Figure 10 about here

Inter-rater Reliability Results

Table 1 provides the inter-rater reliability results for the open-ended items in the Particulate Explanations of Physical Changes (EPC) construct from the LPS 2012 data collection. Three raters scored all the items in this construct. For open-ended items, two raters overlapped on a selection of at least 211 responses. The percentage of exact agreement for the fifteen open-ended items ranges from 73% to 97%, with only one item having less than 80% exact agreement. A common criterion for evaluating the inter-rater reliability is 70% exact agreement (Stemler, 2004; Jonsson & Svingby, 2007).

Insert Table 1 about here

Cohen's kappa ranged from 0.54 to 0.94 for the EPC items. Because Cohen's kappa adjusts for chance agreement among raters, the criterion for evaluation is slightly lower. Literature suggests that a Cohen's kappa between .40 and .60 is moderate and that values above .60 are considered strong (Stemler, 2004; Jonsson & Svingby, 2007).

The outcome space developed from this approach generated strong inter-reliability results. It appeared that independent raters were able to use the outcome space in the same manner and provide consistent ratings.

Conclusion

Establishing a valid and reliable outcome space is important for practitioner use, research, and quality judgment of student performance. Wilson (2005) has defined the outcome space as having the following characteristics: well-defined levels, exhaustive, ordered, context-specific, and research-based. The method we described has provided a clear way for incorporating these characteristics into the outcome space.

Our procedure illustrates the complexity of gathering sound evidence in a systematic way to accurately capture content knowledge. Three features, a construct-driven foundation, multiple sources of information, and an iterative cycle, are the key features to this method. Focusing on these ideas supports the development of high quality tools to capture and code student responses, resulting in sound measurement practices.

References

Allen, D. & Tanner, K. (2006). Rubrics: Tools for Making Learning Goals and Evaluation Critieria Explicit for Both Teachers and Learners. *CBE-Life Sciences Education,* (5), 197-203.

Berson, E., Yao, S., Cho, S., Wilson, M. (2010, April). *The Qualitative Inner-Loop of the BEAR Assessment System.* Paper presented at the International Objective Measurement Workshop (IOMW) in Boulder, CO.

Black, P, & Wilson, M. (2009, April). *Learning Progressions to Guide Systems of Formative and Summative Assessment.* Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Brookhart, S.M. (1999). The Art and Science of Classroom Assessment: The Missing Part of Pedagogy. *ASHE-ERIC Higher Education Report (27, 1).* Washington, DC: The George Washington University, Graduate School of Education and Human Development.

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

Marton, F. (1981). Phenomenography—Describing conceptions of the world around us. *Instructional Science*, 10, 177-200.

Masters, G.N., Adams, R.J., & Wilson, M. (1990). Charting of student progress. In T. Husen & T.N. Postlethwaiter (Eds.), *International Encyclopedia of Education: Research and Studies. Supplementary Volume 2* (pp. 628-634). Oxford: Pergamon Press. Reprinted in T. Husen & T.N. Postlethwaite (Eds.), (1994).

International Encyclopedia of Education (2^{nd} Ed.) (pp. 5783-91). Oxford: Pergamon Press.

Mertler, C.A. (2001). Designing Scoring Rubrics for Your Classroom. *Practical Assessment, Research & Evaluation*, 7(25).

Morell, L.M. & Wilson, M.R. (2013, March). *Uncovering the Learning Progressions in Middle School Science Instruction and Assessment.* Paper presented at BEAR Seminar at the University of California, Berkeley.

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

Moskal B.M & Leydens, J.A. (2000). Scoring Rubric Development: Validity and Reliability. *Practical Assessment, Research & Evaluation*, 7(10).

Nitko, A.J. (1989). Designing Tests That Are Integrated with Instruction. In R.L. Lynn, *Educational Measurement*, 447-474.

Parke, C.S. (2001). An approach that examines sources of misfit to improve performance assessment items and rubrics. *Educational Assessment, 7*(3), 201-225.

