Coordinating Assessments with a Learning Progression

Mark Wilson
University of California, Berkeley

Shih-Ying Yao
Singapore University of Social Sciences

Jonathan Osborne
Graduate School of Education, Stanford University

Acknowledgements: This research is based on work 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. The authors wish to thank all the teachers, fellows, and students from the San Francisco Unified School District who have participated in this research. The authors are also grateful to researchers and colleagues from the Berkeley Evaluation and Assessment Research Center, Stanford University, and Strategic Education Research Partnership for their contributions to the work where this research is based.

Abstract

The goal of this chapter is to illustrate the development and validation of assessments based on a learning progression for a scientific practice—argumentation. Without valid and reliable assessments, science educators will not be able to measure students' locations on a learning progression. The assessments are based on the learning progression in argumentation developed by Osborne et al. (2016) and a set of items developed using the BEAR Assessment System to elicit students' performance indicative of the waypoints in the progression. The goal of this study was to investigate evidence for the validity of the argumentation assessments. Specifically, we applied exploratory factor analysis to examine the dimensionality of the assessments, and then applied item response modeling to examine whether the hypothesized order of waypoints of the argumentation learning progression was confirmed by the set of items used. We also developed software reports for teachers, to give teachers quick and clear feedback from the assessments. We conclude by noting three challenges to the use of learning progressions and summarize how the methods described in this chapter can be used to help address those challenges.

Keywords: learning progression; validity; Wright map; argumentation

Coordinating Assessments with a Learning Progression

The concept of a learning progression has been implicated in curriculum and instructional ideas that have existed in the educational instruction and assessment literatures over many decades[1]. One early definition is as follows:

> Learning progressions are descriptions of the successively more sophisticated ways of thinking about an important domain of knowledge and practice that can follow one another as children learn about and investigate a topic over a broad span of time. They are crucially dependent on instructional practices if they are to occur. (CCII, 2009)

The focus of a learning progression is on the development of the learner's understanding: thus, curriculum activities must fit with the structure of the learning progression—and the topics in a textbook *should* follow the map identified by the learning progression, as should the curriculum standards. The learning progression must document the critical developmental points in a student's growth that are key to designing instruction to bring about change. The developmental points can be seen as defining a "roadmap" of important forms of student thinking (Black et al, 2011). In the long run, as for any other model, a learning progression must be judged by its utility. Its descriptions are ideally situated at a "meso-level" without being either so broad as to provide very little guidance for instruction and assessment (i.e., the "macro-level" that would be

[1] It is close to the concept of a learning *trajectory* as is often used in mathematics education (Clements & Sarama, 2009)

suitable for state and international testing, etc.), or so very fine-grained as to make it difficult for a teacher to use them in planning for a classroom setting (i.e., "micro-level") (Wilson, 2020).

A learning progression is a conceptual device intended to help address a foundational problem in education which is difficult to grasp due to its complexity: that is the interaction of disciplinary knowledge, student learning, and the accompanying appropriate instruction, and assessment to achieve an educational aim. That is, a learning progression is a contribution to the (complex) answer to the four-part question of how we are to establish the *necessary coherence* among: the goals of the education expressed as substantive disciplinary content, whether it is a part of language arts, mathematics, or science (as is the focus here); the principal steps in the development of a more sophisticated understanding of a given concept (i.e., signposts along the prime roads to understanding); the focus of instruction and ordering so as to align with the learning progression; and the foci for the assessment of the progress of students understanding as they learn (i.e., where they are on the "roadmap," and where they are going).

In addition, in any educational approach based on a learning progression, it is crucial to provide 'professional development for teachers on how to use the learning progression, and this should be based in the same quadripartite view—teachers must "know the discipline," "know how students learn," "be able to design and carry out instruction to assist that learning,"[2] and "know how to assess the success of their students."

[2] Although the specifics of that instruction are usually not contained in the learning progression.

In this chapter, we will focus on the fourth aspect of the use of learning progressions, the assessment aspect. To this end, we will not only explore this role for assessment, but also attempt to support our basic proposition. This is that: *Good measurement can provide important evidence in support of a learning progression, and the outcomes can be helpful for teachers, but it requires a systematic and rigorous effort by assessment developers to create materials that can help teachers in the classroom.* We will make this argument, in part, by using an example of assessment of a specific topic, scientific argumentation. As we work through the example below, we will follow these steps:

(i) create a progression as a sequence (or multiple sequences) of waypoints of student understanding (i.e., qualitatively distinct ways of understanding, seen as leading to a particular goal);

(ii) develop ways to assess these waypoints in each construct in this learning progression;

(iii develop reports (and associated professional development activities) to help teachers use the information from the assessments for instructional design and within the activities in a classroom; and

(iv) study whether the empirical information from students taking the assessments is consistent with the hypothesized progression.

One of the challenges of this process is that to do (i) well, one must already have succeeded (at least in part) in (ii) through (iii)—that is, this is a process that is not linear, as it might appear from the labelling above. Rather (i) through (iv), is an iterative process, and frequently it

involves other non-linearities where, for example, steps are skipped, and some may be fed-back or re-visited.

For the next three sections of this chapter we will illustrate, using the example of scientific argumentation, how assessments based on a learning progression can help teachers (a) to understand the waypoints of the learning progression, (b) to make design decisions, and (c) to make practical every-day decisions about learning.

First, we lay out foundational research and ideas for understanding students' growth in scientific argumentation. Following that, we describe (a) the development of an assessment that attempts to measure students' argumentation competency in the context of middle school science, followed by (b) an account of one critical aspect of the validation study of this assessment. The chapter concludes with a discussion of the challenges that are involved in developing assessments that align with a learning progression.

An Exemplary Context: Scientific Argumentation

In the past two decades, there has been an increasing emphasis on incorporating scientific argumentation into the practice of science education (e.g., Driver et al., 2000; Duschl & Osborne, 2002). Extensive research suggests that engaging in the process of argumentation supports students' understanding of scientific content knowledge (e.g., Mercer et al., 2004; Schwarz et al., 2000; Venville & Dawson, 2010; von Aufschnaiter et al., 2008; Zohar & Nemet, 2002). The

need for engaging students in argumentation is addressed explicitly in the national science standards in the US. Specifically, the Framework for K–12 Science Education (National Research Council, 2012), along with the Next Generation Science Standards (NGSS Lead States, 2013), specifies "engaging in argument from evidence" as one of the eight essential science practices in grades K-12.

Following the logic of Henderson et al. (2014), we consider argumentation as a competency demanding both

(i) knowledge, i.e., content knowledge, which is required to adjudicate multiple pieces of evidence, and

(ii) practice, i.e., epistemic knowledge to construct and critique various forms of reasoning.

Thus, with an increasing focus on incorporating argumentation into the teaching and learning of science, an emerging challenge is how to assess students' competency with scientific argumentation. For this work, we take advantage of previous research on how to assess scientific argumentation to provide the waypoints of the learning progression (Henderson, et al., 2014; Osborne et al., 2016). Without valid and reliable assessments, science educators will not be able to readily assess students' argumentation competency.

Toulmin's model of argument. Osborne and colleagues' initial hypothesis about how scientific argumentation is learned (Henderson et al., 2014; Osborne et al., 2016) is based on part on Toulmin's (1958) model of practical argument. In contrast to the logic of more traditional deductive arguments that start from first principles to logically deduce or prove a claim, practical

arguments instead begin with claims and seek justification and evidence to support those claims. This justificatory view of argumentation is more in line with how argument is used in actual scientific practice – scientists begin with a claim (i.e., a hypothesis) and test its veracity against empirical data.

The Toulmin (1958) model begins with a claim as a "conclusion whose merits we are seeking to establish" (p. 90). It follows that in seeking to establish the merit of a claim, one turns to evidence that may lend support. In turn, the relation between the evidence and the claim is provided by a warrant that forms the substance of the justification for the claim. The Toulmin conceptions of claims, warrants, and evidence play a key role in the development of the Osborne et al. construct, just as it has formed the basis of many schemas used in research for analyzing student discourse (Cavagnetto, et al., 2010; Jimenez-Aleixandre et al., 2000; Osborne et al., 2004; Zohar & Nemet, 2002).

The role of critique. In addition to Toulmin's (1958) focus on the justificatory role of argument, the process of critique is also incorporated into the Osborne et al. (2016) construct map for scientific argumentation. As Ford (2008) has argued, the goal of scientific reasoning is knowledge construction, achieved by a dialectic consisting of both construction and critique. More specifically, argument is used to justify the validity of explanatory hypotheses, experimental designs, and interpretations of a given data set. Critique is thereby essential to identifying flaws in such arguments. Without critique, adjudication of competing arguments is not possible.

The ability to engage in critique, however, requires a somewhat different competency than construction. For instance, rather than constructing a claim, critique requires the ability to identify what is the claim in any given argument, what elements constitute the data that is used to support the argument, and what kind of reason is used to relate the data to the claim. Such a competency is reliant on at least a tacit meta-knowledge of the features of an argument and the ability to distinguish its component elements. Taken a step further, critique requires the ability to construct a rebuttal that would explain why the reasoning in a given argument is flawed. Commonly, this may require the cognitive performance of comparing and contrasting the relative merits of two arguments or constructing an argument for why some evidence has higher epistemic validity than other evidence.

Accounting for cognitive load. Taken as a whole, therefore, a competency with scientific argumentation demands a complex orchestration of construction and critique of claims, warrants, and evidence in situations that require scientific knowledge to resolve. Unfortunately, opportunities to engage in argumentation rarely occur in traditional science classrooms in K-12 education (Banilower et al., 2012; Newton et al., 1999; Weiss et al., 2003). Thus, the multitude of information that must be processed to construct and/or critique a scientific argument, argumentation is both a novel and demanding task for many students that can –depending on the complexity of the argument – make substantial cognitive demands.

Given that students have little experience with scientific argumentation, any construct map for scientific argumentation should also account for limitations in human working (i.e., short-term) memory. This is because working on novel tasks places a high cognitive demand on

human working memory. These limitations in working memory are perhaps most famously documented by Miller (1956) who described how, across multiple cognitive experiments, the number of entities that could be simultaneously held in human working memory is seven +/- 2. These findings have served as a foundation upon which cognitive load theory (Chandler & Sweller, 1991; Paas et al., 2003; Sweller, 1994) has been built.

