Examining Validity and Coherence in a Cognitively-Based

Science Performance Assessment

Audrey Rabi Steele Whitaker

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2022

Abstract

Examining Validity and Coherence in a Cognitively-Based Science Performance Assessment

Audrey Rabi Whitaker

The purpose of this research was to explore the coherence and effectiveness of an assessment approach that combined principles of cognitive-based assessment, performance assessment, and the Next Generation Science Standards. By drawing on research on learning progressions and cognition in geoscience to design, implement, and analyze an Earth Science performance assessment at the high school level, I explored the challenges and opportunities inherent in a cognitively-based science performance assessment system. The primary research question for this study was:

How do cognitively-based performance assessments promote coherence between students' understanding, responses, and scoring?

Four subquestions allowed me to compare observations of student thinking with written responses and scores across multiple modalities in order to characterize the overall coherence of the assessment system.

Using a design study approach, an assessment was developed using a two-phase process. First, a construct map was created that outlined a learning progression for each of four geology subdomains: geologic time & stratigraphy; surface processes; plate tectonics; and geologic maps. Second, the construct map guided the development of interconnected performance assessment tasks intended to elicit and measure student thinking within those geology subdomains. Twenty-two high school students engaged in a think-aloud protocol while completing the performance assessment.

Student responses from the performance assessment were scored according to a predetermined scoring procedure that generated scores on individual items as well as holistic scores for each construct. Data from student written responses and think-alouds were quantitatively coded in alignment with the cognitive model for the assessment system. I used these data to examine the correlations between student thinking, written responses, and scores, in both item-by-item and holistic modalities. The strength of these correlations varied by construct, but some overall patterns emerged:

(1) The design of this cognitively-based science performance assessment was successful in eliciting thinking about all four levels of each construct, and there were instances where student thinking went beyond the intended bounds of specific items.

(2) For comparisons of student thinking to written responses or scores, holistic values captured a similar or better level of correlation than individual items, pointing to the important role of holistic scoring in the interpretation phase of this assessment approach.

(3) The performance assessment produced scores for three out of four constructs with statistically significant correlations to student thinking. Together, these results show that fully capturing student thinking remains a formidable challenge for the assessment field, but that cognitively-based science performance assessment tasks have significant potential to reveal the extent and breadth of student thinking beyond traditional assessment approaches.

The findings in this study have implications for the ways in which different stakeholders in science education, including classroom teachers, curriculum writers, and education leadership, can harness the power of cognitively-based assessment tools to better measure and support student learning.

Table of Contents

List of Tables

List of Figures

Acknowledgments

The most important people I would like to thank for their contributions to this work are my former students. Not just those who donated their time and energy to participating in this research project, though I owe them a special debt of gratitude, but all the teenagers who have explored the Earth with me over the past 18 years. There are far too many of them to name, but they are without a doubt the reason I care so much about the intricacies of science teaching, learning, and assessment. In particular, the Young Writers classes of 2010, 2018, and 2022 were instrumental in shaping me as an educator and validating my commitment to this research study.

I am grateful to the adult friends and colleagues who have supported and encouraged my professional growth in this field. This includes my current and former administrators, Carolyn Yaffe, Courtney Winkfield, AAden Stern, and Tanisha Brown; fellow STEM educators Tara Litzenberger and Corinne Cornibe; and astronomers Eric Jensen and David Cohen. Without their contributions to my evolving thought process and logistical needs, this research project could not have come to fruition. I also owe deep appreciation to my best friend of nearly three decades, Matt Dunphy, whose insistent cheerleading has motivated me since I began my dissertation.

Many thanks to my committee members from Teachers College and Columbia University: W. Roger Buck and Oren Pizmony-Levy for their input and perspective, O. Roger Anderson and Felicia Moore Mensah for guiding and encouraging me through a circuitous academic path, and especially my sponsor, mentor, and thought partner in this work, Ann Rivet.

Finally, I would like to thank my family for their ongoing love, pride, and confidence in me: my parents, Bob and Andi Whitaker; my siblings, Zoey and Dylan; my husband, Dave Weber, and most importantly, my daughter Zinnia, whose repeated questions about what it meant to get a PhD served as an ongoing reminder of why I wanted to do this. - ARW

Chapter 1: Introduction

In this dissertation, I examine the design and implementation of a cognitively-based assessment task in geology, as well as the role of such assessments to provide meaningful and valid information about students' understanding of science content and skills.