Parke, C. S. (2002). Mathematics performance assessment: Discovering why some items or rubrics don’t measure up. *Research in Middle Level Education Online, 25*(1).

Popham, W.J. (1997). What’s wrong—and what’s right—with rubrics. *Educational Leadership, 55*(2), 72-75.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children’s learning for standards and assessment: A proposed learning progression for matter and the atomic molecular theory. *Measurement: Interdisciplinary Research and Perspectives, 14,* 1–98.

Smith, C., Wiser, M., Anderson, C. W., Krajcik, J., & Coppola, B. (2004). Implications of Research on Children's Learning for Assessment: Matter and Atomic Molecular Theory. Washington: Center for Education, National Research Council.

Stemler, Steven E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4). Retrieved April 5, 2013 from http://PAREonline.net/getvn.asp?v=9&n=4.

Wiggins, G. (1990). The Case for Authentic Assessment. *Practical Assessment, Research & Evaluation*, 2(2).

Wilson, M. (1989). Empirical examination of a learning hierarchy using an item response theory model. *Journal of Experimental Education, 57*(4), 357-371.

Wilson, M. (1990). Measurement of developmental levels. In T. Husen & T.N. Postlethwaite (Eds.), *International Encylcopedia of Education: Research and Studies. Supplementary Volume 2.* Oxford: Pergamon Press.

Wilson, M. (1992). The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement, 16*(4), 309-325.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Wilson, M. & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*, 181-208.

Appendix

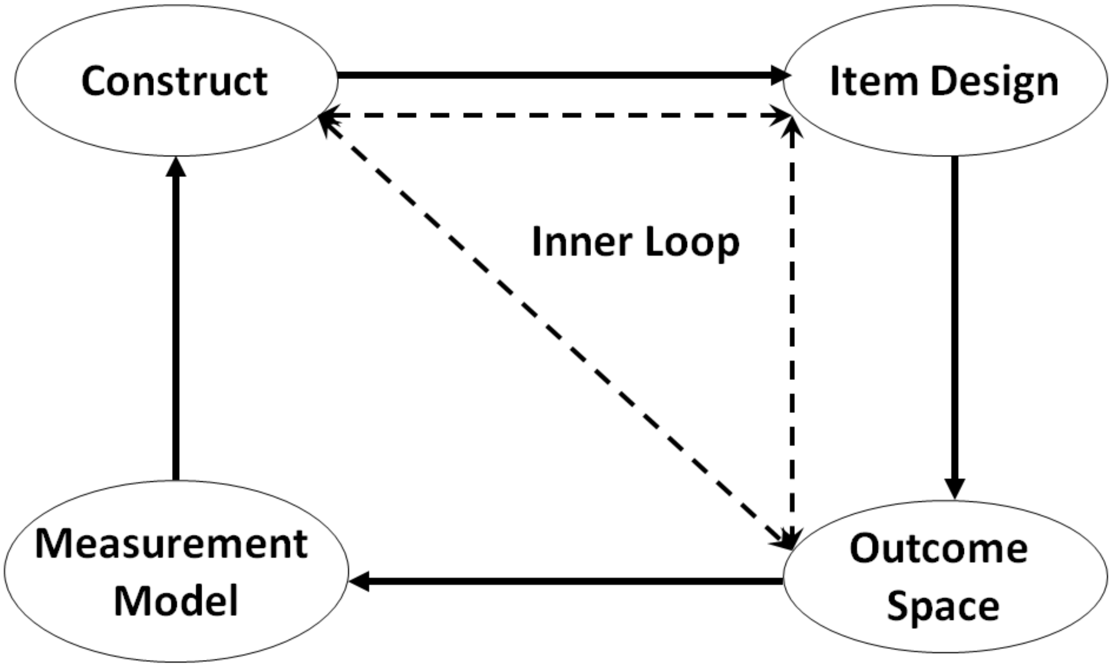

Figure 1. BEAR Assessment System (BAS) Figure.

