Validation of a Learning Progression-based Monitoring Assessment

Jennifer H. Doherty[1], Karen Draney[2], Hyo Jeong Shin[2], JinHo Kim[2], Charles W. Anderson[3]

[1]University of Washington, Department of Biology
[2]University of California, Berkeley, Graduate School of Education
[3]Michigan State University, Department of Teacher Education

Contents

Abstract

In this article we report on our progress towards validating a large-scale, or monitoring, assessment of science learning that utilizes a learning progression-based conceptual model of the progressive nature of science learning. The *Carbon: Transformations In Matter and Energy (Carbon TIME)* assessment is designed to measure student understanding of a key topic in the science curriculum: carbon-transforming processes in socio-ecological systems. We ask, to what extent can we use assessment results to validly distinguish among the explanation practices of students at three scales: individual students, class-size groups of students, or program-size groups of students. We present an empirical validity argument of the utility of this assessment for this purpose using the *Standards for Educational and Psychological Testing* (2014) developed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. We conclude that the Carbon TIME assessment is clearly useful for learning purposes and for program evaluation, but that results comparing individual students or teachers can only be used with great caution.

Introduction

How can we use learning progression research to develop valid monitoring assessments of science learning? Although this is a simply stated question, it does not have a simple answer.

In this article we develop an answer to this question, in three parts.

1. *Defining the problem.* In this section we define the scope of this article using key terms in the opening question: *learning progression, valid, monitoring assessment,* and *science learning*. Our goal is to design a valid large-scale, or monitoring assessment, of science learning that utilizes learning progression-based conceptual models of the progressive nature of science learning.
2. *Reviewing current assessment validation studies.* We briefly review studies of the validity of current science monitoring assessments and of current learning progression-based assessments.
3. *Validity arguments for a monitoring assessment of carbon-transforming processes.* The main section of this paper uses the *Standards for Educational and Psychological Testing* (2014) developed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (AERA/APA/NCME) as a framework for a set of empirical validity arguments for a learning progression-based monitoring assessment, the *Carbon: Transformations In Matter and Energy (Carbon TIME)* assessment, designed to measure student understanding of a key topic in the science curriculum: carbon-transforming processes in socio-ecological systems.

Defining the Problem

In this section we use four authoritative reports, the AERA/APA/NCME *Standards for Educational and Psychological Testing* (2014) and a series of three reports from the National Research Council (NRC)—*Knowing What Students Know* (NRC, 2001), *Systems for State Science Assessment* (NRC, 2005), and *Developing Assessments for the Next Generation Science Standards* (NGSS; NRC, 2014), to define the scope of this article. We first discuss the kind of science assessment we focus on for this article, large-scale or monitoring assessments.

We next discu*ss* standards for validity. We then discuss how learning progressions can be used to assign a particular meaning to science learning. Finally, we pose three research questions around which our validity arguments are organized.

Focus on Monitoring rather than Classroom Assessments

Two NRC assessment reports, *Systems for State Science Assessment* and *Developing Assessments for the Next Generation Science Standards* (2005, 2014) emphasize that standards-based reform will require the development of assessment systems that include many different types of assessments designed for different contexts and purposes.

Classroom vs. monitoring assessments. The NRC report on assessment for the *NGSS* (2014) describes two broad categories for science assessments: classroom and monitoring assessments. These two types of assessments differ in a number of important ways. Classroom assessments are often embedded in instructional programs: the assessments are closely aligned with an instructional program that is organized around specific curriculum goals. Development of student understanding over time is defined by the particular sequence of lessons in the instructional program. Many studies report on the successful use of learning progressions as frameworks for the development of classroom assessments, with the learning progression tightly linked to a particular sequence of lessons (Claesgens, Scalise, Wilson, & Stacy, 2009; Furtak, 2012; Gotwals & Songer, 2010; Lehrer & Schauble, 2012; Schwarz et al., 2009).

Tight coherence among assessments, curricula, and instruction is appropriate for classroom assessments, but for larger-scale monitoring assessments it can be problematic. Monitoring assessments are necessarily at a coarser grain size, so that they are not linked to a particular set of curricular activities; however, they can still be based on the same goals for science education that was used to develop the classroom assessments (and to select appropriate curricula). Monitoring assessments have a number of important uses that transcend evaluation of specific curricula:

> They can be used to answer a range of important questions about student learning, such as: How much have the students in a certain school or school system learned over the course of a year? How does achievement in one school system compare with achievement in another? Is one instructional technique or curricular program more effective than another? What are the effects of a particular policy measure, such as reduction in class size? (NRC, 2014, p. 5-1)

The success of learning progression-based classroom assessments does not assure that learning progressions can also be useful for monitoring assessments. For monitoring assessments, the learning progression framework must describe "how students' science understanding develops over time" for a broader range of students who may not be experiencing common curricula or instructional strategies. This requires using a *learning progression framework* that defines stages or levels of proficiency to characterize student performances but is not tied to a particular instructional sequence.

Design issues for monitoring assessments. Monitoring assessments can conceivably be used for multiple purposes, and those different purposes are associated with design tradeoffs. In this article we focus particularly on the tensions between what Shepard (2013) describes as *accountability* and *learning* purposes of assessments. Accountability is particularly associated with high-stakes summative assessments that have implications for the future success of students, teachers, schools, or programs. When assessments are designed for accountability purposes, *fairness* is a key design goal: the scores that the tests produce must be consistent and accurate and not disadvantage a particular group of people. Learning-oriented assessments, on the other hand are designed more to produce *insight* into the nature and origins of student performances, helping schools and teachers improve the quality of the education that they offer. Ideally our assessments should be both fair and insightful, but as Shepard and others (e.g., Penuel, Confrey, Maloney, & Rupp, 2014) have pointed out, there are many practical and legal considerations that make it difficult to achieve both of these goals with the same assessment.

Standards for Validity Arguments

As Baker describes, validity “addresses the quality of the test and its results and, most simply, whether a parent, a teacher, a scholar, or a policymaker should trust test results and inferences they suggest about school effectiveness, student achievement, teacher evaluation, or changes needed in instruction” (2013). The AERA/APA/NCME *Standards* make two key points about the nature of validity and the validation process.

1. Validity is associated with proposed uses or interpretations of assessment results, not assessments alone. All assessments can be misinterpreted or misused, so we can never simply say that any assessment is “valid.” “Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests.” (AERA/APA/NCME, 2014, p. 11). As the quote from the NRC report above suggests, the proposed interpretations of monitoring assessments and classroom assessments are substantially different and hence require different types of evidence.

2. Empirical arguments for validity can be constructed using different sources of evidence. What evidence is needed to argue for the validity of an assessment? The AERA/APA/NCME *Standards* describes “a sound validity argument” as one that “integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses” (2014, p. 21). The *Standards* provide a set of six major types of validity evidence, five of which are relevant for the current paper:

1. *Evidence Based on Test Content* is “the relationship between the content of a test and

the construct[1] it is intended to measure. Test content refers to the themes, wording, and format of the items, tasks, or questions on a test." (AERA/APA/NCME, 2014, p. 14)

2. *Evidence Based on Response Processes* is "the fit between the construct and the detailed nature of performance or response actually engaged in by test takers." (AERA/APA/NCME, 2014, p. 15)
3. *Evidence Based on Internal Structure* is "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." (AERA/APA/NCME, 2014, p. 16)
4. *Evidence Based on Relations to Other Variables* is the "relationship of test scores to variables external to the test… External variables may include measures of some criteria that the test is expected to predict, as well as relationships to other tests hypothesized to measure the same constructs, and tests measuring related or different constructs." (AERA/APA/NCME, 2014, p. 16)
5. *Evidence for Validity and Consequences for Testing*: "Educational tests … may be advocated on the grounds that their use will improve student motivation to learn or encourage changes in classroom" (AERA/APA/NCME, 2014, p. 20).

We note that there is an order to these five types of evidence. The first two focus primarily on qualities of individual items; the third focuses on the coherence of item sets or instruments containing multiple items; the fourth focuses on relationships among instruments measuring related constructs; the final type of evidence focuses on the role of assessment in education systems.

[1] In our definition of construct, we use the work of Mark Wilson (2005). He defines a construct map as "a well thought out and researched ordering of qualitatively different levels of performance focusing on one characteristic." A construct map defines what is to be assessed in terms general enough to be interpretable within a curriculum and potentially across curricula, but specific enough to guide the development of the other components (items, scoring guides, and the measurement model).

Science Learning and Learning Progressions

The AERA/APA/NCME *Standards* are general, designed to apply to all educational assessments. In this section, we address how these standards apply specifically to science monitoring assessments, and the potential of learning progression research to contribute to improved monitoring assessments.

Conceptual criteria based on learning progression theory and practice: The assessment triangle. The usual procedure for test content validation is a "panel of experts" approach, in which questions of student cognition are not addressed. *Knowing What Students Know* (NRC, 2001) endorsed a cognitive interpretation of "evidence based on test content."

> The central problem addressed by this report is that most widely used assessments of academic achievement are based on highly restrictive beliefs about learning and competence not fully in keeping with current knowledge about human cognition and learning. Likewise, the observation and interpretation elements underlying most current assessments were created to fit prior conceptions of learning and need enhancement to support the kinds of inferences people now want to draw about student achievement. *A model of cognition and learning should serve as the cornerstone of the assessment design process. This model should be based on the best available understanding of how students represent knowledge and develop competence in the domain.* (NRC, 2001, pp. 3-4, emphasis in original)

Both of the NRC science reports (2005, 2014) follow *Knowing What Students Know* in representing the key components of an assessment system with the *assessment triangle* (Figure 1): cognition, observation, and interpretation. Assessment instruments (the observation vertex) must be aligned with the knowledge and cognitive processes one wishes to affect through the instructional process (the cognition vertex), and the scoring and interpretation of student work must reflect measures of the same knowledge and cognitive processes (the interpretation vertex).

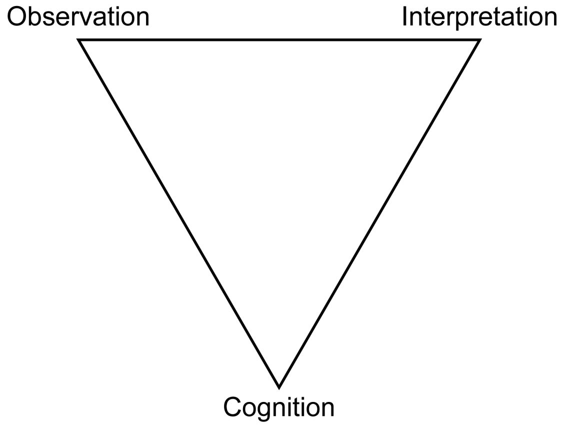


Figure 1: The assessment triangle (NRC, 2005, p. 86)

Learning progressions as models of cognition. Learning progressions are "descriptions of increasingly sophisticated ways of thinking about or understanding a topic that can follow one another as children learn about and investigate a topic over a broad span of time" (NRC, 2007, p. 214). Well-grounded learning progressions can serve as a basis for dialogue among science education researchers, developers of standards documents, assessment developers, and curriculum developers. This learning progression approach is endorsed by both the NRC (2005, 2007) and the National Assessment Governing Board in the framework for the 2009 National Assessment of Educational Progress (NAEP) science test (National Assessment Governing Board, 2006). Conceptual and methodological foundations for learning progressions are described by Briggs, Alonzo, Schwab, & Wilson (2006), Smith, Wiser, Anderson, & Krajcik (2006) and Wilson (2009), among others.

One of the primary reasons that science educators have been interested in learning progressions is their potential as models of cognition for assessment. Indeed, the first reports on science learning progressions (Catley, Lehrer, & Reiser, 2005; Smith et al., 2006) were developed for the NRC panel that developed the report *Systems for State Science Assessment* (2005). This report, and a subsequent report on assessment for the *Next Generation Science*

Standards (2014) advocated systems of assessment that use learning progressions to define standards, develop assessments, and construct curricula.

These documents advocate using learning progression frameworks to define the Cognition vertex of the assessment triangle (Figure 1): Learning progression frameworks provide a particular kind of model of cognition. Rather than simply defining “correct” performance, learning progression frameworks provide descriptions of students’ knowledge and practice at multiple levels—the target performance as well as lower levels of incomplete mastery. Thus the “content” of a learning progression-based assessment is the entire framework, not just the scientifically correct answers. The essential contribution that learning progressions can make is to shift from a binary to a multilevel description of student proficiency.

Characteristics of Current Science Monitoring Assessments

Monitoring assessments in science and other content areas are a multibillion-dollar industry, supported by a large infrastructure in universities and in educational assessment companies. Commentaries about the validity of these tests fill many volumes, including the three NRC reports cited above (2001, 2005, 2014). It is beyond the scope of this paper to add substantively to those commentaries. We will, however, use the framework of the AERA/APA/NCME *Standards* to point out a few of the reasons that the reports have advocated the use of learning progression frameworks in developing monitoring assessments. The discussion below is based on those reports, as well as on Alonzo, Neidorf, and Anderson’s (2012) chapter contrasting current monitoring assessments with learning progression-based assessments.