1.1 The Role of Assessment

In the age of No Child Left Behind (NCLB) and the Every Student Succeeds Act (ESSA), assessment has become an increasingly important issue in education. In order to receive federal funding, state education departments must demonstrate that they have successfully implemented assessment systems, which are to be used in the determination of student achievement (Jones, Carr, & Ataya, 2007). Each state is currently responsible for designing its assessment methodology and defining achievement levels (Linn, Baker, & Betebenner, 2002). Since the enactment of NCLB in 2001, most states have developed accountability systems that rely on standardized testing as the primary means for defining and measuring student achievement levels (Goertz & Duffy, 2003). The growing influence of these tests has raised issues about the nature and purpose of assessment, particularly at the secondary level where their high-stakes nature means they are used to determine whether individual students graduate from high school, to judge the merit of individual schools (or sometimes teachers), and to set policy for school improvement (Popham, 2006). Although the enactment of ESSA in 2015 was intended to give states more flexibility in measuring student learning, few state legislatures have made substantial changes to the nature of their tests (Gewertz, 2018). However, the changes introduced by ESSA have created a pathway for assessment reform at the state level, through the Innovative Assessment Demonstration Authority, making this an opportune time for novel assessment formats to be piloted (Ujifsa, 2019).

In the decades since assessment and accountability systems have been granted an increasingly important role in education, they have also become viewed as an important part of education reform. Assessments are appealing as agents of educational change because, compared to most other reforms (such as reducing class size or increasing teacher education) they are cheap, fast, and easy to implement (Linn, 2000). It's important to keep in mind, however, that large-scale assessment has numerous uses beyond accountability. Black and Wiliam (1998) describe the difference between assessment *of* learning and assessment *for* learning. The latter, typically described as formative, calls for continual and frequent use of assessment, and plays a key role in helping teachers determine the scope and direction of their instruction. In contrast, the primary function of assessment *of* learning is to assign grades, and as such usually takes place after the instruction of a given topic is completed (Black, 1998). This summative approach to assessment is now used by state education departments as a major, or sometimes sole, component of determining student achievement in relation to a set of predetermined standards. However, the assessments used for accountability purposes can produce data that is also able to inform pedagogical decisions. Formative assessment has many benefits (e.g. Treagust et al., 2001); therefore, an effective assessment should allow for both formative uses, such as informing pedagogical decisions made by individual teachers, and summative uses, such as providing data for accountability measures.

The term "performance assessment" points to a major difference from other forms of assessment. This difference is one of form, not function. Performance assessment data may be used for formative or summative purposes. The key difference is that performance assessment requires students to engage with a task, in contrast to assessments that require students to read questions and write or select answers. This is one aspect of assessment design that may improve the quality of science assessment, but a performance approach alone is not sufficient.

1.2 Cognitively-Based Assessment

One of the many ways in which science educators have attempted to improve the quality of assessment is by turning to the findings of cognitive science to guide the development of assessment tasks. As we begin to describe the cognitive processes and activities that underlie different types of performance in problem-solving situations, we can use this knowledge to design tasks that elicit these specific cognitive responses (Linn, Baker, & Dunbar, 1991). The goal of a cognitively-based performance assessment task is to take into account what we know about how students learn a given topic, how their thinking changes as they gain expertise in that topic, and how they demonstrate their knowledge and skills within a given content domain. If this is done appropriately, the assessment task and associated scoring procedure should be able to provide an interpretation of the student's response that describes the nature and complexity of her understanding of that content as well as the related science skills used in solving problems in that realm.

This contribution from cognitive science is an important aspect of creating assessments that are suited to the task of facilitating meaningful judgments about student learning. Poorly designed or enacted assessments have the potential to be detrimental to teaching and learning. Traditional assessments, which are often written tests where most items are multiple-choice or short answer questions, have been criticized for their inability to measure higher-order thinking (Darling-Hammond, 1993), problem-solving skills, or conceptual understanding (Tamir, 1998) due to their near-exclusive focus on recall of facts. By asking students to represent their learning through decontextualized chunks of rote knowledge, traditional assessments fail to provide a complete picture of what students actually understand and can do. High scores on such tests are associated with superficial knowledge (Kohn, 2004) and do not necessarily indicate true mastery of a subject. Educators must be able to measure more than students' superficial knowledge if they wish to use assessments as a tool to support teaching and learning.