Pollock et al. (2002) summarize two basic sources of cognitive load. The first source – intrinsic cognitive load – "is determined by the extent to which various elements interact… An element is the information that can be processed by a particular learner as a single unit of working memory" (p. 62). In contrast, extraneous cognitive load "is generated by the manner in which information is presented to learners and is under the control of instructional designers" (p. 62).

All of these elements derived from the literature on the competency of scientific argumentation were used to develop the first of the BAS building blocks—the construct map of the entity to be assessed, as described in the next section.

Constructing the Assessment of Scientific Argumentation

The work reported here was part of a larger research project designed to examine how students' understanding of the structure of matter and competency in argumentation in the scientific context develop. The development of this scientific argumentation assessment is based

on a construct modeling approach (Brown & Wilson, 2011; Wilson, 2004, 2005, 2009, 2012, in press), known as the BEAR Assessment System (BAS), which consists of four building blocks: construct map, item design, outcome space, and measurement models. The BAS is a general approach and can be applied to both achievement test development (K-12 and higher education) as well as for attitude scales (Wilson, 2005, in press). How each of these four building blocks contributes to the development of the assessment will be explained in more detail below.

This section describes each of the four BAS building blocks (see Figure 1) in the development of the scientific argumentation assessment. The first building block describes the theory underpinning a proposed learning progression for scientific argumentation upon which our assessment is based. The BAS proposes a simple form for the learning progression, one that can be made up of a set of *construct maps*. A construct map consists of a series of increasingly sophisticated ways of understanding a certain idea or process, with a qualitative definition of each successive node in that succession—these nodes are called waypoints (Wilson, in press). The simplest form is where there is just a single ordered set of waypoints, but more complex patterns may be a better representation of some hypothesized structure (see Wilson (2009) for multiple examples of these more complicated structures). The second and third building blocks illustrate the process of designing assessment items and scoring rubrics focused on probing the various waypoints of the hypothesized construct map. The fourth building block describes the measurement model of choice.

<Insert Figure 1 about here>

Building Block One: A Scientific Argumentation Construct Map

As noted above, Osborne et al. (2016) use cognitive load theory as a lens through which the competency of scientific argumentation can be operationalized as a single construct. More specifically, given that Toulmin (1958) resolves argumentation into fundamental components such as claim, warrant, and evidence, the process of constructing or critiquing arguments can be viewed as an orchestration of various combinations of these elements of argument. It follows that for tasks that increase the number of elements that must be processed, the intrinsic cognitive load on the working memory increases, thereby making it more difficult to demonstrate argumentative competency.

Hence conceptualizing the elements of argument as a source of intrinsic cognitive load guided Osborne et al. (2016) as they hypothesized certain levels of sophistication of argumentation to be more difficult than others in their proposed learning progression for scientific argumentation. Consequently, Osborne et al.'s (2016) hypothetical learning progression for scientific argumentation (see Figure 2[3]) consists of three broad sets of waypoints of argumentation differentiated by intrinsic cognitive load (and an initial waypoint with no evidence of argumentation ability), where each higher waypoint is seen as requiring more connections to be made between claims and pieces of evidence.

<Insert Figure 2 about here>

[3] Note that the numbering of the waypoints in Figure 2 has been changed from that given in Osborne et al (2016) to make the scoring clearer—the Waypoints "0a" etc. have been re-labelled "1a" etc., and similarly for Waypoints 1a through 1d and 2a through 2d.

The lowest waypoint of the construct map (at the bottom of Figure 2) is the waypoint where no part of an argument is identifiable in the student's response, and this is labelled "0".

The next set of waypoints are prefixed with the number "1"—this denotes that assessment items probing these stages do not ask for explicit connections between claim and evidence to be made or recognized. These connections – warrants under the Toulmin model – are not specifically asked for. Rather the question is whether the student can identify the individual elements of the argument. and demonstrate competency with identification/critique of a claim and/or evidence without making a logical connection between them. Thus, the 1 prefix for these waypoints indicates that the response is unitary and hence does not show any degrees of coordination. Construction of an explicit claim (Waypoint 1a) is hypothesized to be the most basic demonstration of argumentation competency, as doing so does not technically require any additional knowledge of the features of an argument. Indeed, it is possible for a student to advance a claim without relating it to either evidence or a warrant that might offer justification. Waypoint 2b is the matching ability to recognize a claim.

Items that do indeed require explication of warrants mark the transition from the set of Waypoints 1 to the set of Waypoints 2 in the construct map. These intermediate waypoints are prefixed with the number "2" to denote that they do involve coordination between two elements of an argument – i.e., students need to make at least one explicit logical connection between claim and evidence with a warrant. Thus, Waypoint 2a builds on the set of Waypoints 1 of the construct map, as it requires understanding of not only what constitutes a claim or a piece of

evidence, but also how to construct or critique a relationship between claim and evidence. Assessments probing Waypoint 2a explicitly ask for such a warrant to be constructed. Specifically, at Waypoint 2a, students are able to construct an explicit warrant that connects claim to evidence. Alternatively, Waypoint 2b requires the ability to identify the warrant that another person provided. Waypoints 2c and 2d are concerned with a student's competency in constructing a complete argument or providing an alternative counter argument to rebut a different claim.

The most advanced waypoints of the construct map are prefixed with the number "3" to denote coordination between more than two parts of an argument. At these waypoints, the student not only has to construct their own argument but identify the elements in another argument and compare and contrast the two. Waypoint 3 items require students to explicate two or more warrants. In particular, assessments probing Waypoint 3a require students to critique another person's argument (e.g., explain why an argument is flawed). Waypoints 3b and 3c require students to evaluate the merits of two competing arguments. At Waypoint 3b, students are able to justify the value of one argument and yet do not provide a warrant to show why the other argument is less (or more) superior. Alternatively, at Waypoint 3c, students are not only able to evaluate two competing arguments but also able to explicitly argue why one argument is considered stronger than the other. Last, Waypoint 3d requires the ability to construct a counter claim. This is the most proficient waypoint since it requires all of the capabilities. Specifically, it requires the individual to construct an alternative explanatory hypothesis, compare it with the existing argument, and evaluate why it is superior. The latter requires a reason for why it offers greater explanatory coherence and to refute alternative theories or arguments. Although the

descriptions of the successive waypoints in the paragraphs above have not made explicit references to the relative cognitive loads involved at each waypoint, a clear succession of increasing pieces of the Toulmin argument framework is evident.

Building Block Two: The Items Design

To investigate the validity of the hypothesized construct map described in Building Block One, multiple paper and pencil assessment items were designed to elicit student responses commensurate with the various waypoints of our construct map. Each argumentation assessment in this study began by establishing a scientific scenario relevant to how matter is structured. Students were then presented with a set of items related to the specific scenario. Within each scenario, multiple items were designed to span a range of argumentation skills included in the hypothesized construct map (Figure 2). Argumentation items sharing a common scenario will be referred to as item bundles, following Rosenbaum's (1988) use of the term to denote a subset of items that share a common stimulus. Figure 3 presents an example of an argumentation item bundle called "Sugar in the Water." The choice of specific domain knowledge that is involved in the argumentation items is, of course, a crucial decision. The choice in the examples illustrated here flowed from a decision to investigate the Structure of Matter. See Black et al. (2011)—for a discussion of the Structure of Matter learning progression that the items are based upon.

<Insert Figure 3 about here>

Note that the design of these items was informed by peer review from both educational researchers and teachers. Additionally, cognitive think-aloud interviews (Wilson, 2005) with a group of grade eight students in a local district provided a gauge on whether each item elicited the kinds of cognitive response processes that were intended when the items were originally written. Utilizing an iterative process of item refinement when items did not perform as expected, the resulting modifications were again subjected to additional peer review and think-aloud interviews. It is important to note that the quality of the items is also very important in not only in promoting teacher usage of the assessments and the student outcomes from those assessments, both in terms of teachers understanding how the items reflect the waypoints of the construct map, and how the findings from assessments using those waypoints can be useful to them, but also in promoting teacher acceptance of the assessment.

Building Block Three: The Outcome Space

Given that the preponderance of argumentation items required open-ended written responses, it was important to develop scoring rubrics that facilitated reliable human scoring of written excerpts. More specifically, design of the scoring rubrics was initially based on mock student responses that were hypothetically anticipated for each item. Researchers then used these preliminary rubrics to assign scores on pilot samples of actual student responses for each item. Working in groups of no less than three researchers, each researcher independently scored the same sample of pilot tests. For consistency, researchers focused on scoring only one specific part of one specific item at a time. Each researcher then shared their scores on a specific part of a specific item with the rest of the group. If researchers did not reach a unanimous scoring

decisions for each pilot test, a discussion would ensue, and the scoring rubrics were modified based on this peer review. This moderation process was repeated until the revised scoring rubrics indeed yielded unanimous scores for each pilot test. Along the way, empirical samples of student responses were added to the rubrics to complement the hypothetical mock responses. The materials developed to help the researchers make these judgements are retained as training materials for others who will use these items, as well as the arrangements for the training procedures—these are all considered a part of the outcome space.

In later iterations of this developmental work (Morell et al. 2020) these recorded open-ended responses were mined to use as sources of options for multiple choice and technology-enhanced items (TEIs). This results, eventually in a two-part item bank—an open-ended portion, which can be used by teachers as rich sources of stimuli for formative assessment conversations, and a selected response portion which can be used for teachers to get quick feedback from their students. The remainder of this account focusses on the former portion, the open-ended items.

As described previously, each argumentation item is designed to assess one of the three main waypoints (i.e., Waypoints 1, 2, or 3) in the hypothesized scientific argumentation construct map. Note that partial credit was given to responses where reasoning was largely, yet not entirely correct. In the other words, responses with full credit are regarded as fully demonstrating the argumentation construct waypoint that the item was designed to assess. Alternatively, responses receiving partial credit are considered as demonstrating the construct waypoint in only a partial and incomplete fashion. This logic applies to the scoring of all items.