1. *Evidence Based on Test Content*: Current science monitoring assessments are generally based on state or national standards or on other content frameworks, and their development typically involves elaborate procedures for making sure that the assessments include questions sampling from the domain defined by the standards

documents. The reports above criticize this process on two bases:

a. Inadequate representation of important scientific knowledge and practice, due both to limitations of many standards documents and to the economic decision of many assessments to rely heavily on multiple-choice questions. *Developing Assessments for Next Generation Science Standards* makes this argument particularly forcefully, arguing for "three-dimensional" assessment that incorporates practices, crosscutting concepts, and disciplinary core ideas (NRC, 2014, Chapter 2).

b. Failure to incorporate research on student learning. In contrast with learning progression frameworks *all* current standards documents define constructs in terms of correct scientific knowledge rather than construct maps that include students' non-canonical knowledge and practices.

2. *Evidence Based on Response Processes*: Although many testing organizations include "cognitive laboratories" that use think-alouds, clinical interviews and other methods to check for agreement between students' responses to test items and their reasoning, these checks take a relatively small role in the development process (Padilla & Benítez, 2014). The process is especially restricted when the questions being evaluated are multiple choice. Several studies (De Boer, Dubois, & Herrmann-Abell, 2008; Noble et al., 2012) have found (a) that students' responses are often based on reasoning unrelated to the construct that the item is supposed to measure, and (b) significant numbers of students either chose correct responses based on incorrect reasoning or chose incorrect responses based on correct reasoning.

3. *Evidence Based on Internal Structure*: Current assessments have strong conformity between the measurement model and the observations (two of the vertices in the assessment triangle, Figure 1). They rely extensively on item response theory or classical test theory to make certain of this conformity, including consistency among individual items designed to measure the same construct, monitoring for item difficulty,

monitoring for consistency among different forms of the same test, and other factors. However, the construct (the third vertex of the assessment triangle) is absent or not clearly specified.

4. *Evidence Based on Relations to Other Variables*: Current science assessments are checked extensively for their relationships both to other assessments and to criteria such as future success in high school or college science courses.
5. *Evidence for Validity and Consequences for Testing*: The prolonged debate about the consequences of current assessments for the quality of education has gone far beyond education professionals into the political arena and the popular press. The NRC reports reflect the general consensus among science educators and many measurement specialists that high-stakes assessments based on multiple choice questions generally narrow the science curriculum and distort evaluation of both student learning and teachers' competence.

Thus current science monitoring assessments are designed primarily for what Shepard (2013) refers to as accountability purposes, and the standards they meet with respect to internal structure and relations to other measures of achievement make them *fair* in the sense that fairness is commonly defined in our legal system and intuitively by the general public. The assessments can be administered to all students and scored efficiently; students' scores can be reliably compared across test forms, years, and sites.

However, these qualities are achieved at the cost of compromises that are heavily criticized by both professional educators and the general public (Pellegrino & Chudowsky, 2003; Wong, 2015). They generally elicit a narrow range of student performances, depending heavily on multiple-choice questions. Furthermore, the focus on accountability purposes leaves little room for learning purposes; consistency in scoring is achieved at the cost of insight into student thinking. Thus the deficiencies of current monitoring assessments with respect to other kinds of validity has led the NRC panels to advocate for monitoring assessments that can better serve learning goals, meaning that they must achieve higher standards with respect to content/construct validity and response processes. In this article we report on progress toward developing learning progression-based assessments that meet these standards.

Current Learning Progression Assessments

Current learning progression validation studies, our own and others (Alonzo & Steedle, 2009; Gotwals & Songer, 2013; Gunckel, Covitt, Salinas, & Anderson, 2012; Lehrer & Schauble, 2012; Mohan, Chen, & Anderson, 2009; Plummer & Krajcik, 2010; Rivet & Kastens, 2012; Shea & Duncan, 2013), have been designed with learning rather than accountability as a primary purpose. Thus they are far more careful than current monitoring assessments in collecting and evaluating construct and response process validity evidence. These studies generally use iterative design-based research to carefully align assessment items to the proposed learning progression framework to ensure construct validity. They also use think-alouds and comparisons with student performances in other settings to improve response process validity. This careful work has resulted in empirically supported learning progression frameworks and assessment items.

The evidence presented in these studies, however, falls short of current standards for monitoring assessments with respect to internal structure and relations to other variables. The validity arguments in previous learning progression studies have been for claims about items and learning progression frameworks rather than claims about people (but see Steedle & Shavelson, 2009). For example, internal structure modeling in the recent learning progression validation studies by Gotwals & Songer (2013), Rivet & Kastens (2012), and Neumann, Viering, Boone & Fischer (2013) is used to show that the item difficulties are aligned with the item learning progression levels, but not to evaluate claims about students or groups of students. If we want to use learning progression-based assessments for monitoring purposes—for example, to measure progress of individual students or to compare educational programs—we will need to evaluate the assessments with respect to a broader range of validity evidence. In this article we undertake this kind of evaluation for assessments developed through the *Carbon TIME* program.

Research Questions

We have designed assessments that use learning progression levels to define student achievement and reported on the results of those assessments (Jin & Anderson, 2012a; Mohan et al., 2009). In these articles we provided evidence that the assessments are useful for learning purposes: They provide insight into students' reasoning. But are they also useful for monitoring purposes? Can they be used to make fair comparisons among students, classrooms, or programs? As the AERA/APA/NCME *Standards* make clear, answering this question requires us to attend to multiple forms of validity evidence.

We address the validity issue by specifying three types of claims. We want to use learning progression levels to describe performance and compare achievement at different units of analysis:

1. Individual students: To what extent can we use assessment results to distinguish among the explanation practices of individual students in terms of learning progression levels?
2. Classrooms: To what extent can we use assessment results to distinguish among class-size groups of students in terms of the learning progression levels of students and learning that takes place in different classes taught by different teachers?
3. Programs: To what extent can we use assessment results to distinguish among programs, particularly the difference between the *Carbon TIME* curriculum, designed using learning progression research, to support students understanding of carbon transforming processes, and programs used by teachers before they used *Carbon TIME*?

Validity Arguments for a Monitoring Assessment of Carbon-transforming Processes

In this section we develop validity arguments for the *Carbon TIME* assessment using each of the five types of evidence identified in the AERA/APA/NCME *Standards*. We begin with a summary of the learning progression framework and the kinds of claims about students' knowledge and practice that we aspire to make based on that framework. We then provide a brief overview of the methods we have used to develop and score the *Carbon TIME* assessments. Finally, we address each of the AERA/APA/NCME criteria, presenting and interpreting relevant empirical evidence.

Learning Progression Framework: The Nature of Claims to Be Evaluated

For learning progression-based assessments to be useful as monitoring assessments, we need to be able to make claims about learning progression levels of the students that we are assessing. The domain we focus on is transformations of organic carbon-containing materials at multiple scales, from atomic-molecular to global carbon cycling. In particular we focus on *accounts (explanations and predictions) of carbon-transforming processes* (i.e., photosynthesis, cellular respiration, combustion, digestion, and biosynthesis).

Our empirically-based Carbon Learning Progression Framework for students' evolving language and ideas about matter and energy in carbon-transforming processes has been described previously (Jin & Anderson, 2012a; Mohan et al., 2009) and is represented in Figure 2. The rows in Figure 2 represent four levels of achievement, or stages in the transition from informal or force-dynamic explanations (Pinker, 2007; Talmy, 1988) to scientific reasoning:

- Level 1: Pure force-dynamic explanations: Students' explanations focus on actors and enablers, using relatively short time frames and macroscopic scale phenomena. Events are connected by cause and effect rather than by tracing matter and energy.

- Level 2: Elaborated force-dynamic explanations: Students' explanations continue to focus on actors and enablers, but they add detail and complexity, especially at larger and smaller scales. They include ideas about what is happening inside plants and animals when they grow and respond, for example, and they show awareness of larger scale connections among phenomena, such as food chains.
- Level 3: Incomplete or confused scientific explanations: Students show awareness of important scientific principles and of models at smaller and larger scales, such as atoms and molecules, and relationships among populations in ecosystems. They have difficulty, though, connecting accounts at different scales and applying principles consistently to constrain their explanations.
- Level 4: Coherent scientific explanations: Students successfully apply fundamental principles such as conservation of matter and energy to phenomena at multiple scales in space and time. They give complete and accurate accounts of all of the matter and energy in a system before and after an event and constrain their explanations by laws of conservation of matter and energy. Level 4 explanations are aligned with the 12th grade expectations described in the NRC *Framework for K-12 Science Education* (2012) and Next Generation Science Standards (NGSS Lead States, 2013).

Levels of Achievement	Context				
	Plants	Animals	Decay	Burning	Large Scale
4: Using matter and energy to constrain explanations					
3: Partial success in using matter and energy to constrain explanations					
2: Elaborated force-dynamic explanations					
1: Pure force-dynamic explanations					

Figure 2: The Carbon Learning Progression Framework

Our original work was based on data from students across an age span from upper elementary school through college. *Carbon TIME*, however, is a middle and high school program. Level 1 explanations are relatively rare among these students, and for reasons we describe below, assessment items that distinguish Level 1 from Level 2 performances are of limited use for assessing performances at Levels 3 and 4. Therefore the *Carbon TIME* assessment system does not distinguish Level 1 and Level 2 performances, focusing on distinctions among Levels 2, 3, and 4. The framework also contains five different contexts (the columns in Figure 2) in which we ask students to reason about carbon transforming processes. In each context students have to reason using one or more carbon transforming process (i.e., photosynthesis, cellular respiration, combustion, digestion, biosynthesis).

Thus each level of the learning progression encompasses a wide range of performances: predictions and explanations of different processes, in different contexts, at different scales. A key claim implicit in the structure of the framework is that these performances are all related. Students have underlying approaches to making sense of the world that lead to predictable similarities in their accounts of carbon-transforming processes in different contexts.

Mastering scientific discourse involves learning and applying principles (especially conservation of matter and energy) and models (especially atomic-molecular models) that are applicable across processes. We also claim that lower-level students do not see scientific connections among contexts, but nevertheless provide similar accounts among contexts because they draw on a common pool of linguistic and conceptual resources to construct their accounts. These claims are empirically testable. We present and discuss relevant evidence in the sections on response processes and internal structure below.

The claims above relate mostly to the individual student level of analysis (Research Question 1). We note that since we intend to use the *Carbon TIME* assessments as monitoring assessments, there are related claims at the classroom level and the program level that we want to evaluate (Research Questions 2 and 3).

Methods overview

In this section we briefly describe key aspects of the methods we used to develop the *Carbon TIME* assessments and to collect and analyze validity data. These methods are described in more detail in other publications, including those referenced above.

BEAR Assessment System

Developing and validating an assessment that meets the empirical and conceptual criteria is a complex, iterative process. The NRC report *Knowing What Students Know* (2001) represents this process with an Assessment Triangle. The BEAR Assessment System (Wilson, 2009; Wilson & Sloane, 2000) uses four interrelated components to create a coherent assessment system:

- Construct Maps: Theoretical learning progression framework
- Items Design: Development of assessment items
- Outcome Space: Ordered set of student performances
- Measurement Model: Item response theory (IRT) analysis

These components align with the NRC's Assessment Triangle; the construct maps represent the cognition vertex; the items design the observation vertex, and the outcome space and measurement model the interpretation vertex.

The first and most important of the components are *construct maps*, which are qualitatively ordered levels of performance in a concept or skill of particular importance. Constructs and levels of performance are derived partly from theories about how knowledge and practice are organized and partly from empirical research data. In our work construct maps may be seen as a learning progression framework. For a more detailed description of these building blocks and their relationship to principles and products, see Wilson (2009). This system shares similarities with other assessment design systems, such as Evidence-Centered Design (Mislevy, Steinberg, & Almond, 2003); in fact, the two have been used in concert in the Principled Assessment Designs for Inquiry (PADI) project (Mislevy, Hamel, et al., 2003).

***Carbon TIME* Assessment**

Carbon TIME Assessment development began in 2003. Each year we have developed or refined the assessment, used it to describe student achievement, and collected validity evidence for the learning progression framework. (Jin & Anderson, 2012a, 2012b; Mohan et al., 2009). This current work shifts our focus from the validity of the *framework* as a way of understanding student thinking to the validity of the *assessment* as a way to compare the learning progression levels among individual students, groups of students in the classrooms of different teachers, and of students in different programs.

The current *Carbon TIME* item pool contains items that typically include a series of forced choice questions followed by an open-ended question where the students are asked to explain the reasoning behind their forced choice selections (see Table 1 for the most current version of the assessment; see supplementary information for current and past item text). There are three forms of the assessment, with an average of 15 items per form, 10 items from the

Carbon TIME pool and 5 items about Matter and Energy from the American Association for the Advancement of Science (AAAS) Project 2061 Science Assessment (http://assessment.aaas.org/) we included as external validity evidence. Each form has two *Carbon TIME* (forced choice plus open-ended) and two AAAS forced choice items as linking items, meaning that these items also appear on other forms. The assessments change every year due to the iterative nature of our work.

Table 1: Carbon TIME assessment used in 13-14 with Cohort 3. The assessment has three forms (A, B, C) which are linked by overlapping items and, as a whole, has items from each context in the Carbon Learning Progression Framework (Figure 2). * indicate Carbon TIME items not scored using the learning progression framework[2].