A high-quality science assessment, informed by research in cognitive science, should be able to provide information about a student's understanding of specific content and concepts beyond the superficial level typically measured by traditional tests. Research has shown that experts in a particular content area exhibit not only a deeper and more comprehensive understanding of that content, but also different ways of thinking about the content, which are manifested in how they describe and solve problems (NRC, 1999). Therefore, it is important that science assessments are able to characterize the nature of a learner's thinking in addition to her content knowledge. This need is underscored by the different standards that have been used to guide decisions around science curriculum design and instruction for the past decade. Both the *Next Generation Science Standards* (NGSS) and most state science standards include science practices or inquiry process skills, habits of mind, and an understanding of the interconnectedness of natural phenomena in their descriptions of what a student should gain from her science education (NGSS Lead States, 2013). Without appropriate assessments, it is difficult for teachers, administrators, and policy makers to make informed decisions about how to best support meaningful, worthwhile student learning in science.

In this era of assessment-driven instruction, much time and effort has been spent implementing large-scale summative assessments, many of which take the form of traditional tests. These assessments are increasingly being used in high-stakes contexts. The need for validity, along with various methods of demonstrating validity, is well established. While traditional tests are easily validated in a statistical sense, this does not mean they are valid tools for the current goals of measuring science learning. Research has also clearly shown the benefits of both performance- and cognitively-based assessments, particularly in science. What's missing in science assessment is a combination of these features in a scalable way, that could be implemented in individual classrooms, districts, or even statewide.

1.3 Research Questions

The purpose of this study is to examine an approach to designing, implementing, and scoring a cognitively-based performance assessment task in Earth Science. The intention of my approach is to bring together the principles of cognitively-based assessments and performance assessments in a way that allows scoring to be coherent, and is potentially scalable to different uses and contexts. *Coherence* is key to the way I am conceptualizing validity. Construct validity is, in essence, a measure of how good the match is between what the assessment intends to measure and what it actually measures. If the assessment has good construct validity, then the claims that can be made based on student responses to the assessment tasks are in alignment with the stated purpose and goals of the assessment. The responses given by the students to the task(s) enable educators to make inferences about the students' knowledge and thinking used in their completion of the task. If the entire design—the task itself, the method of capturing student responses, and the scoring procedure and output—is coherent, then the scores are valid and meaningful.

This study is attempting to answer a big-picture question: *How do cognitively-based assessments promote coherence between students' understanding, responses, and scoring?* In order to do this, the following research questions will be addressed:

1. To what extent does the performance assessment task elicit the intended constructs in student thinking?

2. To what extent do the students' recorded responses to the assessment task correlate with their thinking during the task?

3. To what extent do the results of the scoring procedure correlate with the students' recorded responses?

4. To what extent do scores represent the range of student thinking?

In addition to documenting the process and challenges encountered during this initial attempt at a cognitively-based assessment task design, the study examines the construct validity of the assessment products, including the relationship between student responses and interpretations provided by a scoring procedure.

The content areas addressed in this task include topographic maps, surface processes, plate tectonics, and geologic history. The focus on Earth Science is a unique aspect of this study, but its findings should be applicable to content-based science assessment in general.

1.4 Research Methods

In order to answer my research questions, I examine several aspects of the cognitively-based science performance assessment task that I have designed. The assessment task comprises a suite of associated items that address a variety of geology topics. Following the procedure recommended by Mark Wilson (2005), I used a set of clearly articulated ideas around learning in these topics as a guide to constructing components of the assessment task and the scoring procedure that was used to interpret student responses. I evaluate the construct validity of the task using an array of complementary data sources.

In addition to the data embodied in the assessment materials themselves—the construct map, assessment task, and scoring procedure—I also observed student responses to the assessment task. I conducted think aloud interviews with twenty-two high school students as they worked on the performance assessment task. The think aloud protocol was designed to balance the need to encourage students to fully explain their thinking about the assessment task with the need to avoid influencing or interfering with the problem-solving process. The think aloud data reflects each student's thought process while they are directly engaged in the assessment task. These data were coded using a combination of categories defined by the construct map and emergent codes identified during the data reduction process.