As an example, Table 1 presents the final scoring guide of the item 5b in the "Sugar in the Water" item bundle (see Figure 3). This item asks students to identify a piece of evidence that challenges the claim made by a hypothetical student, Paul, in the prompt, and provides a justification for how the selected piece of evidence challenges Paul's claim. As described previously, each argumentation item is designed to assess one of the three main waypoints, i.e., Waypoint 1, 2, or 3, in the hypothesized scientific argumentation learning progression. This particular item was developed to assess Waypoint 2 in the hypothesized argumentation learning progression, which is to identify evidence that contradicts a given claim and build an alternative argument about why that claim may be flawed. Note that partial credit was given to responses, where reasoning is largely yet not entirely correct. Responses with full credit, i.e., code 2 in the current example, are regarded as fully demonstrating the proficiency waypoint an item is designed to assess, i.e., Waypoint 2 in this example. Alternatively, responses with partial credit, i.e., code 1 in the current example, are considered as demonstrating the proficiency waypoint partially. This logic applies to the scoring of all items.

Table 1. The scoring guide of the item 5b in the "Sugar in the Water" item bundle

Code	Descriptions	Mock Student Responses	Empirical Student Responses
2	Correct answer with appropriate reasoning. *A correct piece of counter-evidence is identified and reasoning is provided as to how it counters Paul's claim.*	"Number [anything but 2] challenges Paul's argument, as he claims the sugar disappears for good, but in fact [any piece of evidence besides #2] suggests that something remains."	[PICKED 5] "Paul said the sugar was gone for good but due to the law of conservation, the sugar is still there just that it has dissolved." (Note that this response tied this evidence back to Paul's claim)
1	Reasoning is largely correct, i.e., correct answer with inadequate reasoning or good reasoning that is used to justify an incorrect answer. *A correct piece of counter-evidence is identified, but reasoning is not provided as to how it counters Paul's claim.* <u>*Alternatively,*</u> *A correct piece of counter-evidence is not identified, but at least reasoning is provided as to how the student thinks it counters Paul's claim.*	"Number [anything but 2] challenges Paul's argument, as he claims the sugar disappears for good, but in fact [any reasoning that does not justify why the selected evidence does not support Paul]." <u>*Alternatively,*</u> "Number [2] challenges Paul's argument, as he claims the sugar disappears for good, but in fact [any reasoning that tries to justify why the selected evidence does not support Paul]."	[PICKED 5] "Because you cannot destroy or make new matter." (Note that this response did not tie this back to Paul's claim) [PICKED 1] "Followed by what Paul said, the sugar breaks up and mixed with the water." [PICKED 4] "Because number 4 said that the sugar is still there."
0	Answer fails to address the question. *The student not only fails to identify a piece of evidence that challenges Paul's argument, but no justification is provided for why in fact they believe the piece of they chose counters Paul.*	"Number [2] challenges Paul's argument, as he claims the sugar disappears for good, but in fact [any reasoning that does not justify why the selected evidence does not support Paul]."	[PICKED 2] "Because Paul says the sugar breaks apart and disappears, not stirring the water." [PICKED 2] "Tells more details, and makes his statement sort of clear."
B	No Response		

Building Block Four: The Measurement Model

The fourth building block is represented by measurement models, which allow one to mediate the information from a sample of students back to the original intent in terms of the estimated parameters of the model (both the student and the item parameters). Through a measurement model, scored student responses were calibrated and related to the cognitive model (Wilson, 2005). There have been many measurement models proposed and used in the field of educational and psychological testing. In describing the philosophy of the construct modeling approach (i.e., the BAS approach), Wilson (2005) focused on the special role the Rasch models (Rasch, 1960) play in understanding a construct of interest. In this study, both exploratory factor analysis and Rasch partial credit model were applied to the investigation of the function of the assessment of interest, which are illustrated in detail in the subsequent section.

Among the various aspects of validity evidence, this study focuses on the evidentiary strand of internal structure validity for the assessment of scientific argumentation– the focus of the study. According to the *Standards for Educational and Psychological Testing*, investigation of the empirical manifestations of the internal structure of a test can provide empirical evidence about "the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 2014). The assessment investigated here was designed to measure a unidimensional construct of argumentation. Whether empirical evidence supports this assumption of unidimensionality is, therefore, the first question of interest.

In addition, as each item was designed to target a specific waypoint on the hypothesized construct map of argumentation, the extent to which the empirical pattern of item parameters supports this item design feature constitutes a second focus of interest.

Methods

Data

Twenty items were developed to assess students' argumentation competency, each specifically designed with one waypoint on the argumentation construct map in mind; the items were embedded in four scenarios contextualized in the domain-specific area of Structure of Matter for grade eight. Based on our expectations of the approximate ability distribution of the students, 11 items were designed to measure Waypoint 1, seven items were designed to measure Waypoint 2, and two items were designed to measure Waypoint 3 of the hypothesized argumentation construct map. The 20 argumentation items were administered in paper and pencil format. The students were given what was considered ample time to respond to the 20 items—most were finished well before the available time, though a few were still working at the end. The data used in this study were collected from 347 grade eight students in a local school district in the US. Non-responses were coded as inaccurate responses. Information about five students was incomplete. The remaining sample consists of 188 girls (54.97%) and 154 boys (45.03%). According to the results of the 2011 California Standards Test in Science (California Department of Education, 2013), 195 of these students (57.02%) were classified into the "Advanced"

category, 61 students (17.84%) were in the "Proficient" category, 57 students (16.67%) were in the "Basic" category, 20 students (5.85%) were in the "Below Basic" category, and nine students (2.63%) were in the "Far Below Basic" category[4].

Data Analysis

To assist the investigation into the dimensionality of the assessment, exploratory factor analysis (EFA) was performed. In a study of dimensionality assessment of polytomous items, Timmerman and Lorenzo-Seva (2011) suggested that the parallel analysis with minimum rank factor analysis (PA-MRFA) performed consistently well at identifying the number of major factors when minor factors were present in the simulated data. In the current study, the PA-MRFA procedure described by Timmerman and Lorenzo-Seva (2011) was performed with the software FACTOR (Lorenzo-Seva & Ferrando, 2006). For a detailed account of the PA-MRFA procedure implemented, please refer to Yao (2013). To evaluate the number of factors, the scree plot (Cattell, 1966) was also examined.

The Rasch partial credit model (Masters, 1982) was used as a confirmatory model to evaluate the measurement properties of the instrument as a whole and of each individual item. Under the Rasch partial credit model, the probability of student responses is formulated as follows:

[4] This data was used as part of an earlier analysis which focused on the relationship between the argumentation dimension and the content dimension (Yao et al., 2015). *** NO REFEERNCE FOUND

$$P_{pix}(\theta)=\frac{\exp\left[\sum_{j=0}^{x}(\theta_p-\delta_i-\tau_{ij})\right]}{\sum_{r=0}^{m_i}\left[\exp\sum_{j=0}^{r}(\theta_p-\delta_i-\tau_{ij})\right]},x=0,...,m_i\text{, where }\sum_{j=0}^{0}(\theta_p-\delta_i-\tau_{ij})\equiv 0\ ,$$

where θ_p stands for the latent ability of student p . δ_i represents the overall difficulty of item i, and τ_{ij} refers to the deviation step parameter associated with category j of item i.

The partial credit model was estimated with the software *ConQuest* (Adams et al., 2020) using the marginal maximum likelihood estimator (Gauss-Hermite Quadrature with 15 nodes), where the ability parameter θ_p is assumed to be normally distributed as $N(\mu_\theta,\sigma_\theta^2)$. Item parameters, δ_i and τ_{ij}, are assumed to be fixed parameters. For model identification, the mean of the ability distribution is fixed as 0 and $\sum_j \tau_{ij}=0$ for each item i. In the Conquest outputs used below, the reported Wright maps utilize a calculated value called a "Thurstonian threshold" (Adams et al., 2020). For the cases shown, this value represents the location at which the probability of observing the highest category (in terms of the construct map waypoints shown in Figure 2) is 0.50.

Results

Unidimenionsality--Results from the Exploratory Factor Analysis

The EFA results first provide statistical evidence for whether the studied argumentation items measured one salient latent construct as intended. Figure 4 is the scree plot of the eigenvalues of the reduced correlation matrix (Timmerman & Lorenzo-Seva, 2011). The plot shows that a sharp bend occurs around the second eigenvalue. Although the first two factors both have eigenvalues larger than one, the eigenvalue of the first factor is about three times as large as the eigenvalue of the second factor and the eigenvalue of the second factor is only slightly larger than one. This evidence points to the presence of one major factor.

<Insert Figure 4 about here>

The presence of one salient factor is further observed in the result of parallel analysis of the data (Timmerman & Lorenzo-Seva, 2011). The description of this analysis and corresponding terminologies is beyond the scope of this chapter. Please refer to Timmerman and Lorenzo-Seva (2011) for details. As shown in Figure 5, only Factor 1 has the proportion of explained common variance from the empirical data larger than that from the randomly generated data. This result clearly points to a one-factor solution.

<Insert Figure 5 about here>

To evaluate the fit of the one-factor model, the value of the root mean square of residuals was also checked. The values of the root mean square residuals should be below 0.08, with lower values indicating better model fit (Hu & Bentler, 1999). The one-factor model has the value of

the root mean square of residuals equal to 0.08, which suggests reasonable fit. For estimates of factor loadings, please refer to Appendix A.

Results from the Rasch Analysis

The person separation reliability[5] was estimated as 0.81, and the expected a posteriori reliability was estimated as 0.82. This evidence suggests a reasonable reliability result for the test.

Item fit statistics. The weighted mean square statistic was used to evaluate the fit between the partial credit model and data on each item. The weighted mean square statistic has an expected value of one. Items with values of the weighted mean square statistic larger than one are those with observed variance greater than the expected and contribute less toward the overall estimation of the latent variables. Items that have a weighted mean square statistic outside the range of 0.75 and 1.33 and have the absolute value of the weighted t statistic larger than 1.96 should be considered as problematic (Wilson, 2005; p.129). All of the investigated argumentation items have weighted mean square statistics within the range of 0.75 and 1.33, which provides support for the use of the partial credit model in data calibration. This result should not be surprising, as the 20 items used in this study had already been calibrated once before in an earlier round of data used for item selection and improvement. and found to be reasonably well-fitting in that earlier analysis[6].