Form	A	B	C
Plants	ENERGRASS OAKTREE PLANTDIEDECAY* PLANTMAS	ENERGRASS	OAKTREE
Animals			BODYHEAT2 CRICKETMAS FATLOSS 12-13 GIRLGROW MOUSEDIE*
Decay	BREADMOLD COMPOSTB DECAYTABLE POTATO		
Burning	BRNALCOHOLMAT	BRNMATCHEN BRNMATCHMAT MATERIALS* OCTAMOLE WAXBURN	
Large Scale	FOODCHAIN12-13	BIOMASSPYRAMID KLGSEASON FLBULBS 13-14 FOODCHAIN 12-13 KLGOVERALL	KLGLOCAL* KLGOVERALL KLGSEASON
Inquiry			MIKE* GLUBEX* LUCIA*
AAAS: included for external validity	ANIMGROWTH BUBBLES FOODWEB MILK VINEGAR	ANIMGROWTH INTRAVEN MILK PLANTFOOD PLFDSOURCE	ANIMGROWTH MILK PLANTDIE SUNTREE WATERFOOD

Study Populations

Carbon TIME is an NSF-funded development and research project for which the goal is to develop a curriculum to help middle and high school students learn how to account for the chemical changes—carbon transforming processes—that are responsible for the structure and functions of all living systems and that support our lifestyles. The curriculum will be publically available on the BSCS website in June, 2015. It was piloted during development by three cohorts of teachers who each taught at least three out of six curricular units (at least six weeks of instruction) from pilot versions of the *Carbon TIME* Curriculum (Anderson et al., 2015). Each year piloting teachers administer the *Carbon TIME* Assessment to their students at the end of the school year before they teach the *Carbon TIME* Curriculum (baseline data collection) and as pre- and post-tests.

The teachers and students in Cohorts 1 and 2 were from seven locations: rural southwest Michigan; Seattle, WA; Bellevue, WA; Baltimore, MD; rural and suburban Colorado; and Santa Barbara, CA. The teachers and students in Cohort 3 are from those seven locations plus additional schools in Montana, Pennsylvania, Oregon, Florida, Indiana, and Minnesota. Each cohort included students from a range of grades (i.e., 6th – 12th grade) and in a range of course types (e.g., 7th grade Life Science, 9th grade Biology, AP Biology). We collected over 10,000 written student assessments; this report is based on a sample from these assessments. For Cohorts 1 and 2, we randomly sampled 60 baseline and post-assessments from teachers who had more than 60 students and 30 matching pre-assessments from those post-assessment students (Table 2). For Cohort 3, we only sampled one class period of students per teacher for all three time points. Sometimes we are missing an assessment time point (e.g., baseline, pre, post) for a teacher and that teacher was excluded due to lack of data for relevant comparisons. The results presented below are from the students in Cohort 3, our most diverse sample, but which is also representative of results from the other two cohorts.

Table 2: Summary of written data analyzed. The table contains the number of teachers in each cohort and the number of baseline, pre and post Carbon TIME assessments analyzed.

	# of teachers	**# of Assessments**			
		Baseline	**Pre**	**Post**	**Total**
Cohort 1					
Middle School	5	233	80	200	513
High School	14	492	394	737	1623
Cohort 2					
Middle School	4	101	122	234	457
High School	16	83	446	733	1262
Cohort 3					
Middle School	16	375	436	350	1161
High School	38	642	1,049	682	2373
Totals	93	1926	2527	2936	7389

Assessment Analysis

To analyze student achievement, we coded each student response to the open-ended assessment items according to the Carbon Learning Progression Framework (Jin & Anderson, 2012a; Mohan et al., 2009).[2] For each item, one researcher coded all responses and another, more experienced researcher coded a randomly selected 10% of those responses as a reliability check. If there was less than 90% agreement of assigned codes for this 10% of responses, the coders met to discuss any issues with the scoring rubric or its interpretation. The full data set was then recoded and another 10% of responses coded as an independent reliability check. This process continued until >90% agreement was reached.

[2] While all items on the *Carbon TIME* assessment were related to matter and energy in carbon cycling, 2 items in 11-12, 3 items in 12-13 and 7 items in 13-14 could not be coded using the Carbon TIME Learning Progression Framework. These items can be categorized into three types: 1) items about how energy is stored in organic matter, thus not asking students how matter or energy is transformed, 2) items focusing on inquiry and arguments from evidence about tracing matter, and 3) an item about global scale atmospheric carbon movement. While these items could not be coded using the learning progression, we were able to develop scoring rubrics that we used to provide ordinal codes to use for analysis.

After all open-ended responses were scored, an item response model, specifically a modification of the Partial Credit Model (a Rasch-family model for polytomous data developed by Wright & Masters, 1982), was used in the analysis of the data. Individual forced choice items were summed across each set referring to the same stimulus materials, and the weights of the resulting item scores were adjusted so that all items had the same maximum value (i.e. 2.0) and thus made the same contribution to the total proficiency estimate (e.g., OAKTREE in Figure 3 has five forced choice items that were summed and scaled to determine a total forced choice—TFC—score for OAKTREE). Such models produce estimates of item difficulties, difficulties of steps between levels of an item, ability or proficiency estimates for persons, and measures of fit for individual items and persons. We used the ConQuest 2.0 software for item response modeling analyses (Wu, Adams, Wilson, & Haldane, 2007). In the analysis phase, we first checked whether item difficulties showed drift across years, and concluded that the measurement invariance of the items held over the period of the data collection we use in this study.

ConQuest 2.0 provides several different methods for producing person proficiency estimates; we chose the Weighted Likelihood Estimation (WLE) method, as it produces the least bias at the individual person level (Warm, 1989).

In order to convert proficiency estimates (in logits[3]) to learning progression levels, we computed cut scores as the median of the item thresholds on the logit scale (0.24 between Level 2 and Level 3, and 2.01 between Level 3 and Level 4) (H. Shin, Choi, & Draney, 2012). Once the students had been assigned proficiency levels, we used these cut scores to determine whether they were most likely to be at Level 2, 3, or 4, or at the boundaries between the levels. Readers interested in the details of the models used, and the software used to fit these models, should refer to Adams, Wilson, and Wang (1997) for technical details of the model, and to Wu et

[3] Logits are defined as the log odds of providing a response at a given level versus one lower.

al. (2007) for the ConQuest software. We used the R Package 'WrightMap' to draw Wright maps and item fit plots (Irribarra & Freund, 2014).

Interview Comparison

To collect validity evidence with regards to response processes, we compared students' written assessment responses to their responses to semi-structured clinical interviews to a similar prompt (Padilla & Benítez, 2014). We coded and analyzed pre- and post-interviews of students of Cohort 2 teachers. The teachers conducted the face-to-face interviews during school year before and after *Carbon TIME* instruction. The teachers were instructed to choose two students that represented the range in academic success of the typical students in their classroom. Teachers were provided with semi-structured interview protocols consisting of 12 tasks (see supplementary materials for full text of the interview protocol). One task, TREEGROW, was aligned with the written item OAKTREE, an item that was on two thirds of written tests administered. The interviews were video recorded and the audio transcribed for analysis. Interviews that were performed within one month of a written assessment were analyzed (n= 49 for full interview v. written test comparison, n=31 for TREEGROW v. OAKTREE comparison, some teachers administered the written assessment and interview assessment for the same time point [pre or post] months apart). The interview as a whole and the OAKTREE aligned task were analyzed using the Carbon Learning Progression Framework (Jin & Anderson, 2012a; Mohan et al., 2009). To evaluate how similar students performed in the interviews and on written assessments we 1) tested the correlation between the interview learning progression level estimated student proficiency on the written assessment using a Spearman's rank correlation and 2) compared student learning progression levels on the TREEGROW interview task and the OAKTREE written item.

The following sections on reliability and validity evidence include further methodological details and descriptions of how we used the described methods and analysis to construct a validity argument.

Results: Discussion of Reliability and Validity Evidence

The methods summarized above provide evidence relevant to reliability and each of the five AERA/APA/NCME validity criteria. In this section we present that evidence and discuss its implications for the three kinds of claims we want to make—about learning progression levels of individual students, classes taught by different teachers, and the *Carbon TIME* program as a whole. As noted above, there is an order to the five validity criteria. The first two focus primarily on qualities of individual items; the third focuses on the coherence of instruments containing multiple items; the fourth focuses on relationships among instruments measuring related constructs; the final type of evidence focuses on the role of assessment in education systems.

Reliability

Two types of reliability evidence were collected for these data. First, we determined the internal consistency of the Carbon TIME data, using the IRT model described above. Reliabilities were quite high: about 0.84 for the Carbon TIME assessment. Second, we used the proportion of exact matches from the 10% of responses that were independently double-scored to give evidence of inter-rater reliability. In all cases, the proportion of exact matches was above 90%.

1. Validity evidence based on test content

The first AERA/APA/NCME criterion focuses on the Observation vertex of the Assessment Triangle (Figure 1). For traditional science monitoring assessments, this criterion focuses on the *scientific content* of questions and student responses. For a learning progression-based assessment, the assessment should include questions that assess students' proficiencies in tracing matter and energy through carbon-transforming processes at different scales in multiple contexts—essentially Level 4 in the learning progression framework above. Table 1 shows the alignment between the contexts of the framework and the Carbon TIME item pool for the Cohort 3 assessments (see supplementary material for similar tables for Cohorts 1 and 2 assessments).

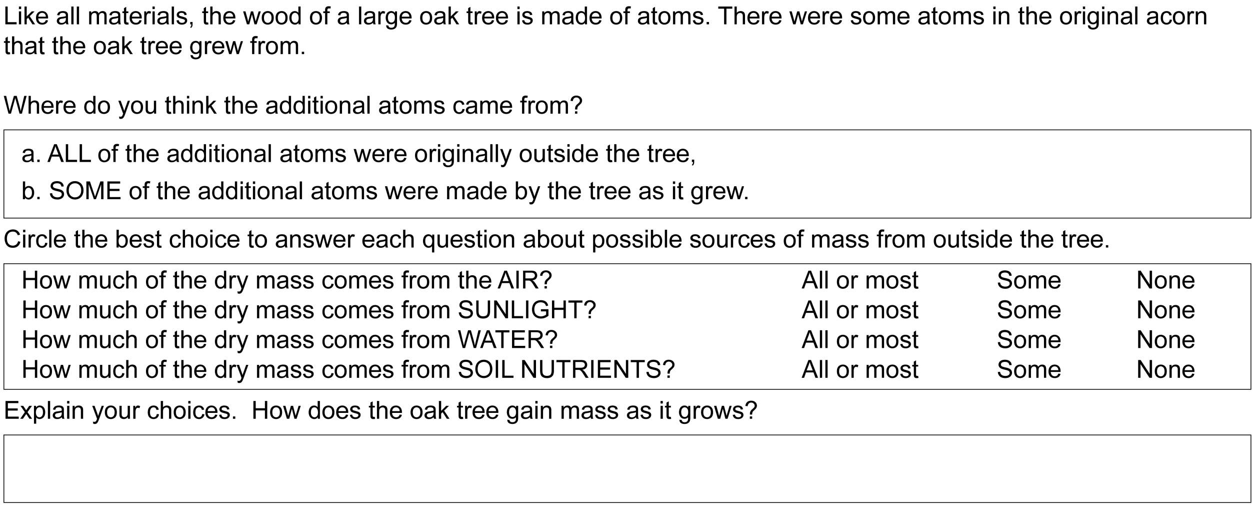
For a learning progression-based assessment, meeting these scientific content criteria is necessary but not sufficient. The Carbon Learning Progression Framework (Figure 2) encompasses performances involving different processes, in different contexts, at different scales, at different learning progression levels. Successful monitoring assessments should elicit this range of performances. That is, when constructing the assessment items, we need to make sure that the responses students provided were at all three levels of interest (Levels 2-4). We have reported on item and assessment design challenges before (Jin & Anderson, 2012b). We used the following learning progression item design criteria to write items that capture this diversity of responses:

- *Items should be designed to elicit a range of possible responses,* not just correct and incorrect answers (N. Shin, Stevens, & Krajcik, 2010). Do items elicit potential indicators for all levels of the learning progression framework or are they only useful for a limited range of students?
- *Items should signal to students that scientific reasoning is preferred.* Our theory is that higher level students are “bilingual:” They are capable of producing higher-level or lower-level responses depending on their judgment of which are appropriate for a particular situation, so items that are too vaguely or informally worded are likely to produce lower-level responses even from students who are capable of higher-level responses.
- *Items should be accessible to lower-level students.* Items must contain vocabulary and situations that are interpretable for students at all levels. It is a challenge to construct items that are valid at all 4 levels (Jin & Anderson, 2012b), so our goal for this assessment was to develop items that elicited responses at Levels 2-4. Therefore our goal is that every item

should give students an opportunity to choose between force-dynamic and scientific discourse, while signaling a preference for scientific discourse.

- *Items should be scaffolded enough to meet the goals above but not too much*. We need to design items that provide sufficient scaffolding as to indicate scientific reasoning is required but that are not over-scaffolded to enable students with lower-level knowledge to produce higher-level responses. In particular we were concerned about (a) forced choice items for which students may be able to recognize higher-level responses that they are incapable of producing, and (b) questions that can be answered through procedural display, such as producing memorized chemical equations whose meaning the students do not understand.
- *Items should be efficient to assess*. This means in particular finding an appropriate balance between easily-scored forced choice questions and short-answer questions that elicit performances closer to our learning progression construct, which focuses in particular on explanation and prediction practices (Figure 2).