I also collected each student's written response to the task prompts, which represent the type of product that would come out of a standard administration of the assessment. These written products were coded for common or significant themes using a partially emergent coding scheme, and then scored according to a predetermined scoring procedure. By collecting these different types of data related to the assessment task, and then coding them in order to quantify the nature and range of student responses, I am able to look at multiple correlations between them.

1.5 Significance

This work is significant because it looks at the nature of science assessment in an era when major efforts to reform standardized assessment practice are still in their early stages. It can therefore inform continuing attempts to improve performance assessment at the secondary level. There are multiple facets of these reform efforts related to the different purposes of assessment; this work has potential implications for several of these.

Overall, improved performance assessment tasks have applications for formative, ongoing, and summative purposes. They can be used to measure and describe what students have learned and can do in science, either before, during, or at the conclusion of an instructional sequence. The more effective these assessments are, the more appropriate it will be to use them in support of moderate- to high-stakes decisions such as assigning grades or determining promotion. This work will provide insights into the practicality of designing and implementing cognitively-based performance assessment tools. Future iterations of both the design and implementation of such assessments will allow for the continued refinement of this approach and work towards scalability.

The goal of cognitively-based performance assessment to characterize student responses in relation to a progressive model of conceptual understanding makes such tasks useful for formative assessment as well. When based on research, these cognitive models may be codified as science learning progressions, which take into account "inaccurate, yet productive, understandings that can foster learning of more sophisticated understandings" (Duncan & Rivet, 2013, p. 396). Science learning progressions provide the guideposts for the student understandings a cognitively-based assessment should measure. By more effectively observing the ways in which students are reasoning about a given topic or within a given problem context, educators are able to better determine a productive direction for future instruction for both individuals and groups of students. Ultimately, this assessment approach has the potential to support student learning while providing teachers with data to guide decisions around classroom instruction.

Similarly, this work may contribute to the growing understanding of how students learn, think, and develop understandings about particular topics within geology. This supports continued efforts to characterize learning progressions in science, both through the examination of student responses to cognitively-based performance assessment tasks and through the development of tools to empirically measure and describe those responses. The data collected from the students' think aloud interviews and written work on the assessment task provides insights into common patterns of thinking by novices and early learners in geology, as well as the range of student ideas around these topics. At the same time, increasing our ability to identify distinctions between different levels of sophistication in student understanding, and the relationships between various student ideas across topics, can bring a new level of refinement and validity to current learning progression work in this area.

Chapter 2: Review of the Literature

2.1 Goals of Assessment

Ultimately, the underlying goal of assessment is to gather information about whether and to what extent students have learned a given idea, topic, skill, or way of thinking. The reasons for doing so, however, are myriad and not always complementary. The appropriate format and scope of an assessment are in many ways dependent on how the results of that assessment will eventually be used; the intention behind an assessment tool is therefore a vital component of its design.

Historically, in the context of traditional education, assessment was used to make judgments about individual students' knowledge. This type of assessment typically carried out after instruction for a given topic is completed (Black, 1998), is now known as *summative assessment,* and is designed to assign a single grade to each student. Although this practice remains common today—typically in the form of written exams—it has expanded to provide the large quantities of data that drive modern accountability systems. This information allows judgments to be made by governing agencies about the effectiveness of different programs, schools, and even entire districts (Linn, Baker, & Betebenner, 2002; Yeh, 2006). Teachers and administrators working in teams within a given school can use summative assessment to judge the effectiveness of curriculum materials or instructional policies, allowing them to target areas that need improvement and make decisions about professional development, resource allocation, cohort tracking, and long-term planning (Herman, Aschbacher, & Winters, 2002).

The role of assessment in accountability measures has also grown to inform decisions about student promotion and graduation from high school. In some states, a student's scores on standardized tests are used to determine not only whether or not she will receive a diploma, but the nature of her degree. This type of summative assessment, when used to determine whether or not students will be awarded a diploma, is for obvious reasons referred to as high-stakes assessment. Questions about the validity of high-stakes assessments have made them controversial and unpopular in many educational circles (Koretz, 2008). This is not to say that accountability is unimportant; it is both politically and ethically necessary to have a system in place that ensures students are, in fact, learning. However, the significant consequences attached to high-stakes assessments makes it all the more important that the assessments provide data that allows for clear, fair interpretations and comparisons. Although these purposes have typically been served by traditional written tests, performance assessment can also address the needs of external accountability systems, particularly in science.