[5] The person separation reliability is the equivalent of Cronbach's alpha, following the translation of the raw scores (used in Cronbach's alpha) into estimates on the logit scale, based on the Rasch model calibration (Masters, 1982),
[6] The results of this earlier data collection and analysis were formative of the instrument, and hence were not published at that time.

Item discrimination. Students who demonstrate a higher proficiency waypoint on the latent construct should also tend to score higher on each item than their counterparts-this is evidence for the discrimination of the item. To check this consistency between each single item and the whole instrument, Wilson (2005) recommended examining the mean ability of students within each score category for a given item. Specifically, the mean ability of students within each score category is expected to *increase* as the score increases. All of the studied argumentation items met this expectation, except for item A2_5d. Figure 3 presents the prompt for this item and Table 2 presents the information about each score category of this item. As shown in the Table, the average ability estimate increases as the score increases for item A2_5d, except for the category of score 2. For this item, categories of score 1 and score 2 are responses that earned partial credits. According to the scoring guide for this item, responses as described in the category of score 2 (i.e., recognize uncertainty but DO NOT provide reasoning) got a higher overall estimate than responses as described in the category of score 1 (i.e., provide reasoning but DO NOT recognize uncertainty), because recognizing uncertainty in evidence is considered to be more sophisticated than providing reasoning to back a claim on the hypothetical argumentation progress map. The empirical evidence, however, shows that the average ability of the students that scored 2 (n=80) on this item is lower than the average ability of the students that scored 1 (n=82). The reason why item A2_5d did not show as displaying large misfit statistic is probably because there are only 80 students in the misbehaving response category (i.e., score 2). Nevertheless, the disordering in the average ability estimates of the students across response categories is indeed manifested on the weighted mean square statistic. Specifically, the weighted mean square statistic of item A2_5d is 1.21, which is close to the upper boundary of the acceptable range (i.e., 1.33). When the value of the weighted mean square statistic is larger than

one, it suggests that the observed residual variance is greater than the expected. This is the only argumentation item that assessed students' ability to recognize uncertainty in evidence. More items of this kind are needed in the future to investigate whether recognizing uncertainty is indeed more sophisticated than providing reasoning to back a claim as hypothesized.

Table 2. Information about each score category of Item A2_5d.

Score	Descriptions	Mean Ability (logit)	Number of Students
3	Recognize uncertainty AND provide reasoning. *The student recognizes that the evidence is not conclusive and makes a case for why this uncertainty does not allow them to side with either Laura or Mary*	0.38	28
2	Recognize uncertainty but DO NOT provide reasoning. *The student recognizes that the evidence is not conclusive but does not make a case for why this uncertainty does not allow them to side with either Laura or Mary.*	0.05	80
1	Provide reasoning but DO NOT recognize uncertainty. *Here students, failing to recognize the uncertainty in the evidence, conclusively side with either Laura or Mary, but at least they provide a plausible reason why they chose to side with Laura or Mary.*	0.30	82
0	DO NOT provide reasoning/ DO NOT recognize uncertainty. *The student not only fails to see the uncertainty in the evidence, but they are unable to correctly reason for the claim (i.e., side with either Laura or Mary) in which they feel the evidence is indeed sufficient.*	-0.27	157

The Wright map. A Wright map displays the student ability distribution and item locations on the same logit scale. When we produced the full set of item locations for the Wright map for these items, including those related to the partially-correct responses (e.g., partially correct at Waypoint 2), the pattern of results was not clear—the conclusion from this analysis was that these partially-correct categorizations were not as consistent across items as we would

desire them to be. Hence, it was decided to focus attention on only the "full credit" categorizations. With this focus, we found that the pattern of results was much clearer.

Thus Figure 6 presents the Wright map based on the *full-credit response category* for each item. The first column on the left of Figure 6 shows values of the logit scale. The second column on the left represents the distribution of the estimated student ability from the most able at the top to least able at the bottom of the map, with mean equal to 0 (by convention) and variance equal to 0.39. Each "x" in this column represents approximately 2.1 students and the location of the "x"s (with respect to the logit scale) indicates the waypoint of their ability estimate. The third to the fifth columns present the item Thurstonian thresholds (Wilson, 2005), which are arranged in columns 3, 4 and 5 depending on the construct waypoint an item step was designed to measure. The j^{th} Thurstonian threshold represents the point at which the probability of scoring below j is equal to the probability of scoring j and above. If the ability estimate of a student and the full-credit Thurstone threshold of an item are located at the same position on the map, the probability of this student obtaining full credit on the item is 0.5.

<Insert Figure 6 about here>

Checking the internal structure validity evidence.

Each item is hypothesized to measure one specific construct waypoint. Items measuring a lower construct waypoint are expected to have lower difficulty waypoints, compared to items measuring a higher construct waypoint. Ideally, item thresholds tapping the same construct

waypoint will be located at similar positions on the Wright map. When the locations of the thresholds for a certain construct waypoint fall close together, we can specify this by a partitioning of the scale on the Wright map; this is called "banding." Once a set of "bands" representative of construct waypoints are drawn, the empirical ordering of these "bands" can then be compared to the theoretical ordering of the waypoints as hypothesized in the construct map. Examining Figure 6, we can readily see that the locations of the thresholds for a given construct waypoint do not fall closely together on the Wright map, and thus a set of "bands" representative of construct waypoints cannot readily be drawn in Figure 6.

Given that the empirical pattern of the locations of the item thresholds on the Wright map does not initially support the expected ordinal structure of the construct waypoints, one plausible step is to identify items with threshold locations that do not conform to the expectation and investigate possible causes for their unexpected functions.

One reason for this lack of clear structure may be that some items did not perform as well as the others on measuring the common latent argumentation competency, and thereby brought "noise" into the Wright map in Figure 6. Five of the items that do not conform to the banding are among those that have small factor loadings (please see Appendix A): items A1_3b, A3_14c, A3_14bb, A4_7a, and A4_7b. To focus on items that more clearly are centered on the argumentation construct, it was decided to drop these five items from the remaining analysis. Figure 7 presents the banding of the Wright map based on the smaller set of items, where these five items with small factor loadings are excluded. Although not perfect, an ordinal structure

among locations of item thresholds associated with different construct waypoints more readily emerges in Figure 7.

<Insert Figure 7 about here>

Waypoint-1 item*s*. First, all Waypoint-1 item thresholds fall in the range between -0.47 logit and 0.41 logit on the Wright map, except for the thresholds of the following three items: items A1_3c, A1_3a, and A2_5c. Getting full credit on these three items appears to be much more difficult than getting full credit on the other Waypoint-1 items. An examination of item A1_3c (see Figure 8) reveals a potential influence of language on the difficulty waypoint of the item. Bundle A1 establishes a scenario with three paragraphs: the first paragraph presents the issue of California being short of fresh water, and each of the remaining two paragraphs presents a method of desalination to obtain fresh water from the oceans. Hypothetically, items A1_3c and A1_3d (see Figure 8) are expected to have similar difficulty waypoints, since they ask students to identify an argument against each of the two desalination methods. However, item A1_3c appears to be much more difficult than item A1_3d in the empirical data. One possible reason is the ambiguity of the language. Item A1_3d explicitly specifies the desalination method students should identify an argument against, which helps students to locate the critical information in the prompt. In contrast, item A1_3c does not clearly specify what students should look for.

<Insert Figure 8 about here>

An examination of item A1_3a provides another implication for the assessment of scientific reasoning. This item asks students to identify the main claim that is being made in the scenario of Bundle A1. According to the scoring guide, students received full credit only if their responses referred to both (a) water shortage and (b) how the water shortage is to be addressed. 55% of the students mentioned only one of the two pieces in their responses and, thus, were awarded partial credit. These pieces of evidence suggest that, identifying parts of a claim may be easy as hypothesized yet it is difficult to identify *all* pieces in each claim. This finding calls for reflection on two general assessment issues: (a) what waypoint of sophistication of student responses should be sufficient to be considered as demonstrating performance at a particular construct map waypoint, and (b) how item design can be improved to prompt students to demonstrate their understanding and proficiency waypoint to the fullest.

Osborne and colleagues (2016) have provided a discussion about why item A2_5c was surprisingly difficult. As shown in Figure 3, item A2_5c asks students to identify a piece of evidence that best supports the claim the hypothetical student Mary made and to provide an appropriate explanation. Item A2_5a is identical to item A2_5c, except that item A2_5a is about identifying evidence for the claim the other student Paul made. By design, these two items were both regarded as requiring Waypoint-1 competency, as they only ask for identification of evidence supporting a given claim. However, the empirical evidence suggests that it is much more difficult to get full credit on item A2_5c than on item A2_5a. An examination of the question shows that, only one of the five pieces of evidence given in the prompt could support Paul's claim, while four pieces of evidence could support Mary's claim. Osborne and colleagues (2016, p.11) argued that "in the case of Paul, the question merely demanded an assessment of

which single element was the most significant piece [of] evidence for his claim. However, in the case of Mary [,] the question demanded the assessment of the relative significance of multiple pieces of evidence in relation to her claim." As described previously, the competency for scientific argumentation is conceptualized as a complex orchestration of various elements of argument (e.g., claim, evidence, etc.) in the current construct map. Although item A2_5c and item A2_5a were initially designed to assess the same type of coordination (i.e., identifying evidence supporting a claim), item A2_5c imposes a higher waypoint of the intrinsic cognitive load as the pieces of information students need to process is larger.

Waypoint-2 items. As to Waypoint-2 item thresholds, the four Waypoint-2 item thresholds cluster between the range of 0.70 logit and 1.30 logit and are located higher on the Wright map than most of the Waypoint-1 item thresholds. This pattern matches with the expectation that item thresholds targeting the same construct waypoint have similar difficulty waypoints and item thresholds of a higher construct waypoint should be located higher on the Wright map than those of a lower construct waypoint. Development of more Waypoint-2 items will be needed in the future to obtain more evidence about this pattern.

Waypoint-3 items. For Waypoint 3, item A1_3e is surprisingly easy. Also embedded in Bundle A1, item A1_3e asks students to decide which desalination method is the best and explain their choice. It was hypothesized that this item assessed a student's ability to evaluate and compare competing arguments. However, according to the scoring guide, students get full credit if they successfully give a claim (i.e., choose one of the two desalination methods) and a reason provided in the prompt. Cognitively, succeeding on this item thus only requires the ability

to identify a claim and a relevant piece of supporting evidence. We would therefore re-classify item A1_3e as Waypoint 1. This item is a good example of how the way an item is written may affect the cognitive demand it requires. Also, more Waypoint-3 items will be needed for a further examination of the "banding" and establishing the typical difficulty locations of the Waypoint-3 items.