Figure 3 contains an example item, OAKTREE, which meets these criteria. It is a multiple forced choice plus open explanation item that asks students to explain where the mass of the plant comes from. OAKTREE elicits a range of responses, words like mass and atoms and molecules cue student who are able to provide higher level responses to provide them, and the familiar scenario of tree growth allows lower level students to provide answers.



Like all materials, the wood of a large oak tree is made of atoms. There were some atoms in the original acorn that the oak tree grew from.

Where do you think the additional atoms came from?

a. ALL of the additional atoms were originally outside the tree,

b. SOME of the additional atoms were made by the tree as it grew.

Circle the best choice to answer each question about possible sources of mass from outside the tree.

How much of the dry mass comes from the AIR?	All or most	Some	None
How much of the dry mass comes from SUNLIGHT?	All or most	Some	None
How much of the dry mass comes from WATER?	All or most	Some	None
How much of the dry mass comes from SOIL NUTRIENTS?	All or most	Some	None

Explain your choices. How does the oak tree gain mass as it grows?

Figure 3: Text of OAKTREE item on Carbon TIME assessment

2. Validity evidence based on response processes

This criterion focuses in particular on the Interpretation vertex of the Assessment Triangle (Figure 1). If we have items that elicit student responses covering all the levels and dimensions of the learning progression framework (Figure 2), we still have to ask how well those responses represent students' understanding of the underlying construct. For example, forced choice items are often excessively scaffolded: Students can recognize the "best answer" even if they don't really understand it. On the other hand, open explanation items are often not scaffolded enough: Students who are capable of writing higher-level responses assume from the wording of the question that a simpler response is sufficient. For this determination we rely mainly on students' responses; we need to empirically determine that the assessment items do elicit responses that can reliably and validly sorted into levels. To do this we must carefully construct scoring rubrics and compare students' written and interview responses.

Constructing scoring rubrics. Coding reliably is an important step to producing good measures of student understanding. Therefore we designed our scoring rubrics (see Table 3) to be clearly interpretable with little training and routinely achieve 90% exact matches with undergraduate coders. However, such measures of exact match are internal to a single item. They answer the question, how similar are different coders in applying the same rubric to the same set of responses?

To answer questions about validity of learning progression-based scoring rubrics, where we are making an argument that students consistently respond in a way described in our learning progression framework, we need to align item scoring rubrics to the learning progression framework and with each other. This is an iterative process which includes editing items, rubric construction and alignment, coding, IRT, and developing a shared understanding of the reasoning that students at each level are relying on to produce their answers (Yao, Berson, Ayers, Choi, & Wilson, 2010). This type of validity evidence is based on a qualitative judgment by researchers—not necessarily based on statistics or cut-off values.

In Table 3 you can see the final scoring rubric for OAKTREE with specific indicators of Level 2, 3 and 4 reasoning aligned with the general level descriptions. You can also see student exemplar responses matched to each level. For example, in the level 2 exemplars, students are not describing changes in mass of the tree as due to the movement of atoms into or out of the tree. Whereas, Level 4 responses trace mass changes to the movement of atoms and rearrangement of atoms into new molecules.

Table 3. OAKTREE Scoring Rubric

	General Description	**OAKTREE Specific Description**	**Student Exemplars**
Level 4	Correct descriptions of matter and energy transfers through processes at multiple scales (from molecules and cells to organisms and ecosystems).	Indicators: 1. Explain that most of the oak's dry matter comes from air/CO_2 (might include other complementary sources like water and soil, but definitely recognizing CO_2/air) with no portion of them becoming energy OR 2. Trace the carbon (and maybe complementary sources like water and soil, but definitely recognizing CO_2/air) into part of tree's weight, glucose, or other organic molecule after undergoing photosynthesis/metabolic process with no portion of them becoming energy	(4.1) I choose that the mass comes from the air and some from the soil because, most or all of the mass comes from carbon which is in the air, but some also comes from the nutrients in the soil, but not a lot does. No mass comes from the water or sunlight. (4.2) The tree takes in air, water, sunlight, and nutrients. It then converts the air and water into food which gives it mass.
Level 3	Attempts to describe matter and energy transfers, but with errors (such as confusing matter and energy, forgetting to account for the mass of gases).	Indicators: 1. Explain tree's matter comes from air/CO_2, but with portion of tree mass coming from sunlight (energy to matter conversion) OR 2. Describe photosynthesis as contributing to weight gain (do not need to use the word photosynthesis. i.e., photosynthesis or sunlight as making food, glucose, or sugar) without tracing CO_2/air into the plant. OR 3. Explain that enablers transform into cells/tree structure/cellulose (must show understanding that enablers become tree)	(3.1) The tree took air and sunlight and through photosynthesis changed it to glucose which gave the tree its mass. (3.2) The oak tree completes photosynthesis and creates carbohydrates to help it grow. (3.3) An oak tree has cells, cells multiply and make matter. Some of the oaks matter comes from the environment (like dirt and/or water), but the main increase in mass comes from growth and development

Level 2	Elaborated force-dynamic accounts: Students' accounts continue to focus on actors, enablers, and natural tendencies, but they add detail and complexity, such as the idea that different organs have different functions.	Indicators: 1. List various enablers (e.g., sunlight, soil nutrients, water, air) as contributing to plant growth, but does not differentiate air from other enablers as a main source for matter OR 2. Explains that the tree will gain mass because it grows and/or becomes larger.	(2.1) An oak tree gains mass as it receives sunlight and water. It also gains mass as it receives minerals from the soil surrounding its roots. (2.2) As the oak tree grows, the larger the tree will become and that will cause the mass to increase.

Comparing written responses with interviews. Comparing written responses with interview transcripts provides evidence for written items eliciting responses that are aligned with student learning progression level as we consider our interview analysis as the gold standard for classifying student understanding.

Using interviews of students from Cohort 2 we are able to make two types of interview-written comparisons. First we can look at overall performance on the written assessment and compare it to a students' overall interview learning progression level (Figure 4). While there is some spread, the correlation demonstrates that students' performance on the interview and assessments are highly related (Spearman rank correlation = 0.81, *p*>0.01, *n*=49).

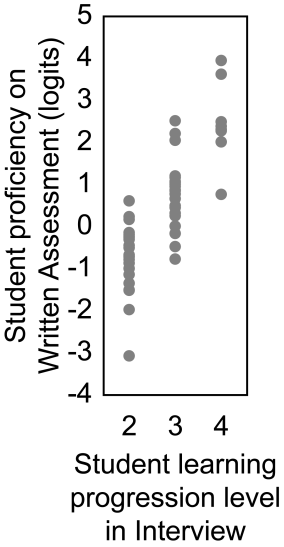
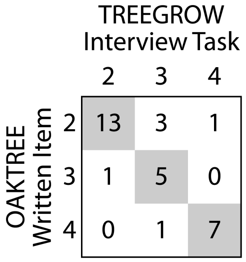


Figure 4: Relationship between student proficiency measured by interviews and written assessments

In addition to looking at overall performance, we can investigate students' responses on similar interview and written items. Figure 5 and Table 4 report on one such comparison. The teachers interviewed 31 students who had the OAKTREE written item on their assessment form. Of those 31 students, 81% had responses at the same learning progression level for the written item and interview task (gray cells of Figure 5). You can see in the first three rows in Table 5 (Students A-C) examples from these 25 students that demonstrate good alignment between the written and interview items. The last two rows (Students D-E) provide examples of comparisons for the 20% of students whose written and interview levels did not align. When students were misaligned they often added more detail to their answer in an interview, such as Student D, who described how glucose was made in his interview and revealed his misunderstanding of chemical energy. Student E demonstrates another common reason for misalignment, a student being vague in writing but giving a rich description when prompted in an interview. While the *Carbon TIME* written items sometimes over- or under-estimate a student's true performance, generally due to students' brevity in writing, the interview-written comparisons provide good evidence for the *Carbon TIME* written item validity.



	TREEGROW Interview Task		
OAKTREE Written Item	2	3	4
2	13	3	1
3	1	5	0
4	0	1	7

Figure 5: Comparison of learning progressions levels for interview and written student responses to similar items, TREEGROW and OAKTREE

Table 4: Comparison of student responses to a similar written and interview items.

Student	**Response to OAKTREE Written Item**	**Response to TREEGROW Interview Task**
A: Consistent Level 4	*Level 4:* The tree needs the CO_2 in air to perform photosynthesis, sunlight gives energy, but no mass. Water provides some mass, while soil nutrients provide very little.	*Level 4:* the soil only has minerals that are not used for growth….They're taking in the carbon dioxide which is part of photosynthesis. … It's a process where they're using the carbon dioxide atoms and sunlight and water to make glucose so they can grow…. The sun provides the energy that's needed to perform photosynthesis but doesn't directly give them any mass.
B: Consistent Level 3	*Level 3:* atoms from other objects turn into the tree	*Level 3:* Maybe it like helps to like give the tree more mass because it couldn't grow if it didn't have things that are going into it to like give it more material to make more of itself I guess.
C: Consistent Level 2	*Level 2:* The soil nutrients act as food for the oak tree which makes the oak tree gain mass.	*Level 2:* The growth of the tree because it grows. Like everything grows and I think it might be in the branches if that's what you're trying to ask me. Like the leaves grow. But I'm not sure exactly where it comes from….Fertilizer.
D: Inconsistent higher level	*Level 4:* An oak tree gains mass by absorbing CO_2 in the surrounding air and using it to build new molecules like the walls of the cell. There is really not water in the dry mass of the tree. Also light energy does not contribute to the mass of the tree, just its chemical energy. Some of the oak trees mass comes from nutrients in the soil, which the tree absorbs through water.	*Level 3:* It needs oxygen, carbon dioxide, some nutrients; nitrogen, phosphorous, and it's mostly made up of carbon dioxide which it gets from the CO2 and air.… [The tree] absorbs the light energy from the sun and uses that power turning the other atoms into chemical energy in the form of glucose.
E: Inconsistent lower level	*Level 2:* something to grow and I did an experiment that showed that plants do not need sunlight to grow.	*Level 3:* with carbon dioxide and water, basically when they make into a chemical reaction which would create glucose which is $C_6H_{12}O_6$ … it would create food for the tree… doesn't need any sunlight to come down for it, even though we know photosynthesis but that doesn't really matter right now at that moment…I've talked about glucose and it won't really need that sunlight. Mostly the sunlight is just maybe to maybe it will be absorbed to make it feel better or something like that but they just, it just doesn't really need that for that moment.

3. Validity evidence based on internal structure: Item response theory analyses

Validity evidence based on internal structure is also concerned with the "interpretation" vertex of the assessment triangle (Figure 1), but at a larger scale. The first two kinds of validity evidence focused primarily on qualities of individual items. Evidence based on internal structure focuses on how well the individual items "fit together" to produce a more general picture of the proficiency of an individual student or of the students in a class. We use IRT models to examine internal structure validity, specifically Rasch-family models. The use of such models allows us to empirically evaluate the relationship between the learning progression level scores on the items, and the overall person proficiency distribution, as well as to evaluate the fit of the model to the data at the individual item and student level. In this section we use these models to address three kinds of questions:

- How well individual items and individual students fit within the overall structure of our measurement model and align with one another?
- To what extent can we use assessment results to distinguish among the explanation practices of individual students in terms of learning progression levels? (Research Question 1)
- To what extent can we use assessment results to distinguish among class-size groups of students in terms of the learning progression levels of students and learning that takes place in different classes taught by different teachers? (Research Question 2)

Item and student fit to the measurement model

Rasch-family models provide several ways to analyze individual items and students in comparison with one another and in the context of the model as a whole, including Wright maps, item fit statistics, and person fit statistics. We discuss each below.

Wright maps. Figure 6 shows a Wright map based on the learning progression aligned open explanation-based items given in the *Carbon TIME* Assessments. In these data, item responses were coded into Levels 2 through 4. The red dots indicate the thresholds from Level 2 to 3 for the items, and the blue dots from Level 3 to 4. Items are arranged from empirically easiest overall, to most difficult.