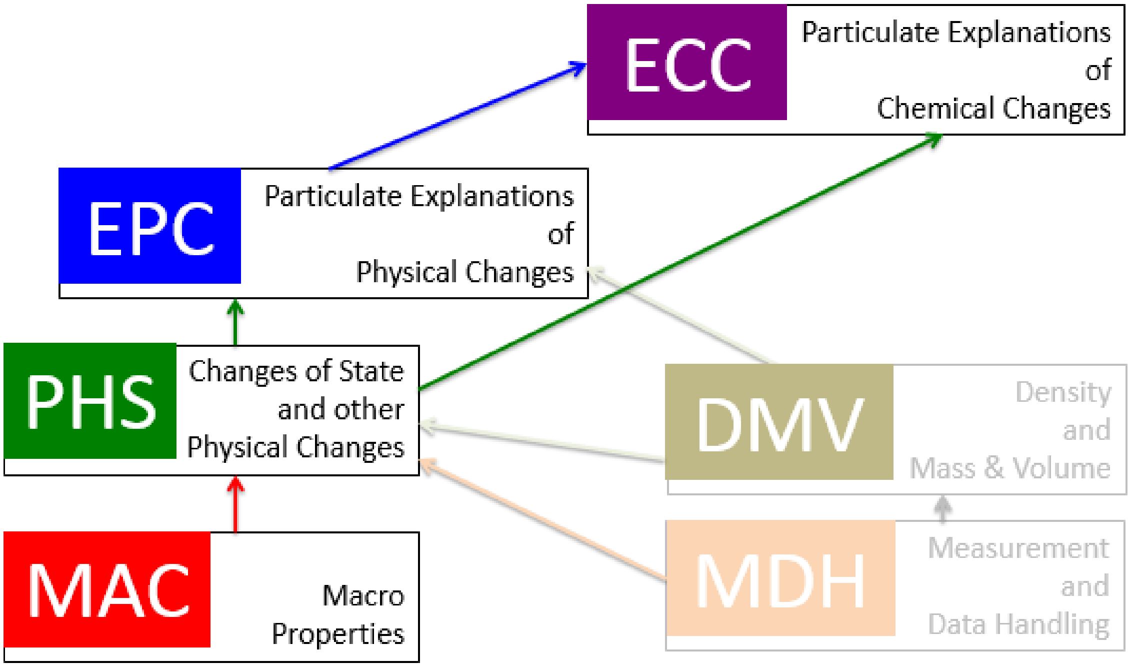
ECC
Particulate Explanations of Chemical Changes
EPC
Particulate Explanations of Physical Changes
PHS
Changes of State and other Physical Changes
DMV
Density and Mass & Volume
MAC
Macro Properties
MDH
Measurement and Data Handling

Figure 2. Learning progression for the Structure of Matter domain.

Level	Description	Explanation
5	Conceptions of particles	The micro-scale constituents of matter are now pictured as consisting of particles. Students understand that these “building blocks” are not merely miniature versions of macro-scale objects. That is properties such as physical change, density and temperature can be explained in terms of the way the particles are packed or move but that the particles themselves do not share all the same properties as the macro-scale objects they constitute.
4	Particulate Explanations	Students now understand that many macro-scale properties of matter in a given state, or at a given temperature, can be explained in terms of the existence of tiny particles that are packed and move in a certain manner. Changes in volume and density, and in other physical properties, when a material dissolves, warms up, melts, evaporates or condenses are caused by changes in the way the particles are packed or in the way in which they move. The differences in density between different substances can be explained by differences in the masses and spacings of constituent particles.
3	Particulate representations of matter and of the difference between phases	Students understand that matter is composed of atoms or molecules which are very small particles. Differences between solid, liquid and gas phases can be represented by differences in the way that particles of a substance are arranged, their ability to move freely and their distance from each other.
2	Reversibility	Students understand that transitions between different states of matter, dissolving in solutions, or thermal changes, can be bi-directional. Matter is not only able to change from one form to another while being conserved, but can change back to a previous state.
1	Conservation of Matter	Students understand that while state changes are possible (i.e., BOX 4), matter is not created nor destroyed (e.g., when a material dissolves, the mass of the solution remains the same, suggesting that the matter remains, just in a different form).
0	Matter not conserved	Students think that when a sample of substances is heated or cooled, or is dissolved, or changes between solid, liquid and gas phases, the sample can no longer be the same substance.