Discussion of Results

The overall picture of alignment, following the investigation of the non-aligned items above, as shown in Figure 7 (on the far right-hand side) allows one to see a potential banding for the scientific argumentation construct. The banding was based on an application of the "construct mapping" standard setting technique. (Draney & Wilson, 2011; Wilson & Draney, 2002). This is shown by the horizontal lines through the Wright map separating the scale into four segments, each labelled on the right-hand side by the relevant Waypoint of the construct map (Figure 2). Bearing in mind that the location of a student at the same point as an item means that a student is tending to succeed at about 50% of the time, these bands can now be interpreted as indicating the regions of the scale where a student is learning at a certain Waypoint of the construct map. Thus, for example, a student at, say, 0.0 logits, is currently learning about claims and data (i.e., Waypoint 1 on the construct map), while a student at, say, 1.0 logits, is learning about arguments and rebuttals (i.e., Waypoint 2). In contrast, students at, say, -1.0 logits, are not yet at the point where they can consistently make a claim. And, if any student were located at 2.0 logits, then they would be working on articulating different arguments (but no students were consistently doing that!). This banding shows the way that the foundational effort to create the construct map

waypoints, can then be wound around the BAS to allow one to interpret the quantitative estimates on the logit scale.

In summary, we chose to use this example of how to calibrate a learning progression not because the results are clean and perfect, but as discussed above, because the results show the somewhat consistent, and somewhat inconsistent patterns that we typically find as we start to explore applications of this new paradigm. What one would need to do next is to *iterate* the procedures carried out above, with new items (as well as the old ones), and seek to conform that the patterns of results can be replicated. One important part of this strategy is to examine the items that did not conform to the overall pattern (such as the "non-aligned" items mentioned above, as well as the five items that did not load well on the construct) and seek to understand what it was about the items that brought this about.

Reporting Results to Teachers.

It is crucial for the success of the project that the assessment results be reported to teachers in ways that allow them to make efficient use of the information. The work to develop these is currently underway and has been partially reported in Morell et al., 2019; Morell et al., 2021; Wilson et al., 2019, and are currently under submission to appropriate journals. Examples of the reports we have developed are shown and discussed in this subsection.

First, Figure 9 shows a display of results form a (fictitious) set of students—this is called the "Group Report.". The waypoints of the construct are shown advancing from left to right

across the page, from Waypoint 0 to Waypoint 3. Students are shown down the page, ordered from lowest performers at the bottom, to highest at the top. Each student has a separate row, and his/her location is indicated by the black dot. Stretching to the left and right of the student are 67% confidence intervals for that student, expressed as the thin black bars. This report can give teacher several useful pieces of information, that is:

(a) the range of the students, in this case from Waypoint 0 ("ARG 0" in Figure 9) to Waypoint 3 ("ARG 3)";
(b) a quick indication of the relatively more frequent Waypoints in the class (in this case Waypoints 1 and 2);
(c) identification of the students at the top (in this case 3 students), and at the bottom (in this case 2 students);
(d) some indications of uncertainty about specific student locations (for example, note that the two students at the bottom have relatively wide, probably associated with some missing data).

In our discussions with teachers, they have indicated that the Group Report (as shown in Figure 9) would be most useful in grouping students for subsequent instruction (Morell et al., 2021). The information in this graph is also augmented by tables, which show all the above information in explicit detail.

<Insert Figure 9 about here>

Teachers can also see, for example, a report on an individual student, as shown in Figure 10—this is called the "Individual Report." Here the layout and graphical features are the same as for Figure 9, but the focus is on just one student. In this case, the student is performing solidly in Waypoint 2, with sufficient information that the student is beyond Waypoint 1 and not yet at Waypoint 3 (as indicated by the observation that the error bars do not overlap with those regions of the graph). Thus, a teacher observing this performance, could interpret that the student was making reasonable progress in constructing an argument. Further information is also available in the "Scores Report," which shows each student's responses on each item that was attempted. This can give further depth to a teacher's interpretation of a student's performance.

<Insert Figure 10 about here>

Conclusion

With the formal inclusion of argumentation as an essential practice in science classrooms in national standards (e.g., NGSS Lead States, 2013), how to evaluate a student's scientific argumentation proficiency becomes a necessity. And, without valid argumentation assessments, science educators will not be able to diagnose their students' learning needs readily nor provide appropriate instructional interventions. Consequently, argumentation may not be adequately incorporated into science classrooms as desired. This study presented the development of an assessment for learning progression based on scientific argumentation for middle school students and illustrated the procedure of assessment validation. Specifically, (i) we used the scholarship

of Osborne et al (2016) as a basis for a progression as a sequence of waypoints of student sophistication in argumentation; (ii) we applied the BEAR Assessment System (BAS) to develop ways to assess the waypoints in this progression; (iii) we developed reports intended to help teachers use the information from the assessments for instructional design and also within the activities in a classroom; and (iv) we used both exploratory factor analysis and Rasch modeling approach to study whether the empirical information from students taking the assessments is consistent with the hypothesized progression. The process of developing assessments that align to a learning progression shaped as a set of construct maps, has been described in several other contexts. Recent examples illustrating this can be found in Dray et al. (2019), Junpeng et al. (2018), Metz et al. (2019), Morrell et al. (2017), Siddiq et al. (2017), and Wilson (2018).

Findings of the Internal Structure Validity Evidence

In the final sections, we discuss findings and challenges of the current research. Comparing the empirical pattern of the item parameter estimates to the theoretical structure of the progression waypoints, the empirical pattern of item thresholds on the original Wright map did not support the expected structure of the learning progression waypoints. As Wilson (2005) has discussed, negative evidence for internal structure validity could originate from at least three sources: (1) the theory of the construct (i.e., learning progression in the current case) which was inaccurate in some way, (2) the items did not work as intended, or (3) the outcome space was not developed appropriately. In this study, we traced the sources of the problems through the examination of the result of exploratory factor analysis and through the content analysis of the items. Items that did not conform to the expected structure were found to either measure the

argumentation construct poorly (i.e., have small factor loadings in the exploratory factor analysis) or measure nuisance variables in addition to the argumentation construct of interest.

Except for these erratic results in the initial development of items, the remaining items in this pilot study were found to have locations of thresholds that conformed to the expected pattern: (a) item thresholds of a lower progression waypoint have lower difficulty locations than item thresholds of a higher progression waypoint, and (b) item thresholds of the same progression waypoint locate at similar positions on the Wright map. This we see as supporting the Osborne et al. learning progression.

Challenges of Developing Learning-Progression-Based Assessments

Challenge 1: Designing items capable of eliciting student performance indicative of progression waypoints. To accomplish the goal of diagnosing where a student stands in the path of a learning progression, it is fundamental to develop a set of items that are capable of eliciting student performance indicative of progression waypoints on the hypothetical learning progression. In this study, we attempted to assess a student's scientific argumentation competency with items in the paper and pencil format. We found that it was most challenging to develop items that could elicit student performance indicative of high-level competency. For example, item "A1_3e" was originally expected to assess a student's ability to evaluate and compare competing arguments, which was regarded as the Waypoint-3 performance. However, it was found that succeeding on this item requires only the ability to identify a claim and a relevant piece of supporting evidence, which was essentially only Waypoint-1 performance. That is, this

item was not capable of differentiating students at Waypoint 3 from students at lower locations as expected and appeared as a surprisingly easy item in the analysis output. This example illustrates how the way an item is written may affect the cognitive demand it requires and shows the importance of evaluating students' cognitive or test taking strategies in the process of responding to a specific item. Thus, although it is not too hard to write very demanding open-form items, it can be difficult to write items that assess the specific higher order waypoints of a competency

Challenge 2: Linking student performance on an item to a specific progression waypoint. Wilson (2005) has suggested that *"*Although qualitative levels [on a learning progression] are definable, we assume that the respondents can be at any point in between--- that is, the underlying construct is continuous*."* (pp.6-7). Based on the hypothesis that a learning progression is essentially a latent continuum, variation in student performance *within* a learning progression waypoint should be expected. Following this idea, in this study student responses demonstrating partial understanding of a Waypoint, and which could not be mapped to a specific lower progression Waypoint, were regarded as incomplete or imperfect performances at that Waypoint, and hence given an intermediate score. However, the Wright map results from this tactic were not conclusive. Instead, it was found that only the results from the full-credit scores gave a clear pattern of results. Our conclusion from that evidence is that we should not code out partial success at a Waypoint but stick to full credit in the development of future items.

Challenge 3: Handling unexpected variables that affect student performance on an item. As discussed previously, an assessment is never strictly unidimensional. While some variables,

such as the use of language in item design, may be regarded as "nuisance" variables, other variables may systematically influence student success on items. For example, as discussed in detail above (in the section *Checking the internal structure validity evidence*), items "A2_5a" and "A2_5c" were both initially expected to assess Waypoint-1 competency. However, they differed significantly in terms of cognitive demand due to the specifics of the item design beyond the amount of detail provided in the description of Waypoint 1 of the Argumentation Construct Map. In this case , the item design issue had to do with the nature of the options in a multiple-choice item—most likely this would never rise to the importance of being a specific part of a Waypoint descriptor but is nevertheless important in understanding the relationship between the item design and its difficulty.

A Broader Perspective.

Our argument has been that a rigorous and systematic effort is needed to investigate the alignment of a learning progression with empirical data from an assessment. The question remains, however, about whether the alignment illustrated in Figure 7 constitutes useful evidence for the learning progression hypothesized in Figure 2. Specifically, what we have found is that over half of the items in this pilot study have locations of thresholds that conform to the expected pattern: (a) item thresholds of a lower progression waypoint have lower difficulty locations than item thresholds of a higher progression waypoint, and (b) item thresholds of the same progression waypoint locate at similar positions on the Wright map. In fact, we see this finding as a critically important finding, pointing towards a confirmation of a learning progression in the area of scientific argumentation. This finding needs the usual scientific support in the shape of

further data from new samples (which we are in the process of gathering; Morell et al., 2020; Wilson et al., 2019). We expect that there will be adjustments to the progression as we add new data and studies, but we are also expecting that the major outline of the learning progression will prove to be resistant to changes in assessment formats, such as differences between selected response and constructed response items (Wilson, Morell, Osborne, Dozier & Suksiri, 2019).