[

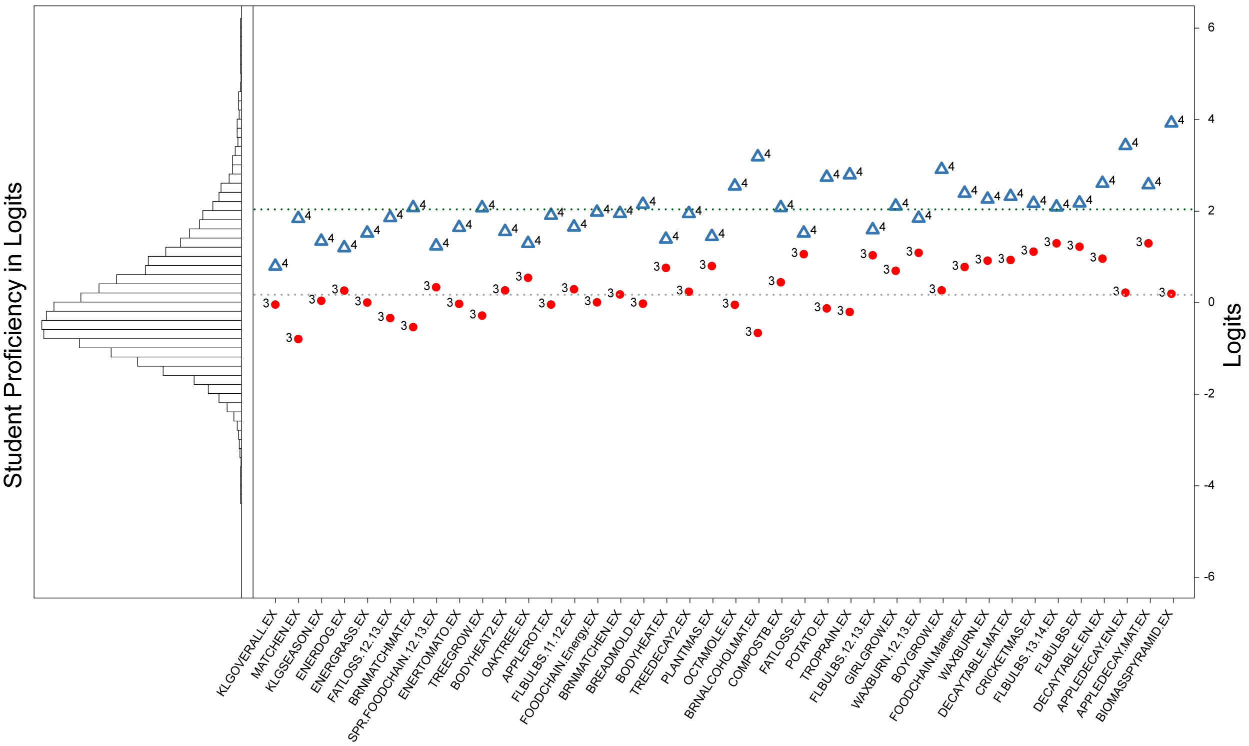


Figure 6: Wright map showing learning progression level thresholds for the explanation-based learning progression level

In general, most of the thresholds from Level 2 to 3 are in the same logit range, and well separated from the thresholds from Level 3 to 4. This implies that our theoretical predictions about students' thinking are mostly well-supported by the data. However, there is overlap at the extremes, such that the thresholds from Level 2 to 3 for the most difficult items overlap somewhat with the thresholds from Level 3 to 4 for the easiest items. Interestingly, patterns can be observed in the difficult items (most of which have to do with decay or large scale carbon cycling), and the easiest items (many of which provide more scaffolding). Observation of such patterns can lead to revision of items (e.g., rewriting the easy items to be less scaffolded), or to new understandings of student cognition (e.g., that understanding of decay is particularly difficult for students in middle and high school).

There are also several items for which the thresholds from Level 2 to 3, and Level 3 to 4, are not well differentiated (i.e. they have little vertical distance between them). This means that relatively few students are scoring at the middle Level, Level 3. This may mean either that attaining a Level 3 is difficult enough that, once students have attained this level, it is only a small step to attain Level 4; or alternatively, that attaining Level 4 is relatively easy, and thus once students have attained Level 3, they very quickly attain Level 4.

The majority of the students in the distribution at the left-hand side of Figure 6 fall at or below the Level 2 to Level 3 threshold for most of the items. This implies that the most common responses are either Level 2, or Level 3, and that Level 4 responses are uncommon. The distribution is slightly right-skewed, indicating there are some students who are at or above the Level 3 to Level 4 thresholds for most items. These are the students who are most likely to give Level 4 responses to most of the items. There are some items for which a Level 4 response may be difficult even for these students, however. The most difficult threshold – associated with the item labeled BIOMASSPYRAMID, is interesting. This item (see Supplementary Materials for full text) asks students about why there is a greater biomass of producers than herbivores than carnivores. This item has a relatively easy 2 to 3 threshold, many student are able to describe that there is less food available as you move up the food chain, which is the Level 3 indicator, but very few students can describe the loss of matter to carbon dioxide via cellular respiration as the reason there is less matter available, the Level 4 indicator for this item.

Item fit statistics. A second set of tools for examining the internal structure of the assessments is fit statistics. Fit statistics provide information about how well the data for an individual item or an individual student's performance are represented by the model we have chosen (i.e. the modification of the partial credit model). Fit statistics provide a measure of the random variation that is present in the data for a particular item or person. As this is a probabilistic model, some amount of randomness is expected. Items that show less random variation than expected have low fit statistics, and are usually not a concern. However, items with more random variation than expected have a relatively large number of low-performing students are doing better than expected on this item, and/or a relatively large number of high performing students doing worse than expected. This may indicate several possible issues with the item: it may be measuring another dimension in addition to the one of interest (e.g. perhaps it has a large reading load or requires more advanced math skills); it may be affected by current events (e.g. perhaps the issue addressed by the item has been in the news recently); perhaps the wording leads many students to misinterpret the item, and so forth.

Figure 7 shows a graph of the overall item fit statistics for the Carbon TIME open explanation items. Each dot represents a fit statistic, calculated from the empirical data, for one item. The gray area represents control limits (0.75~1.33) for the fit statistics (based on Adams & Khoo, 1996); dots inside the gray area show acceptable fit. In this case, only two items fell outside the control limits and may have some fit issues that should be investigated. When we examined the items, the multiple choice for APPLEDECAY, and TREEGROW, and student performances on the item, to see if we could find any problems that should be addressed, we did not uncover any issues. However, just as in any set of statistics based on random variation, one or two items out of a large group of items showing significant random variation should not be cause for concern.

In addition to the item fit information provided by IRT, we have also calculated item-total correlations (point-biserial correlations for dichotomous items, Pearson correlations for polytomous items). These are provided in supplementary materials. Item-total correlations are used to determine item fit in classical test theory; they measure the relationship of a given item to the rest of the test. Various minimum standards for item-total correlations have been suggested. One limitation with such statistics is that they will tend to be low for items that are either very easy or very difficult (in essence, these items show little variation, and hence there is a ceiling on the size of the correlation they can show). However, IRT fit statistics often show that such items have acceptable fit, and in an assessment that is to be used for pretest-posttest analysis, it is a good idea to have some items that are very easy, and some that are very difficult, to capture the full range of proficiencies in both the pretest and the posttest.

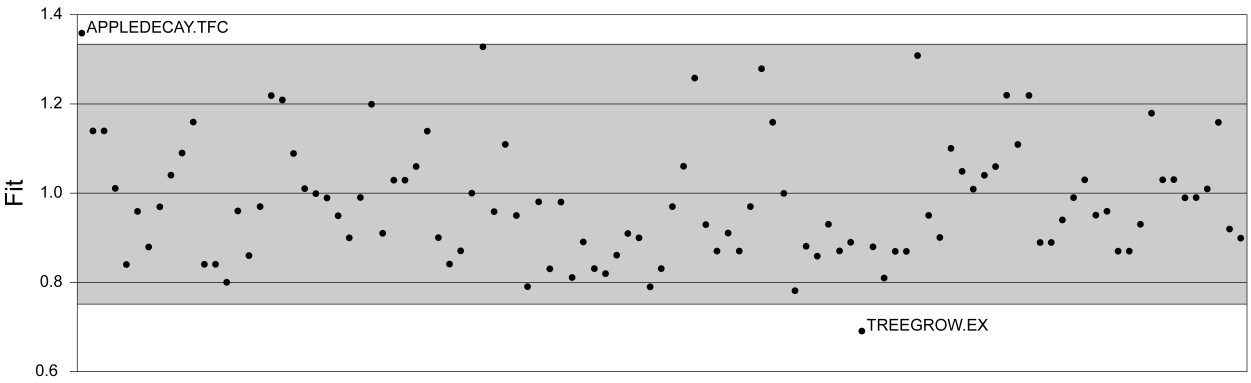


Figure 7: Plot of item fit

Person fit statistics. Just as fit statistics can be calculated for each item, they can also be calculated for each person. The interpretation of such statistics is similar to those for items; low fit statistics indicates sets of responses that are very regular; high fit statistics indicate more random variation than expected.

A distribution of fit statistics across all assessment events for students (*n*=7,389) is shown in Figure 8. In this figure, students with fit statistics below the left red line (the lower bound of the acceptable range) show very regular responses. It can be seen that a large proportion (38.7%) of students falls in this range; this is additional evidence of the consistency of student reasoning across a wide variety of contexts; since the items were designed to perform in this way, this is not cause for concern. In addition, the distribution of fit is right-skewed; 10.5% percent of students fall above the right red line (the upper bound of the acceptable range). These students are showing more random variation in their responses than expected. More research should be done to determine what might be causing this phenomenon. In particular, interviews for a number of these students about what they are thinking as they answer items might be useful.

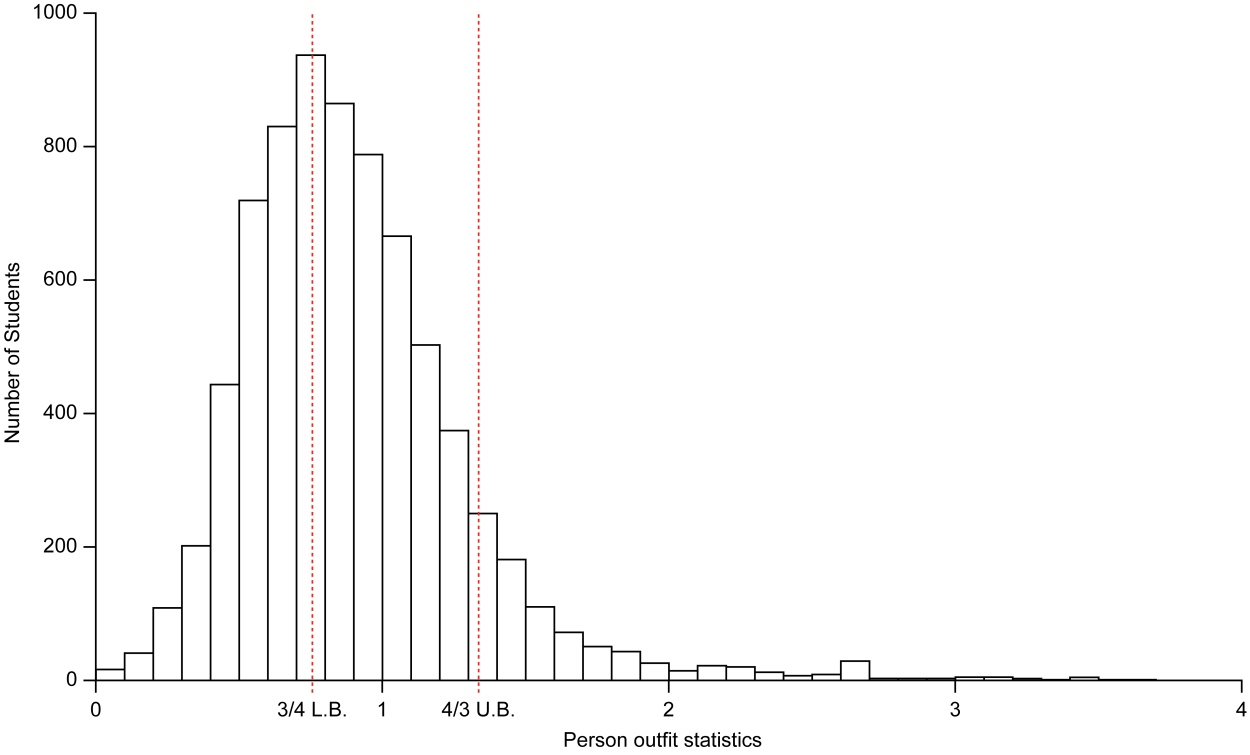


Figure 8: Plot of person outfit

Research Question 1: Assigning Learning Progression Levels to Students

A detail to note on the Wright map (Figure 6) is the horizontal lines drawn through the sets of item thresholds. These represent “cut scores” between the learning progression levels, based on the empirical results of our data analyses. We have chosen to put them at the median thresholds for each transition (the median of the thresholds between Level 2 and 3, and between Level 3 and 4). These median thresholds serve as the cut points for each transition and allow us to characterize students' overall performances in terms of their learning progression levels. Students below the first cut point are most likely to perform at Level 2 for most items; students between the two cut points are most likely to perform at Level 3, and students above the second cut point at Level 4. This allows us to characterize students in terms of their most likely performance, rather than simply the raw score on the items they took (without considering the difficulty of that particular set of items).

One issue that must be considered when determining which learning progression level to assign to a student is the issue of measurement error. Although we know the responses a student gave to a particular set of items, we also know that these responses may be affected by a number of issues other than simply the student's overall proficiency. These may include issues affecting the student (e.g. tiredness), issues affecting the scorer (e.g. distraction), and issues involving the particular item (e.g. the student misunderstanding the item text). We therefore need to take this measurement error into account. We do this using a standard error associated with each person's proficiency estimate. This functions very much like the standard error of the mean in classical statistics – i.e. there is a 95% probability that the true but unknown proficiency value lies within a 2 standard error margin around the estimate (assuming a normal distribution). Especially for students whose proficiency estimate lies near one of the cut points, this may affect their assignment to a learning progression level.