Figure 3. The initial construct map for the *Particulate Explanations of Physical Changes* (EPC) Construct.

The arrangement of molecules in **water vapor** is…

____

____

____

Figure 4. Example item from the *Particulate Explanations of Physical Changes* (EPC) Construct.

Construct Map Level	Code	Description
3	1	Student describes the arrangement of molecules as more spread out than both water and ice.
	0	Student provides an incorrect or off-topic answer.
	9	No response.

Figure 5. Sample draft outcome space for selected item in Figure 4.

Source	Description
Literature Review	Helped provide information on what should be included in the outcome
Content Expert	Feedback on construct map, codes, levels, and correctness.
Teachers / Co-Developer Meetings	Feedback on codes, levels, what they would find useful, common responses.
Student: Interviews	Researchers conduct think-aloud interviews with students to see how students “think” when solving items and ask clarifying questions.
Student: Pilot Administration	Small sample of students answer questions under test-like conditions.
Student: Field Administration	Administration of revised test to a large number of students.

Figure 6. Description of the multiple sources of information used in the Learning Progressions in Science (LPS) study.

Construct Map Level	Code	Description	Examples
3	2	Student describes the arrangement of molecules as more spread out than both water and ice.	
	1	Student describes the distance between particles as spread out, random, separated, expanded, or having no specific arrangement.	“all spread out.”
	0	Student provides an incorrect or off-topic answer.	“like evaporation.” “freely moving molecules.” “the arrangement is outrageous!”
	9	No response.	BLANK

Figure 7. Revised outcome space for selected item in Figure 4.

Construct Map Level	Code	Description	Examples
3	2	Student describes the arrangement of molecules as more spread out than both water or ice.	“Farther away than molecules in water.”
3	1b	Student describes the distance between particles as spread out, random, separated, expanded, or having no specific arrangement.	“they are more scattered.” [borderline] “spread out.” “all spread out.”
3	1a	Student describes the arrangement of molecules in water vapour as moving freely.	“freely moving molecules.” “very free”
	0	Student provides an incorrect or off-topic answer.	“like evaporation.” “the arrangement is outrageous!” “loose” “not packed”
	9	No response.	BLANK

Figure 8. Finalized outcome space for item in Figure 4.

Four students are trying to explain why ice is harder than water.

Student **A** says, "Liquid water and ice are both made of water, but the molecules of water in ice are harder than the molecules of water in liquid water."

Student **B** says, "Liquid water and ice are both made of water, but the molecules in ice have sharp edges and the molecules in liquid water have round edges."

Student **C** says, "Liquid water and ice are made of different substances. Liquid water is made of soft material, and frozen water is made of hard material."

Student **D** says, "Liquid water and ice are both made of the same molecules, but it is the different ways they are arranged that causes the differences in hardness. "

a. Which student do you agree with?

I agree with Student ______

b. Give ***another reason*** why the student you agree with is right:

____

____

____

Figure 9. Sample problematic item from the *Particulate Explanations of Physical Changes* (EPC) construct.

The arrangement of molecules in **water vapor** is......

A. Adjacent in a repeating pattern.

B. Stretched apart in an ordered pattern.

C. Tight in a random pattern.

D. Far apart in a disordered pattern.

Figure 10. Updated item from Figure 4 for next iteration of BAS.

Table 1

Inter-rater Reliability Results

Item Name	**Percentage of Exact Agreement***	**Cohen's Kappa***
1	85%	0.78
2	97%	0.93
3	94%	0.88
4	96%	0.70
5	80%	0.66
6	87%	0.76
7	76%	0.85
8	88%	0.79
9	97%	0.94
10	92%	0.74
11	92%	0.81
12	84%	0.69
13	94%	0.83
14	73%	0.54
15	94%	0.91

**Note*: Missing data for items were not included in this calculation.