As we move forward with this research, we will also engage directly with the teacher-use of the construct map and the assessment results (Morell et al., 2021), which will add to our understanding of implications of the development of this learning progression and its value for improving the quality of information that we can provide about student learning and progression to teachers. Looking beyond its application in science education, it would be interesting to try the learning progression out on other educational areas: in particular, argumentation can be important in mathematics education, but it is not clear that the same structure will be generated by the sorts of argumentation tasks that mathematics teachers ask of their students. As Hattie has argued, high quality feedback is one of the most effective tools for improving teaching and learning (Hattie and Timperley, 2007; Hattie, 2008). We see our work as a contribution to this goal.

References

Adams, R. J., Wu, M. L., Cloney, D., & Wilson, M. R. (2020). *ACER ConQuest: Generalised Item Response Modelling Software, Version 5* [Computer software]. Australian Council for Educational Research.

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Banilower, E. R., Smith, P. S., Weiss, I. R., Malzahn, K. M., Campbell, K. M., and Weis, A. M. (2013). *Report of the 2012 National Survey of Science and Mathematics Education*. Horizon Research.

Black, P., Wilson, M., & Yao, S.-Y. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research and Perspectives*, *9*(2-3), 71–123.

Brown, N. J. S., & Wilson, M. (2011). A model of cognition: The missing cornerstone of assessment. *Educational Psychology Review*, *23*(2), 221–234.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245–276.

Cavagnetto, A., Hand, B. M., & Norton-Meier, L. (2010). The nature of elementary student science discourse in the context of the science writing heuristic approach. International *Journal of Science Education, 32*, 427–449.

Center for Continuous Instructional Improvement (CCII). (2009). *Report of the CCII Panel on Learning Progressions in Science*. CPRE Research Report, Columbia University.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*(4), 293–332.

Draney, K., & Wilson, M. (2011). Understanding Rasch measurement: Selecting cut scores with a composite of item types: The Construct Mapping procedure. *Journal of Applied Measurement*, *12*(3), 298-309.

Dray, A. J., Brown, N. J. S., Diakow, R., Lee, Y., & Wilson, M. (2019). A construct modeling approach to the assessment of reading comprehension for adolescent readers. *Reading Psychology*, 40(2), 191-241. DOI: 10.1080/02702711.2019.1614125

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84*, 287–312.

Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. *Studies in Science Education, 38*(1), 39–72.

Ford, M. (2008). Disciplinary authority and accountability in scientific practice and learning. *Science Education, 92*(3), 404–423. doi:10.1002/sce.20263

Hattie, J. (2008). *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement*. Routledge.

Hattie, John, & Timperley, H. (2007). The power of feedback. *Review of Educational Research 77*(1), 81-112. https://doi.org/10.3102/003465430298487. http://rer.sagepub.com/cgi/content/abstract/77/1/81.

Henderson, J.B., Osborne, J., MacPherson, A., & Szu, E. (2014). A new learning progression for student argumentation in scientific contexts. In C. P. Constantinou, N. Papadouris & A.

Hadjigeorgiou (Eds.), *E-Book Proceedings of the ESERA 2013 Conference: Science Education Research for Evidence-based Teaching and Coherence in Learning: Part 7* (co-ed. M. Evagorou & K. Iordanou) (pp. 26-42). European Science Education Research Association.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55.

Jimenez-Aleixandre, M. P., Rodriguez, A. B., & Duschl, R. A. (2000). "Doing the lesson" or "doing science": Argument in high school genetics. *Science Education, 84*, 757–792.

Junpeng, P., Inprasitha, M., & Wilson, M. (2018). Modeling of the open-ended items for assessing multiple proficiencies in mathematical problem solving. *The Turkish Online Journal of Educational Technology*, Special Issue for INTE-ITICAM-IDEC 2018 (Volume 2), 142-149.

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods, 38*(1), 88–91.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

Mercer, N., Dawes, L., Wegerif, R., & Sams, C. (2004). Reasoning as a scientist: Ways of helping children to use language to learn science. *British Educational Research Journal, 30*(3), 359–377.

Metz, K. E., Cardace, A., Berson, E., Ly, U., Wong, N., Sisk-Hilton, S., Metz, S. E., & Wilson, M. (2019). Primary grade children's capacity to understand microevolution: The power of leveraging their fruitful intuitions and engagement in scientific practices. *Journal of the Learning Sciences, 28*(4-5)*,* 556-615. DOI: 10.1080/10508406.2019.1667806

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97. https://doi.org/10.1037/h0043158

Morell, L., Collier, T., Black, P., & Wilson, M. (2017). A Construct-Modeling approach to develop a learning progression of how students understand the structure of matter. *Journal of Research in Science Teaching*, *54*(8), 1024-1048. doi:10.1002/tea.21397.

Morell, L., Suksiri, W., Dozier, S., Osborne, J., & Wilson, M., (2019, January). *Addressing the NGSS practice of Arguing from Evidence using forced-choice item formats: Challenges and successes*. Paper presented at the IES Annual Principal Investigators Meeting, Washington, DC.

Morell, L. Dozier, S., Suksiri, W., Osborne, J., & Wilson, M. (2021, Feb). *Gaining insight into teachers' interpretations of computer-based feedback for the purpose of valid inferences.* Paper presented at the IOMW 2020 Virtual Conference.

Morell, L., Suksiri, W., Dozier, S., Osborne, J., & Wilson, M. (2020, September). *An exploration of selected-response items compared to constructed-response item types in science educatio*n. Presented at the National Council on Measurement in Education (NCME), online Annual Meeting.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. The National Academies Press.

Newton, P., Driver, R., & Osborne, J. F. (1999). The place of argumentation in the pedagogy of school science. *International Journal of Science Education, 21*(5), 553–576.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. The National Academies Press.

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*, 994–1020.

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S. Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching, 53*(6), 821–846. https://doi.org/10.1002/tea.21316

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*(1), 1–4.

Pollock, E., Chandler, P., & Sweller, J. (2002). Assimilating complex information. *Learning and Instruction, 12*(1), 61–86.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danmarks Paedagogiske Institut.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 53*(3), 349–359.

Schwarz, B. B., Neuman, Y., & Biezuner, S. (2000). Two wrongs may make a right... If they argue together! *Cognition and Instruction, 18*(4), 461–494.

Siddiq, F., Gochyyev, P., & Wilson, M. (2017). Learning in digital networks—ICT literacy: A novel assessment of students' 21st century skills. *Computers & Education, 109*, 11-37.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16(*2), 209–220.

Toulmin, S. (1958). *The uses of argument*. Cambridge University Press. Retrieved from http://www.dlpdfs.com/pdf03/uses%20of%20argument.pdf

Venville, G. J., & Dawson, V. M. (2010). The impact of a classroom intervention on grade 10 students' argumentation skills, informal reasoning, and conceptual understanding of science. *Journal of Research in Science Teaching, 47*(8), 952–977.

Von Aufschnaiter, C., Erduran, S., Osborne, J., & Simon, S. (2008). Arguing to learn and learning to argue: Case studies of how students' argumentation relates to their scientific knowledge. *Journal of Research in Science Teaching, 45,* 101–131.

Weiss, I.R., Pasley, J. D, Smith, P. S., Banilower, E. R., & Heck, D. J (2003). *A Study of K–12 Mathematics and Science Education in the United States*. Horizon Research.

Wilson, M. (2004). A perspective on current trends in assessment and accountability: Degrees of coherence. In, M. Wilson, (Ed.). *Towards coherence between classroom assessment and accountability. 103rd Yearbook of the National Society for the Study of Education, Part II (*pp. 272-283). University of Chicago Press.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates, Inc.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, *46*(6), 716–730.

Wilson, M. (2012). Responding to a challenge that learning progressions pose to measurement practice: Hypothesized links between dimensions of the outcome progression. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions.* (pp. 317–343). Sense Publishers.

Wilson, M. (2018). Making measurement important for education: The crucial role of classroom assessment. *Educational Measurement: Issues and Practice, 37*(1), 5-20.

Wilson, M. (2020, April). *Promises and perils for classroom assessments in the digital era*. E. F. Lindquist Award presentation, AERA Annual Meeting San Francisco, CA http://tinyurl.com/wnxmsxl (Conference Canceled).

Wilson, M. (2023). *Constructing measures: An item response modeling approach*., Second Edition. New York, Routledge.

Wilson, M., & Draney, K. (2002). A technique for setting standards and maintaining them over time. In S. Nishisato, Y. Baba, H. Bozdogan, & K. Kanefugi (Eds.), *Measurement and multivariate analysis* (Proceedings of the International Conference on Measurement and Multivariate Analysis, Banff, Canada, May 12-14, 2000), pp 325-332. Springer-Verlag.

Wilson, M., Morell, L., Osborne, J., Dozier, S., & Suksiri, W. (2019, April). *Assessing higher order reasoning using technology-enhanced selected response item types in the context of science*. Paper presented at the 2019 Annual Meeting of the National Council on Measurement in Education in Toronto, Ontario, Canada.

Yao, S-Y., Wilson, M., Henderson, J. B., & Osborne, J. (2015). Investigating the function of content and argumentation items in a science test: A multidimensional approach. *Journal of Applied Measurement, 16*(2), 171-192.

Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching, 39(*1), 35–62.

Appendix A. Pattern of the Factor Loadings

Table A1 below presents the pattern of factor loadings for the 20 argumentation items. As a common criterion, 0.3 is used to determine the strength of the relationship between an item and a factor. To make the factor loading pattern more readable and interpretable, factor loadings with absolute values lower than 0.3 are reported in gray scale. As shown in the Table, all of the argumentation items load positively on the factor as desired. Seven out of the 20 argumentation items have loadings smaller than 0.3.