This is illustrated in Figure 9. This figure shows a modification of the person distribution shown on the left side of the Wright map for baseline and posttests for students of teachers in Cohort 3, still using the logit scale as the vertical axis and including the vertical lines as cut points between levels (explained above). The curved line illustrates the person proficiency estimates, ordered from smallest to largest, on the open explanation and forced choice items (red portions indicate baseline data; blue portions indicate posttest data). We use the combination of forced choice and open explanation items together as the gold standard in this figure, as the combination produces the highest internal consistency reliability, and thus reduces the standard errors of the person estimates.

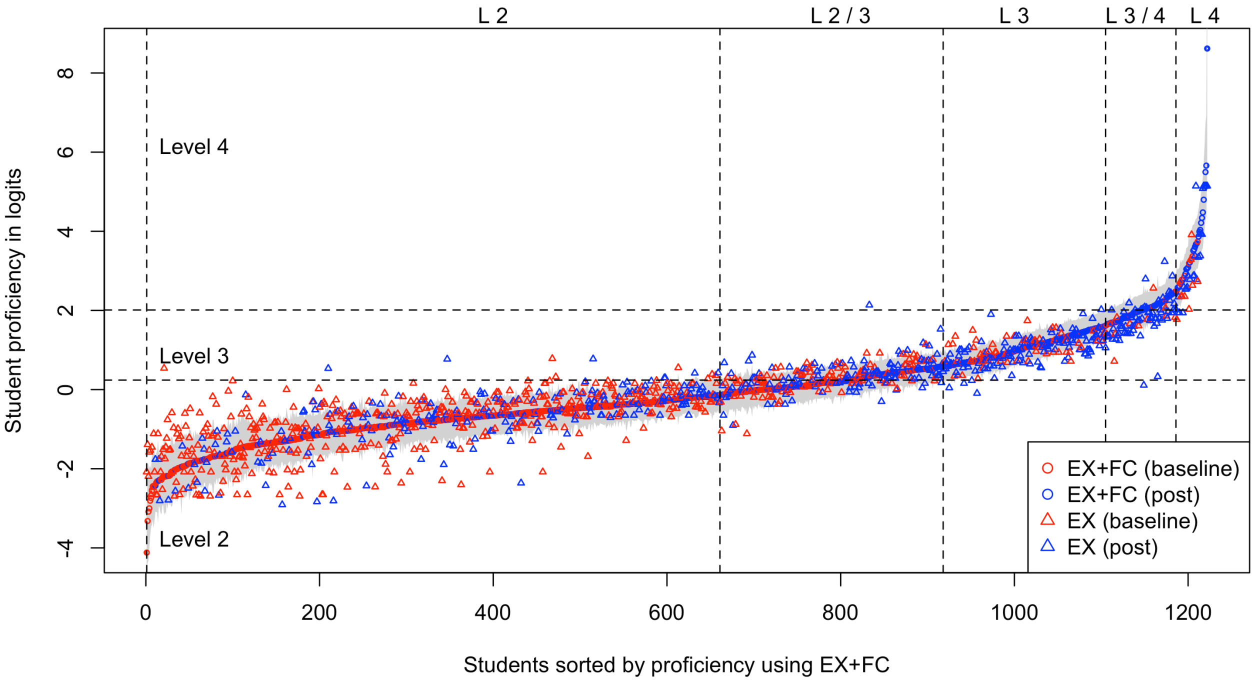


Figure 9: Modified Wright map showing student proficiency estimates, sorted from lowest to highest (using both explanation [EX] and forced choice [FC] items), along with standard error bands and explanation (EX) estimates

The gray band around this line illustrates the 68% confidence intervals, assuming a normal distribution (i.e. one standard error in each direction) for all of the proficiency estimates. The gray band shows some fluctuations depending on the response pattern of the persons. For example, if a person has not responded to one or more of the items on the test, the standard error band will be larger for that person. As illustrated by the vertical lines, sometimes the confidence intervals fall entirely in one of the learning progression regions (for example, the first approximately 675 student estimates fall entirely in the Level 2 logit region), while sometimes the confidence intervals overlap one of the cut points (for example, from about student 676 to about student 925). In the latter case, we are less certain about the learning progression level into which a student should be classified – students 676 to 925 may be at learning progression

Level 2, but they may also be at learning progression Level 3. Figure 9 also contains triangles; these are the estimates of student proficiency for students based on their performance on only the corresponding constructed-response explanation items (again, red are baseline data, blue are posttest).

Figure 10 is similar to Figure 9, with the same red and blue circles and shading, only now the red and blue triangles illustrate performance only on the forced choice items (using the same, anchored item difficulties as Figure 9). It is clear from comparison of the locations of the triangles in the two figures that estimates of student proficiency based only on the forced choice items deviate more from the overall estimates than estimates based on only the open explanation items, particularly at the higher end of the person proficiency distribution. This likely contributes to the drop in the variance of the person distribution that has been noted when comparing student performance on open explanation items to forced choice items (Draney & Shin, 2014). Therefore, the use of exclusively forced choice items to assign students to learning progression levels entails a higher degree of uncertainty, and should only be done in low-stakes settings that include additional information about student level, such as formative assessment in classrooms.

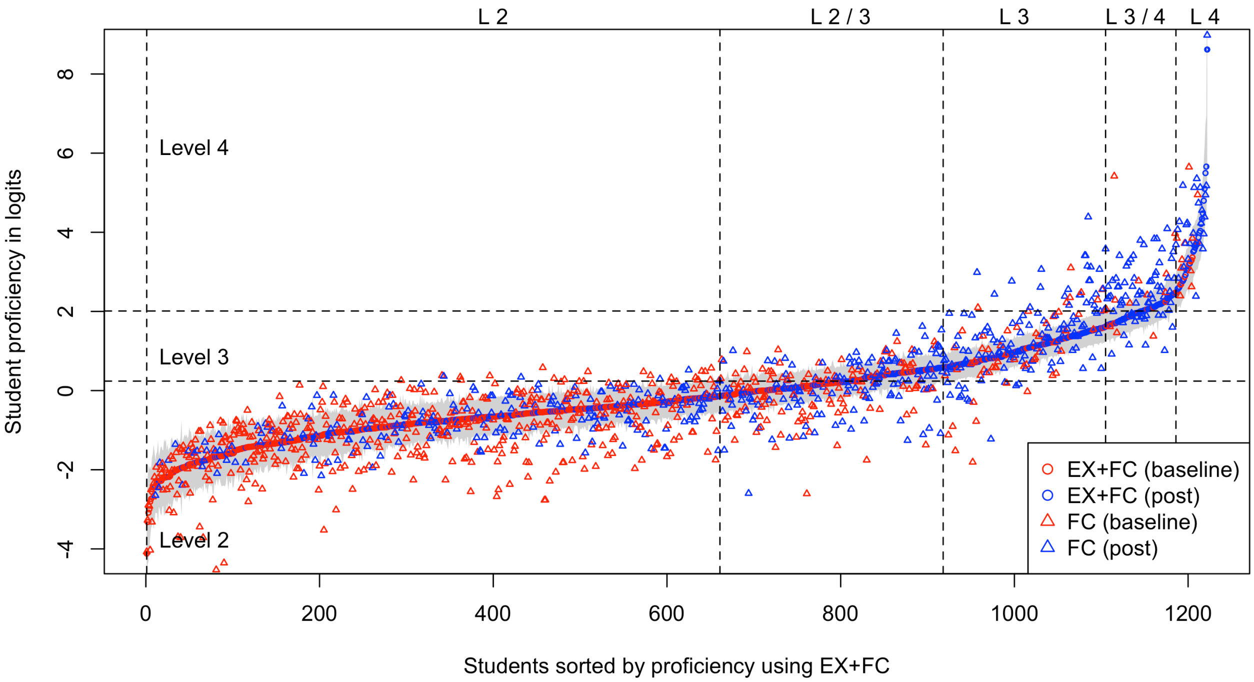


Figure 10: Modified Wright map showing student proficiency estimates, sorted from lowest to highest (using both explanation [EX] and forced choice [FC] items), along with standard error bands and forced choice (FC) estimates

Research Question 2: Assessing learning in classroom-sized groups of students

We can also summarize individual students' proficiency estimates into classroom-sized groups and assign these classrooms to a learning progression level. Figure 11a-c displays the average classroom performance, with 95% confidence intervals, for Cohort 3 teachers who were using the *Carbon TIME* units for the first time. The figure shows three time points: a) "Baseline," at the end of the school year the year before the teachers participated in the *Carbon TIME* program, b) "Pre", at the beginning of the school year the teachers participated in *Carbon TIME*, and c) "Post", at the end of *Carbon TIME* instruction. As in the case of individual students in Figures 7 and 8, the lower horizontal dotted line indicates the Level 2/3 cut point and the upper dotted line represents the Level 3/4 cut point. Figure 11a shows that only 3 of 25 classrooms had an average student proficiency for which the entire 95% confidence interval lay in the Level 3 band at the end of the school year before *Carbon TIME* implementation, while the other 22 classrooms were at Level 2 or on the border. Figure 11b indicates very similar results for beginning of year proficiency in these same classrooms, here only one classroom had an average student proficiency for which the most of the 95% confidence interval lay in the Level 3 band. You can contrast these proficiencies with Post Carbon TIME instruction scores in Figure 11c: 9 classrooms have means and confidence intervals which have moved into the Level 3 range and one entirely into the Level 4.

We can calculate student learning gains from pre to post test, and then calculate the average student gain score in each teacher's classroom, with appropriate standard errors. Such a graph is shown in Figure 11d, with teachers ordered from lowest to highest in terms of the average learning gains of the students in their classrooms. Learning gains were calculated using only students with matching pre and posttests. In this panel, rather than show the cut points between Levels 2/3 and 3/4 (as this graph shows learning gains and not simple performance), we have shown horizontal lines at zero, and at the average learning gain for the entire group (about 1.0 logits). Thus, we can see that the 4 classrooms at the far left have learning gains that are indistinguishable from zero. 8 of the 25 classrooms, shown to the far right, have learning gains that are statistically higher than the overall average. Such an ordering allows us to investigate the possible reasons for more or less success in different classrooms.

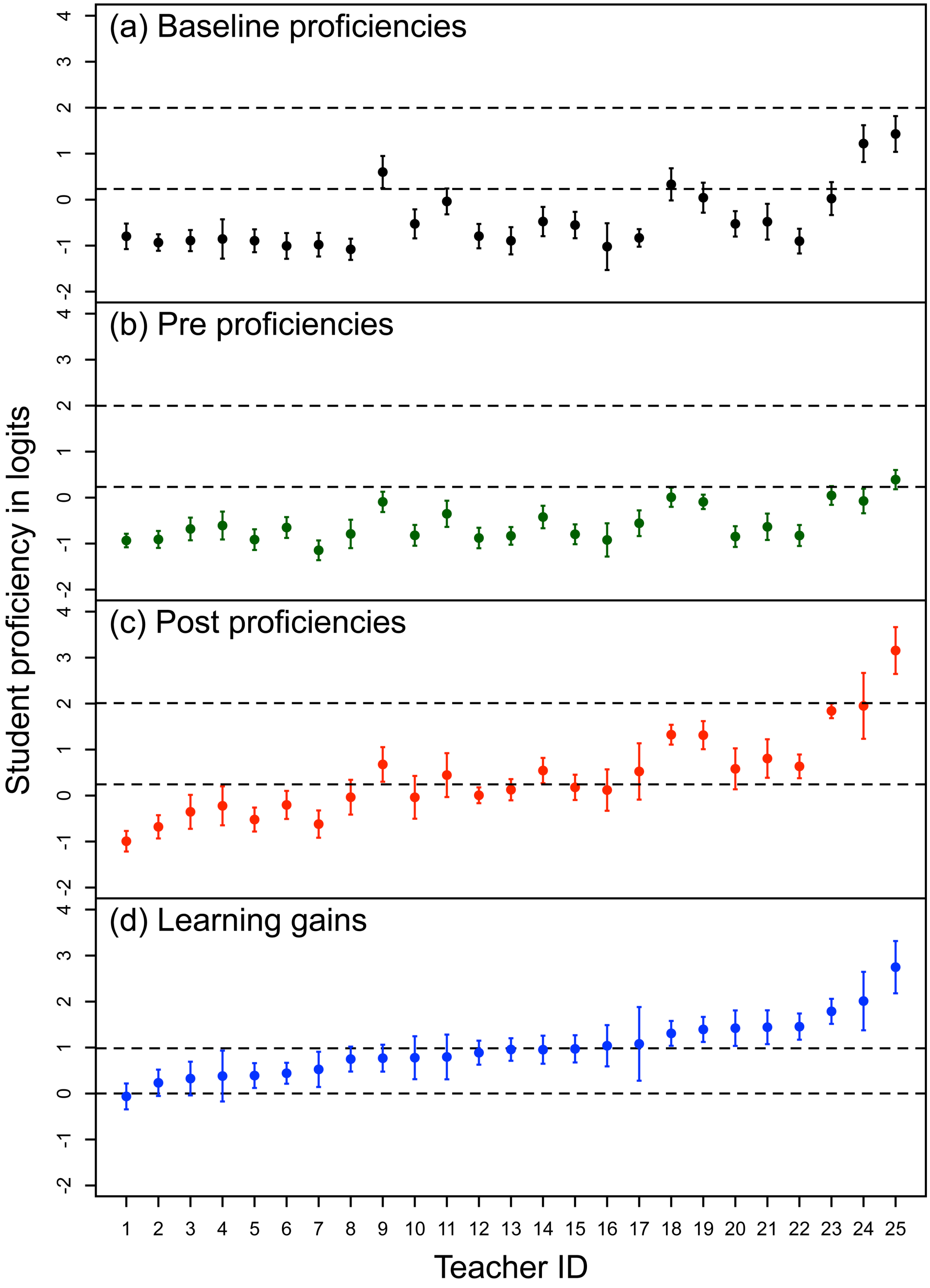


Figure 11a. Classroom learning gains on the Carbon TIME assessment (pre v. post) for the same classrooms. The lower dotted line indicate no learning gain and the upper dotted line represents the average learning gain. 11b-d. Classroom-level student performance on the

Carbon TIME assessment for Cohort 3 classrooms of teachers that did not participate in Cohorts 1 or 2 at three time points: a) “Baseline,” at the end of the school year the year before teachers participated in the Carbon TIME program, b) “Pre”, at the beginning of the school year teachers participated in Carbon TIME, and c) “Post”, at the end of Carbon TIME instruction. The lower and upper dotted lines indicate the Level 2/3 and Level 3/4 cut scores. Baseline students are different from those in pre and post. Data are plotted as means with 95% CI.