Table A1. Factor Loadings Obtained from the One-Factor Model

Variable	Factor Loading
A1_3a	0.23
A1_3b	0.28
A1_3c	**0.39**
A1_3d	**0.53**
A1_3e	**0.42**
A2_5a	**0.59**
A2_5b	**0.50**
A2_5c	0.29
A2_5d	0.22
A3_14aa	**0.64**
A3_14ab	**0.49**
A3_14ac	**0.54**
A3_14ba	**0.41**
A3_14bb	0.13
A3_14bc	**0.54**
A3_14c	0.28
A4_7a	**0.30**
A4_7b	0.28
A4_7c	**0.52**
A4_7d	**0.54**

Figure 1. A representation of the BEAR assessment system (BAS).

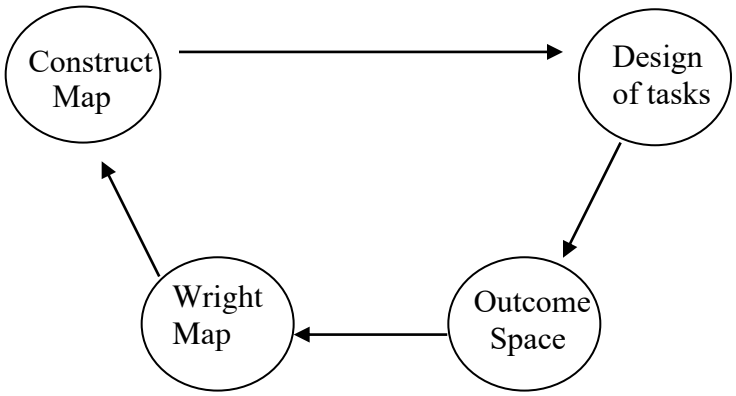

Figure 2. The hypothesized construct map for Argumentation

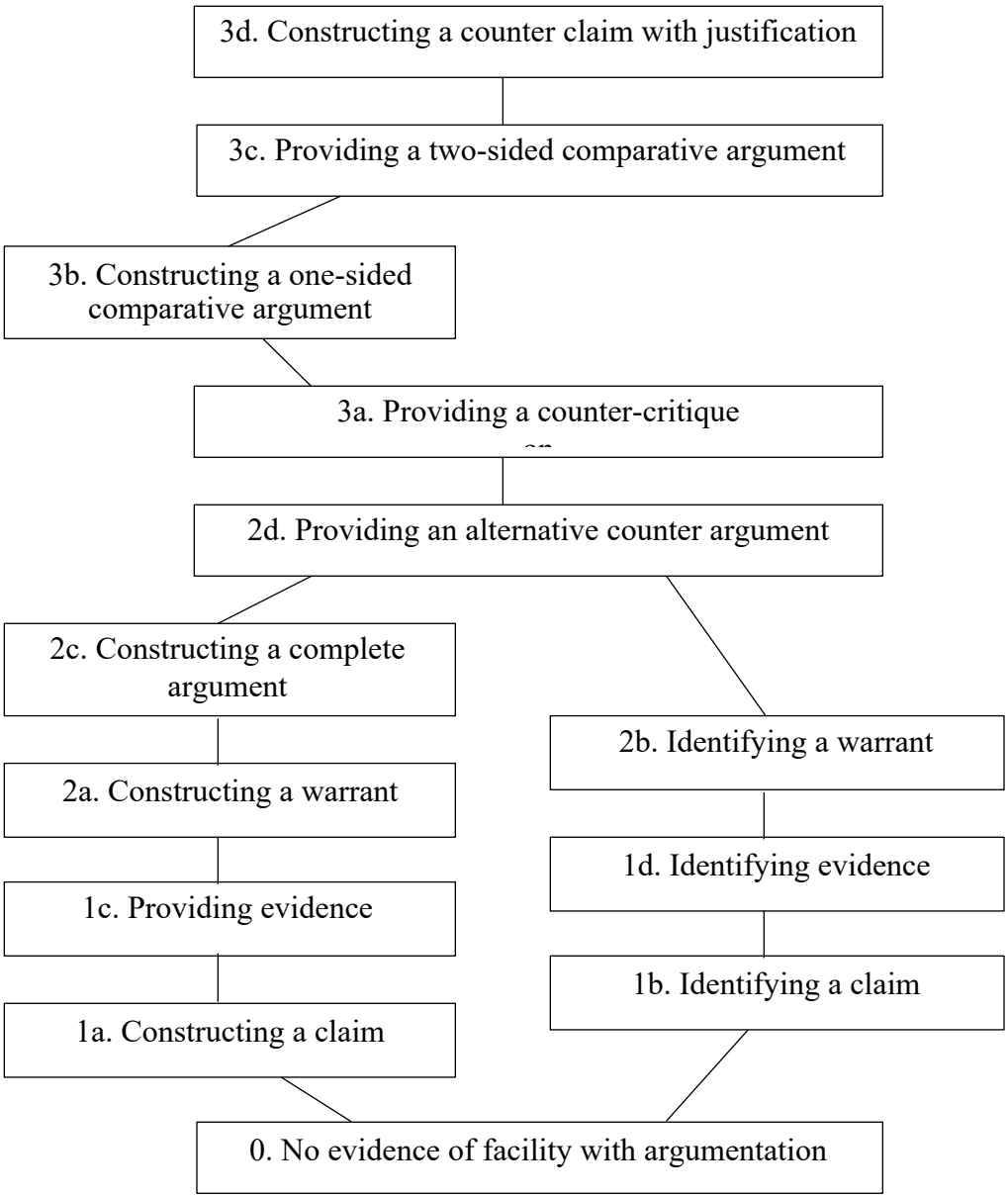

Figure 3. The "Sugar in the Water" Argumentation Item

Item A2. What Happens to Sugar in the Water?

Three students are discussing what happens to sugar grains when the sugar is added to a cup of hot water.

Paul says: *The sugar breaks up and disappears for good.*

Mary says: *The sugar breaks up but is still there and has not changed.*

Laura says: *The sugar has changed to make a new substance.*

They have several pieces of evidence for their argument.

1. The water tastes sweet.
2. After stirring the sugar can no longer be seen.
3. They leave the cup for several days. All the liquid evaporates leaving a sticky, solid substance at the bottom of the cup. This tastes sweet but is one large lump rather than separate grains.
4. They weigh the cup of water and the sugar separately. The weight of the cup of water and the sugar added together is the same as the weight of the cup of water with the sugar added.
5. They have been told that you cannot destroy or make new matter.

5a. Which evidence do you think best supports Paul's argument?
(a) Number ____
(b) Explain why you think this evidence supports Paul's argument.

5b. Which evidence challenges Paul's argument?
(a) Number____
(b) Explain why you think this evidence challenges Paul's argument.

5c. Which piece of evidence best supports Mary's argument?
(a) Number ____
(b) Explain why you think this evidence supports Mary's argument.

5d. Laura claims that the evidence available is not enough to decide whether she or Mary is right. Do you agree? Why?