4. Validity evidence based on external variables: Comparison with AAAS items

We embedded 11 items about matter and energy, 5 per form with some as linking items, from the AAAS Project 2061 Science Assessment into the Carbon TIME assessment in order to collect validity evidence based on external variables. The AAAS Project 2061 Science Assessment items are designed to assess students’ conceptual understanding of key science ideas as well as test for common misconceptions and alternative ideas (http://assessment.aaas.org/). They are well designed items that have undergone rigorous and extensive review. Although these items have been designed to measure key ideas in science, they have not been specifically designed to target the levels of the learning progression in which we are interested; they should correlate well with our measure but would not provide the additional learning progression information we want.

An assessment should be well correlated with other measures that are designed to measure similar things. In the multidimensional item response theory analysis we obtained the latent correlation between the *Carbon TIME* items and the AAAS items assessments. The correlation between the scores for students on each assessment is 0.60, an indication that the two sets of items are indeed measuring related abilities, but not the same.

5. Validity evidence based on consequences: Learning gains from *Carbon TIME* project

The final form of validity evidence focuses on consequences of using the assessments: Does using the assessments lead to better educational outcomes? It is on this criterion that evaluations of current science monitoring assessments such as the NRC reports NRC, 2005, 2014) render some of their harshest criticisms, arguing that the focus of these assessments on a restricted range of student performances scored in mechanical ways encourages a focus on test preparation at the expense of broader goals of education.

The *Carbon TIME* monitoring assessments are part of an instructional system which includes the assessment, learning progression framework, curriculum with embedded formative assessments and professional development. Figure 12 shows evidence about the consequences of using this system for the same sample of students as in Figures 8, 9, and 10. At the end of a year of traditional instruction (Baseline, *n* = 667) and before *Carbon TIME* instruction (Pre, *n* = 574) mean student performance is below the Level 2/3 cut point indicating students are on average reasoning at Level 2.

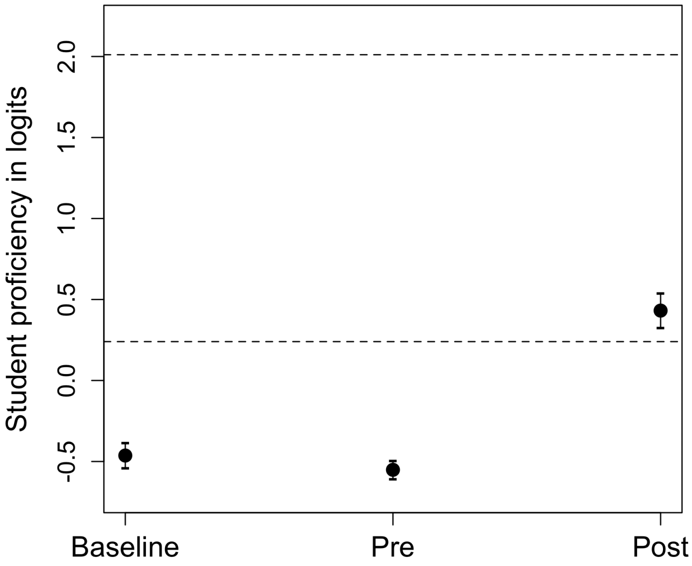


Figure 12. Program-wide student performance on the Carbon TIME assessment at three time points: "Baseline," at the end of the school year the year before teachers participated in the

Carbon TIME program, “Pre”, at the beginning of the school year teachers participated in Carbon TIME, and “Post”, at the end of Carbon TIME instruction. The lower dotted line indicates the Level 2/3 cut off and the upper dotted line represents the Level 3/4 cut off. Data are plotted as means with 95% CI.

Compare these performances to performance of the same students after *Carbon TIME* instruction (Post, *n* = 574), where students, on average, are reasoning at Level 3. The error bars show 95% confidence intervals. Students who experienced Carbon TIME instruction perform significantly better than 1) they did at the beginning of the school year (*t*=19.78, *p*<0.001, effect size=0.94) and 2) students in the same teachers’ classes from the year before (*t*=13.26, *p*<0.001, effect size=0.77). While these differences are significant and the effect sizes are large, you can see from the relationship of the Post *Carbon TIME* performance and the Level 3/4 cut off line we still must make additional progress before the majority of students achieve Level 4 reasoning, the level of reasoning described in the NRC *Framework* and the *Next Generation Science Standards*.

Discussion

This article begins with a discussion of what Shepard (2013) described as dual purposes of assessment: accountability and learning. Penuel, et al. (2014) similarly discussed the need for “balancing the goals of supporting and assessing student learning.” Our current system has developed different kinds of tests for these two purposes.

- For accountability we rely on large-scale assessments designed with careful attention to matching the measurement model to responses and relationships with other large-scale assessments. These tests are “fair” in the sense that they can be administered to all students and scored efficiently; students’ scores can be reliably compared across test forms, years, and sites.

- For learning we have a variety of assessments that have been developed with careful attention to the nature of knowledge and practice that they assess and to cognitive and sociocultural models (the “Cognition” vertex of the assessment triangle—Figure 1). Learning progression-based assessments, including our own prior work (e.g., Alonzo & Steedle, 2009; Gotwals & Songer, 2013; Gunckel et al., 2012; Jin & Anderson, 2012b; Lehrer & Schauble, 2012; Mohan et al., 2009; Plummer & Krajcik, 2010; Rivet & Kastens, 2012; Shea & Duncan, 2013) are included in this category. These assessments have been shown to be valuable for the learning functions of assessment—they produce valuable insights into the nature of students’ knowledge and practice.

On the surface, this system of “different tests for different purposes” may seem reasonable, but as many critics, including the NRC reports cited above, argue that this system is dysfunctional. The mismatch between the knowledge and practices assessed, and between the cognitive models that guide development of the learning progression assessments and the statistical models that guide development of the large-scale assessment, mean that the assessments work at cross purposes.

So this article reports on a test of concept: The *Carbon TIME* units include formative and summative assessments that are used for individual lessons and units. Is it possible to develop an assessment system that is instructionally sensitive to specific curriculum units, and also includes monitoring assessments that can be used to assess the impact of the program as a whole? In this article we use the AERA/APA/NCME *Standards* (2014) to organize evidence evaluate the validity of *Carbon TIME* monitoring assessments for both accountability and learning purposes. We conclude with (a) a summary of our arguments for each of our three research questions and (b) a discussion of dilemmas or tradeoffs that we have faced in designing and enacting these assessments.

Summary of Findings for Research Questions

Evaluations of the items and scoring rubrics using the first two AERA/APA/NCME Standards—test content and response processes—provide strong evidence that the *Carbon TIME* assessments are useful for the learning purposes of assessment. They elicit student responses encompassing the full range of performances that we targeted for the assessments, and there is good alignment between students' performances on interviews and written tests (Table 4 and Figure 3). Evidence aligned with the third standard—internal structure—also supports positive evaluations of the items included in the assessments. The items fit the learning progression-based statistical model (Figure 7) and the scoring rubrics differentiate student responses consistently and reliably with respect to learning progression levels (Figure 8).

But what about the accountability purposes of monitoring assessments? How well do the *Carbon TIME* assessments differentiate more and less successful students, teachers, or programs? Ideally we would like assessments that produce both qualitative descriptions of learning progression levels and quantitative comparisons for individual students, for teachers or classrooms, and for the program as a whole. To what extent do the *Carbon TIME* monitoring assessments meet the AERA/APA/NCME *Standards* for claims made at each level?

Individual students.

To what extent can we use assessment results to distinguish among the explanation practices of individual students in terms of learning progression levels? There is substantial error in assigning students to a particular level. The proportion of students for who cannot be reliably placed in a level (those in sections for "Level 2 or 3", "Level 3 or 4" in Figures 8 and 9) is sizable. Hence, our ability to distinguish individual students in terms of learning progression level is limited. A possible method for reducing this error would be to increase the number of items each student takes, thus reducing the estimate error (Figures 8 and 9); however, we feel that this is not a practical solution; the *Carbon TIME* assessment already takes most students between 20 and 45 minutes to compete.

Figure 8 also shows that a number of students do not fit the statistical model predicting their performance on individual items. This suggests that for these students (about 10%) there are patterns in their responses that are not well predicted by statistical models based the learning progression-based frameworks and scoring rubrics.

On balance, the available evidence supports the notion that the tests of individual students can be useful for learning purposes: They provide useful insights into how individual students reason about carbon-transforming processes that can guide teachers and curriculum developers. However, there are significant numbers of students who are on the borders between learning progression levels, or who do not fit the model. Thus the *Carbon TIME* tests might be useful as part of a system for high-stakes purposes such as assigning course grades or determining student placements or rankings, but they would not be sufficient.

Classrooms. To what extent can we use assessment results to distinguish among teachers in terms of the learning progression levels of students in their class and the success of different teachers in helping students to advance to higher learning progression levels? Our results with respect to this research question are presented in Figure 11 and the accompanying text. They show that there are significant differences among classrooms in terms of student pretest performance, even larger differences in posttest performance, and significant differences in pre-post learning gains. Thus the assessments are useful for many learning purposes. For example, they can be used to identify classrooms where students are doing significantly better and worse than average. This type of monitoring could be used to identify effective teaching practices or to track individual teachers' progress over multiple years.

As with students, though, using these results for accountability purposes is more problematic, particularly if the goal is to evaluate the success of individual teachers. The error bars for each individual teacher are large enough that there are significant uncertainties in that teacher's ranking. Furthermore, these statistics say nothing about the many other factors (e.g., student grade level, proficiency in writing and mathematics, school-level support) that affect students' learning.

Programs. To what extent can we use assessment results to distinguish among programs, particularly the difference between the *Carbon TIME* curriculum and programs used by teachers before they used *Carbon TIME*? As Figure 12 shows, the assessments provide unambiguous evidence with respect to this research question. With large samples of students, the differences between baseline results (the same teachers using a different program) and *Carbon TIME* posttest results are many times larger than the error bars showing 95% confidence intervals. There is no doubt that students learned more of the knowledge and practices tested from *Carbon TIME* than they learned from the previous curricula.

This is, of course, in part evidence of successful alignment between the *Carbon TIME* curriculum and assessments: The curriculum taught what the assessments tested. Thus the data provide evidence for the consequential validity of the curriculum and assessment system, but not for the value of the knowledge and practices assessed. That requires a different kind of argument that is beyond the scope of this article.

Tradeoffs in the Design and Development Process

Our experience in developing and evaluating the *Carbon TIME* assessments leads us to conclude that although continuing improvement is possible on all the criteria, there are also inevitable tradeoffs in the design and development process—no assessment can serve all purposes well. In this section we discuss three of these tradeoffs.

Grain size for learning progression levels. There are striking differences among learning progression frameworks in terms of the precision in their descriptions of levels. Some frameworks are fine-grained, describing multiple levels of achievement and sometimes multiple strands associated with different topics or practices (e.g., Alonzo & Steedle, 2009; Lehrer & Schauble, 2012). In comparison, the *Carbon TIME* framework is coarse-grained, using just three levels to describe differences among students in a broad content domain, across a wide range of proficiencies.

Our contention is that learning progression-based monitoring assessments will inevitably need to be relatively coarse-grained. If the assessments are to be useful for students experiencing different curricula, then they will need frameworks describing learning trajectories that are not curriculum dependent. Furthermore, considerations of internal structure, in particular, tend to dictate fewer and more broadly defined levels. Penuel, et al. (2014) describe a development process similar to ours: the developers of a mathematical learning trajectory wanted a fine-grained framework for diagnostic precision (learning purposes), but they also wanted results that would enable comparisons among classrooms, teachers, and programs. Like us they found that they could not do both with same test.

Limitations of forced-choice questions. The contrast between Figures 9 and 10 has important implications for assessment design. In comparison with our best estimates of students' proficiency, estimates based on students' written explanations alone (Figure 9) were substantially more accurate than estimates based on students forced-choice responses alone (Figure 10). This was especially true for more proficient students taking the posttest. This result comes after many years in which we have experimented with and evaluated different approaches to constructing forced-choice items.

Our conclusion is this: *No assessment consisting of forced-choice items alone has been found to accurately assess students' levels on the Carbon Learning Progression Framework.* This is partly due to the nature of the practices we wish to assess: We are particularly interested in students' explanations and predictions, so the content validity of forced-choice items alone is always suspect. Choosing the correct explanation or prediction is not the same as generating it. It is also partly due to the nature of the Carbon Learning Progression Framework: Students who are not proficient in using scientific discourse can still identify scientific discourse when they choose among possible responses. These conditions prevail, however, in most current work on learning progressions and science assessment generally.

We note a couple of qualifications to this largely negative conclusion:

- Forced-choice items contribute significantly to the overall quality of the *Carbon TIME* assessments. In particular, we have found the multiple forced-choice plus explanation format to be powerful (see the OAKTREE example in Figure 3). The forced-choice questions both scaffold students' responses and make it clear what the students are explaining in their written explanation.
- We are encouraged by recent work, by our project and others, focusing on automated scoring of constructed responses (Haudek, Prevost, Moscarella, Merrill, & Urban-Lurain, 2012; Nehm & Haertig, 2011). It appears to us that with appropriate item design and construction of scoring rubrics, it will be possible to develop machine-scoring algorithms that are as accurate as trained human coders.

Aligning cognitive and statistical models. Current science monitoring assessments can achieve high standards with respect to student and item fit to model in part because the models are purely statistical; they are not associated with any notions about the meaning or significance of the students' responses. As noted above, building cognitive models such as learning progression frameworks into assessment systems is a primary reform recommendation of the NRC reports (National Research Council, 2005, 2014). This recommendation entails costs, though, for statistical fit. There are items that provide good information about students' learning progression levels that do not fit the model well, as well as items that align well with the statistical model but not the learning progression framework. Scoring rubrics need to align with the learning progression framework as well as differentiating student responses to a particular item. Thus we conclude that assessment systems built around cognitive models will generally be more difficult to design with high levels of statistical fit.

Conclusion

We conclude with a brief comment on the implications of this research for the role of assessment in science education systems. In the United States there is currently a movement to reform our school systems by using assessments for high-stakes accountability—for example, making school funding and teacher evaluations dependent on assessment results. The reform documents cited above argue convincingly that current science monitoring assessments are inadequate for these purposes. In this article we present a careful evaluation of a learning progression-based alternative using the AERA/APA/NCME *Standards.* Our evaluation of this evidence is that the system is clearly useful for learning purposes and for program evaluation, but that results comparing individual students or teachers can only be used with great caution. Thus rigorous evidence-based scrutiny leaves us doubtful of any system that relies on current science assessments as a primary tool for test-based high stakes accountability. We need to be aware of these limitations as we make policy and seek to improve practice in science education.

Acknowledgements

The authors would like to thank several people for their invaluable contributions to the work presented in this paper. We would like to acknowledge the contributions made by Kathryn Oleszkowicz, Emily E. Scott, Hannah Miller, David van Deurs, Etiowo Usoro, Courtney Lannen, Brian Grieves, Carly Atkinson, Liz Thompkins, Cara Morrison, Anthony Machniak and Nick Verbanic from Michigan State University, Jenny Dauer from the University of Nebraska, Jinnie Choi at Pearson and Tian Xia from the University of California, Berkeley. We would also like to thank Hannah Miller, Wendy Johnson, Jay Thomas and Mark Wilson for comments on earlier versions of this manuscript. This research is supported in part by grants from the National Science Foundation: Learning Progression on Carbon-Transforming Processes in Socio-Ecological Systems (NSF 0815993), and Targeted Partnership: Culturally relevant ecology, learning progressions and environmental literacy (NSF-0832173), and CCE: A Learning Progression-based System for Promoting Understanding of Carbon-transforming Processes (DRL 1020187). Additional support comes from the Great Lakes Bioenergy Research Center. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the United States Department of Energy.

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement*, *21*(1), 1–23. http://doi.org/10.1177/0146621697211001

Adams, R., & Khoo, S. T. (1996). Quest : the interactive test analysis system. Retrieved from http://works.bepress.com/ray_adams/36

AERA/APA/NCME. (2014). *Standards for Educational and Psychological Testing*.

Alonzo, A. C., Neidorf, T., & Anderson, C. W. (2012). Using Learning Progressions To Inform Large-Scale Assessment. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning Progressions in Science* (pp. 211–240). SensePublishers. Retrieved from http://link.springer.com/chapter/10.1007/978-94-6091-824-7_10

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education*, *93*(3), 389–421. http://doi.org/10.1002/sce.20303

Anderson, C. W., Miller, H. K., Johnson, W., Freed, A. L., Dauer, J. M., Doherty, J. H., … Scott, E. E. (2015). Carbon TIME (Transformations in Matter and Energy) Curriculum. Retrieved from http://carbontime.bscs.org

Baker, E. L. (2013). The Chimera of Validity. *Teachers College Record*, *115*(9), 1–26.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic Assessment With Ordered Multiple-Choice Items. *Educational Assessment*, *11*(1), 33–63. http://doi.org/10.1207/s15326977ea1101_2

Catley, K., Lehrer, R., & Reiser, B. J. (2005). *Tracing a Prospective Learning Progression for Developing Understanding of Evolution*.

Claesgens, J., Scalise, K., Wilson, M., & Stacy, A. (2009). Mapping student understanding in chemistry: The Perspectives of Chemists. *Science Education*, *93*(1), 56–85. http://doi.org/10.1002/sce.20292

De Boer, G. E., Dubois, N., & Herrmann-Abell, C. F. (2008). *Assessment linked to middle school science learning goals: using pilot testing in item development*. Presented at the National Science Teachers Association (NSTA) National Conference, Boston, MA.

Draney, K., & Shin, H. J. (2014). *An effect of item type on proficiency distribution.* Presented at the annual meeting of the National Council on Measurement in Education (NCME), Philadelphia, PA.

Furtak, E. M. (2012). Linking a learning progression for natural selection to teachers' enactment of formative assessment. *Journal of Research in Science Teaching*, *49*(9), 1181–1210. http://doi.org/10.1002/tea.21054

Gotwals, A. W., & Songer, N. B. (2010). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Science Education*, *94*(2), 259–281. http://doi.org/10.1002/sce.20368

Gotwals, A. W., & Songer, N. B. (2013). Validity Evidence for Learning Progression-Based Assessment Items That Fuse Core Disciplinary Ideas and Science Practices. *Journal of Research in Science Teaching*, *50*(5), 597–626. http://doi.org/10.1002/tea.21083

Gunckel, K. L., Covitt, B. A., Salinas, I., & Anderson, C. W. (2012). A learning progression for water in socio-ecological systems. *Journal of Research in Science Teaching*, *49*(7), 843–868. http://doi.org/10.1002/tea.21024

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J., & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid-base chemistry in introductory biology. *CBE Life Sciences Education*, *11*(3), 283–293. http://doi.org/10.1187/cbe.11-08-0084

Irribarra, D. T., & Freund, R. (2014). Wright Map: IRT item-person map with ConQuest integration, R Package "WrightMap" (Version 1.1). Retrieved from https://cran.r-project.org/web/packages/WrightMap/WrightMap.pdf

Jin, H., & Anderson, C. W. (2012a). A learning progression for energy in socio-ecological systems. *Journal of Research in Science Teaching*, *49*(9), 1149–1180. http://doi.org/10.1002/tea.21051

Jin, H., & Anderson, C. W. (2012b). Developing Assessments For A Learning Progression on Carbon-Transforming Processes in Socio-Ecological Systems. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning Progressions in Science* (pp. 151–181). SensePublishers. Retrieved from http://link.springer.com/chapter/10.1007/978-94-6091-824-7_8

Lehrer, R., & Schauble, L. (2012). Supporting inquiry about the foundations of evolutionary thinking in the elementary grades. In J. Shrager & S. Carver (Eds.), *The Journey from Child to Scientist: Integrating Cognitive Development and the Education Sciences* (pp. 171–205). Washington, DC, US: American Psychological Association.

Mislevy, R. J., Hamel, L., Fried, R. G., Gaffney, T., Haertel, G., Hafter, A., … Wenk, A. (2003). *Design patterns for assessing science inquiry (PADI Technical Report 1)*. Menlo Park, CA: SRI International.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus Article: On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3–62. http://doi.org/10.1207/S15366359MEA0101_02

Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching*, *46*(6), 675–698. http://doi.org/10.1002/tea.20314

National Assessment Governing Board. (2006). *Framework for the 2009 National Assessment of Educational Progress Science Test*.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

National Research Council. (2005). *Systems for State Science Assessment*. Washington D.C.: National Academies Press.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: National Academies Press. Retrieved from http://www.nap.edu/catalog.php?record_id=13165

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards*. Washington D.C.: National Academies Press.

Nehm, R. H., & Haertig, H. (2011). Human vs. Computer Diagnosis of Students' Natural Selection Knowledge: Testing the Efficacy of Text Analytic Software. *Journal of Science Education and Technology*, *21*(1), 56–73. http://doi.org/10.1007/s10956-011-9282-7

Neumann, K., Viering, T., Boone, W. J., & Fischer, H. E. (2013). Towards a learning progression of energy. *Journal of Research in Science Teaching*, *50*(2), 162–188. http://doi.org/10.1002/tea.21061

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. Achieve, Inc. on behalf of the twenty-six states and partners that collaborated on the NGSS.

Noble, T., Suarez, C., Rosebery, A., O'Connor, M. C., Warren, B., & Hudicourt-Barnes, J. (2012). "I never thought of it as freezing": How students answer questions on large-scale science tests and what they know about science. *Journal of Research in Science Teaching*, *49*(6), 778–803. http://doi.org/10.1002/tea.21026

NRC. (2007). *Taking Science to School: Learning and Teaching Science in Grades K-8*. (R. A. Duschl, H. A. Schweingruber, & A. W. Shouse, Eds.). National Academies Press. Retrieved from http://www.nap.edu/catalog.php?record_id=11625

Padilla, J.-L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136–144. http://doi.org/10.7334/psicothema2013.259

Pellegrino, J. W., & Chudowsky, N. (2003). Large-Scale Assessments That Support Learning: What Will It Take? *Theory Into Practice*, *42*(1), 75–83.

Penuel, W. R., Confrey, J., Maloney, A., & Rupp, A. A. (2014). Design Decisions in Developing Learning Trajectories–Based Assessments in Mathematics: A Case Study. *Journal of the Learning Sciences*, *23*(1), 47–95. http://doi.org/10.1080/10508406.2013.866118

Pinker, S. (2007). *The Stuff of Thought: Language as a Window Into Human Nature*. Penguin.

Plummer, J. D., & Krajcik, J. (2010). Building a learning progression for celestial motion: Elementary levels from an earth-based perspective. *Journal of Research in Science Teaching*, *47*(7), 768–787. http://doi.org/10.1002/tea.20355

Rivet, A. E., & Kastens, K. A. (2012). Developing a construct-based assessment to examine students' analogical reasoning around physical models in Earth Science. *Journal of Research in Science Teaching*, *49*(6), 713–743. http://doi.org/10.1002/tea.21029

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D., … Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching*, *46*(6), 632–654. http://doi.org/10.1002/tea.20311

Shea, N. A., & Duncan, R. G. (2013). From Theory to Data: The Process of Refining Learning Progressions. *Journal of the Learning Sciences*, *22*(1), 7–32. http://doi.org/10.1080/10508406.2012.691924

Shepard, L. A. (2013). Validity for What Purpose? *Teachers College Record*, *115*(9), 1–12.

Shin, H., Choi, J., & Draney, K. (2012). *Using item response theory models for classifying students onto levels of achievement.* Presented at the International Objective Measurement Workshop (IOMW), Vancouver, BC, Canada.

Shin, N., Stevens, S. Y., & Krajcik, J. S. (2010). Tracking student learning over time using construct-centered design. In S. Rodrigues (Ed.), *Using analytical frameworks for classroom research: Collecting data and analyzing narrative*. New York: Routledge.

Smith, C., Wiser, M., Anderson, C., & 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*, *14*(1-2), 1–98.

Steedle, J. T., & Shavelson, R. J. (2009). Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching*, *46*(6), 699–715. http://doi.org/10.1002/tea.20308

Talmy, L. (1988). Force Dynamics in Language and Cognition. *Cognitive Science*, *12*(1), 49–100. http://doi.org/10.1207/s15516709cog1201_2

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Lawrence Erlbaum Associates.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, *46*(6), 716–730. http://doi.org/10.1002/tea.20318

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, *13*(2), 181–208.

Wong, A. (2015, October 28). Testing, So Much Testing. *The Atlantic*. Retrieved from http://www.theatlantic.com/education/archive/2015/10/testing-testing/412735/

Wright, B. D., & Masters, G. (1982). *Rating scale analysis*. Chicago: Mesa Press.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2.0. Generalized Item Response Modelling Software*. Melbourne: Australian Council for Educational Research.

Yao, S. Y., Berson, E., Ayers, E., Choi, S., & Wilson, M. (2010, April). *The Qualitative Inner-Loop of the BEAR Assessment System*. Presented at the International Objective Measurement Workshop, University of Colorado, Boulder.