Figure 4. Scree plot for the EFA analysis

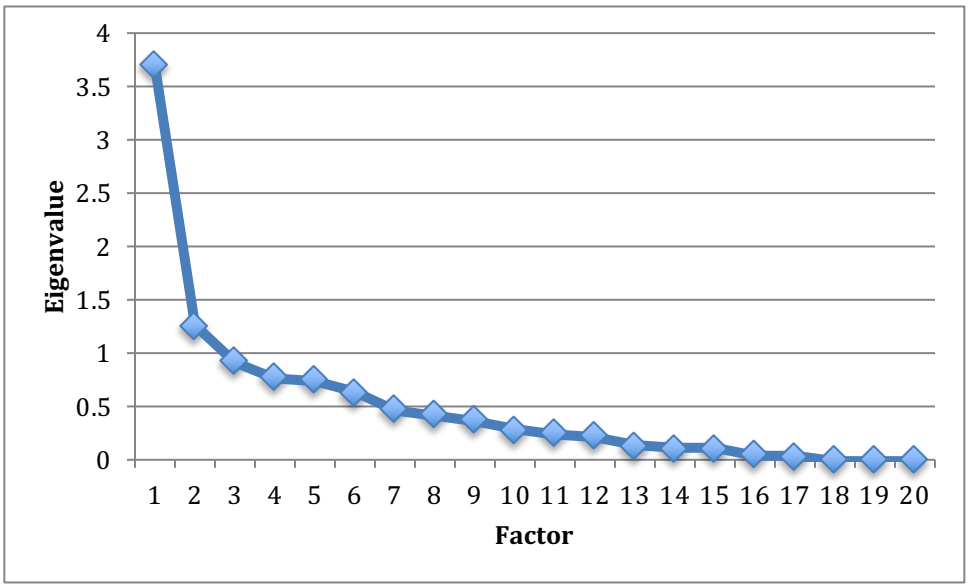

Figure 5. Results of the parallel analysis

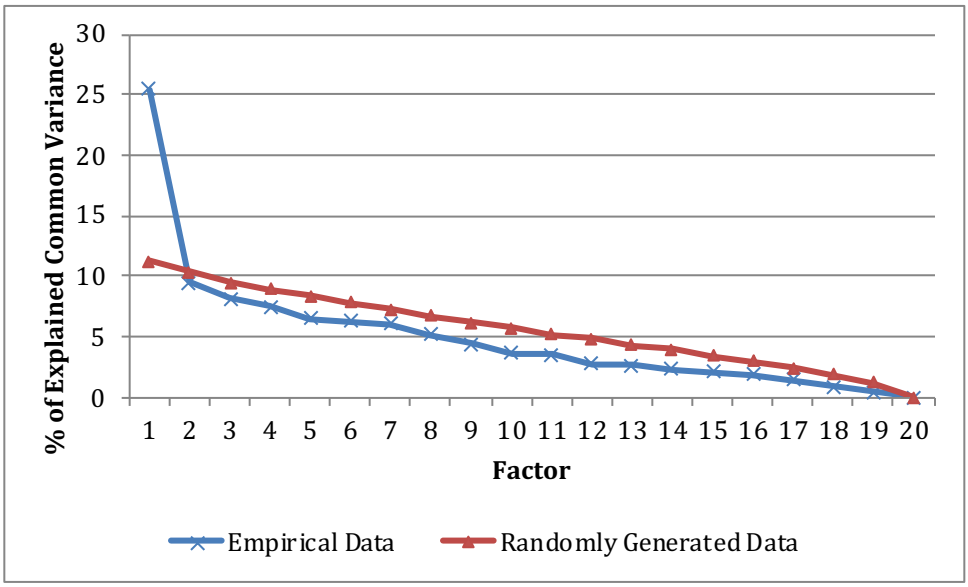

Figure 6. Wright map with item thresholds of the full-credit response categories

logit	Ability	Item Thresholds		
		Level 1	Level 2	Level 3
		A3_14bb.2 A1_3c.2		
2				
		A2_5c.2		
			A3_14c.2	
				A2_5d.3
	X		A4_7d.2	
	X		A3_14bc.2	
	X	A1_3a.2		
1	XX			
	XX	A1_3b.2		
	XXX		A3_14ac.3	
	XXXXX			
	XXXXXX		A2_5b.2	
	XXXXXXXX			
	XXXXXXX			
	XXXXXXXXX			
	XXXXXXXX	A3_14ab.2		
	XXXXXXXXX			
	XXXXXXXXX		A4_7b.2	
	XXXXXXXX	A4_7c.3 A2_5a.2		A1_3e.2
	XXXXXXXXXX			
0	XXXXXXXXXX	A3_14aa.2		
	XXXXXXXX	A1_3d.2	A4_7a.2	
	XXXXXXXX			
	XXXXXXXX			
	XXXXXXX			
	XXXXXXXX	A3_14ba		
	XXXXXX			
	XX			
	XXX			
	XXX			
	XXX			
	XX			
-1	XX			
	XX			
	X			
	X			
	X			
	X			
	X			
	X			
-2				
	X			

Figure 7. Edited version of the Wright map with item thresholds of the full-credit response categories

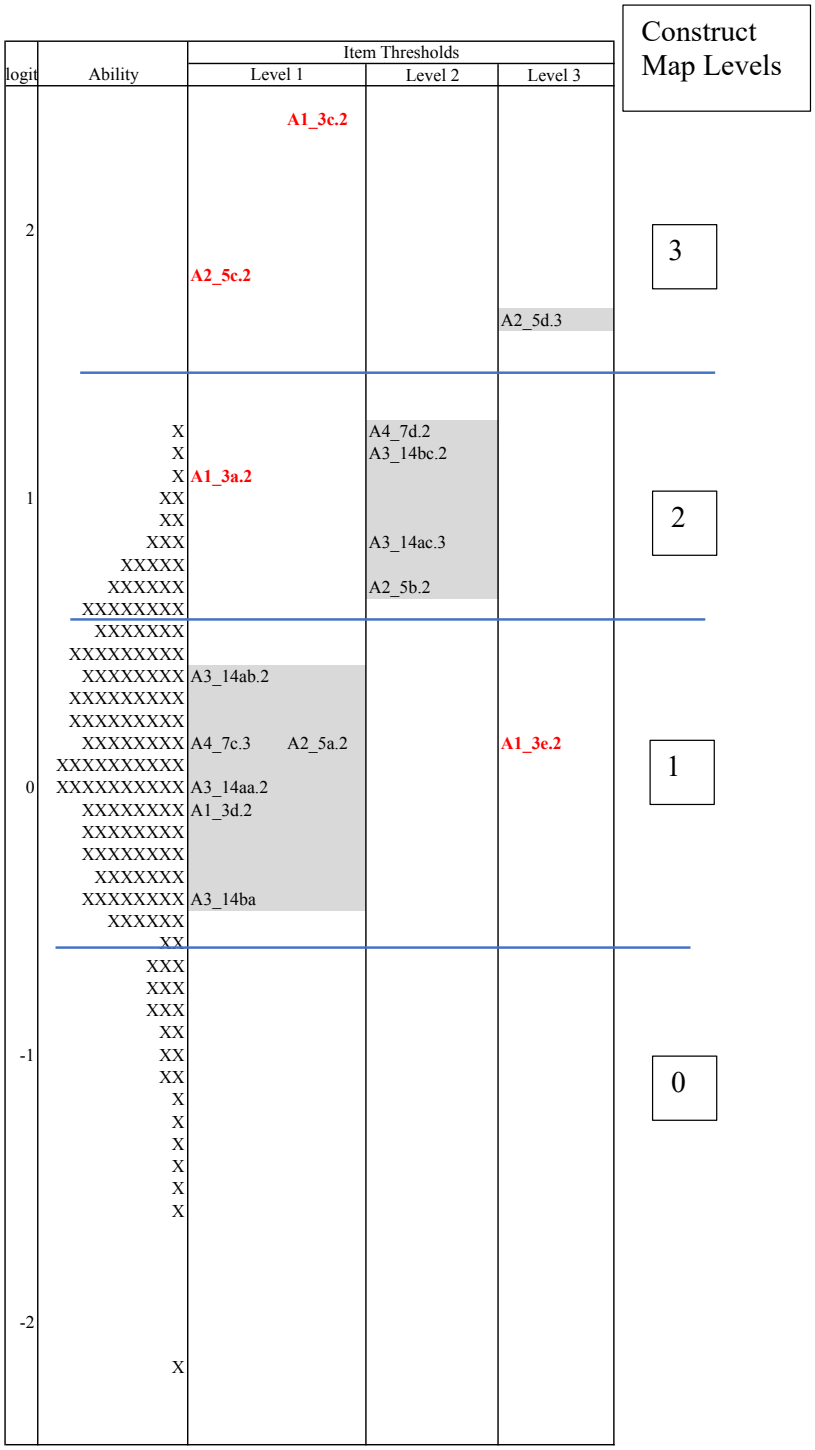

Figure 8. Items A1_3c and A1_3d

Item A1. Desalinating Water

California is increasingly short of fresh water. Lack of fresh water limits the amount of food that can be grown. Clearly, there is a need to better manage this increasingly valuable resource. 97% of the world's water is in the oceans. The trouble is that it is salty and cannot be drunk. The process of removing salt from water is called *desalination*.

One method of desalination is to use energy to heat seawater so that it evaporates leaving the unwanted salt behind. The water vapor is then passed over a cold object where it turns back to pure water. A great deal of energy is lost during this process.

Another method involves pushing the seawater through a sheet with tiny holes in it called a membrane. Only the water particles pass through leaving all the salt on the other side. This uses half the energy, but the membranes get clogged and have to be cleaned regularly. All the sludge produced then has to be taken away.

Figure 9. Graphic display of the locations of a class of students on the argumentation learning progression

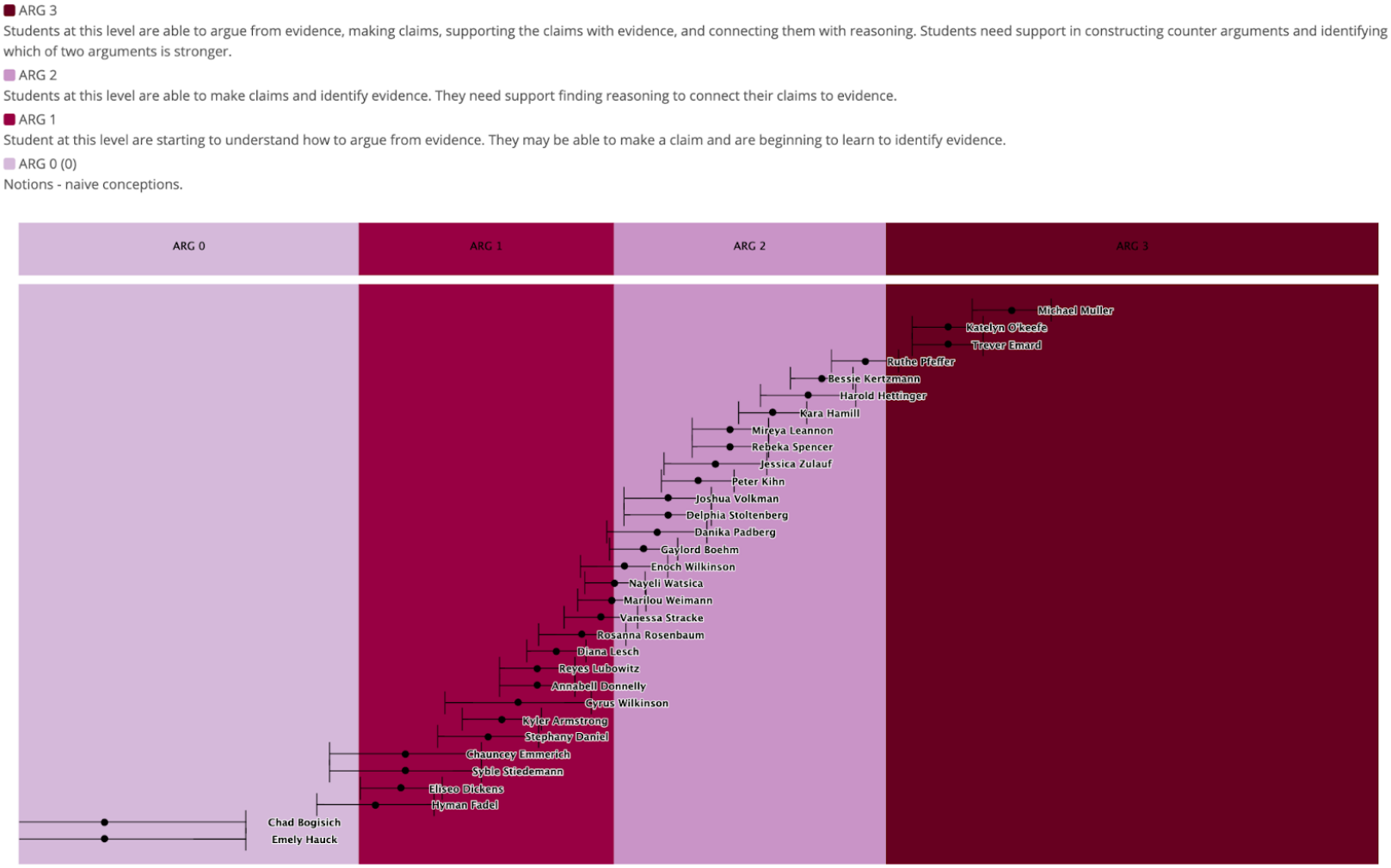

Figure 10. Graphic display of an individual student's location on the argumentation learning progression