Assessment has also increasingly become used as a tool to support student learning. In general, the purpose of this approach, known as *formative assessment,* is twofold: first, to find out precisely what students have learned; and second, to make instruction more effective by informing decisions about curriculum design and enactment. Wynne Harlen (2006) found that, when assessment is used to drive such decision-making, its effectiveness is increased by the use of teacher-designed tasks and scoring instead of using externally designed and implemented tests.

Teachers use formative assessment to monitor the learning of individual students, but they must make their ensuing decisions based on the performance of student groups, such as the members of a given class. Formative, embedded, and ongoing assessments that are intertwined with daily classroom activities provides a teacher with diagnostic information that will help her plan future instruction, make modifications to upcoming lessons, and adapt to the needs of her students. When used effectively, these types of assessments lead to “more focused approach… in which adjustments are made in response to students’ ideas” (Treagust et al., 2001, p. 139).

For these decisions about instruction to be worthwhile they must be supported by data that is rich, accurate, individualized, and descriptive. The data must be able to provide the teacher with insight into productive adaptations of future plans, rather than whether or not a fully executed plan has been successful. This means that an assessment should be able to elicit information about the nature of a student’s existing knowledge and skills within a given content area. Science educators have begun to take advantage of the growing body of research around learning progressions and early learners’ ideas in order to design assessments that can target developing knowledge (NRC, 2014). For example, geoscience educators at Purdue University found that the use of scenario-based constructed response items given in a formative assessment allowed them to target specific misconceptions and naïve assumptions about watershed dynamics held by participants in an in-service environmental science program (Cooper, Shepardson, & Harber, 2002). The scenario-based assessment was designed to gather information about the respondents’ problem-solving approach to a given scenario as well as about their knowledge of the relevant environmental science concepts.

Students themselves may also use assessments to track their progress in developing their understanding of a given topic, to identify their strengths and weaknesses in learning, and to guide their efforts for improvement (Herman, Aschbacher, & Winters, 1992). The *Next Generation Science Standards* (NGSS) cite this use of assessments as one of the reasons they should be fully integrated with other classroom activities. This type of assessment practice can help develop self-directed learners who are able to monitor, consider, and adjust their engagement with a subject in order to make it as productive as possible. The NGSS front matter also explains the necessity that assessments reflect not only the content (referred to as "Disciplinary Core Ideas") but also the science practices and skills as well as unifying themes ("crosscutting concepts") that are valued by the science education community (NGSS Lead States, 2013).

What we have seen, then, is that the purpose of assessment is at least partly dependent on its audience. Administrators, politicians, classroom teachers, and students all have different, though related, uses for assessment and therefore find value in different types of assessment data. A 2007 study comparing stakeholders' beliefs about different types of assessment found that, although most educators had similar beliefs overall, classroom teachers and administrators had differing perceptions about the validity of formative and summative assessments. Teachers were more likely than administrators to consider formative assessments, such as daily work done within the classroom, to be highly valid indicators of student learning. Conversely, the administrators had stronger positive beliefs about the validity of summative, external assessments like standardized multiple-choice tests as indicators of student learning than teachers did (Guskey, 2007). This difference among stakeholders underscores the differing, sometimes divergent, purposes of assessment. The teachers valued data that could guide the strategies and decisions about ongoing instruction, leading them to see formative assessment as useful in ways that administrators, who were more interested in normed comparisons between larger groups of students, typically did not.

It may not always be possible for a single assessment (or form of assessment) to meet the needs of all these different stakeholders. However, if a summative assessment is able to provide student-specific data that helps teachers understand the progress of their learners and plan future instruction, those assessments are more likely to be seen as worth the time it takes to administer, score, and interpret them. More importantly, assessments that are able to serve multiple purposes will have a greater potential to have a positive effect on student learning. While the performance assessment described in this work is designed with summative assessment goals in mind, the scoring procedure (detailed in chapter 3) provides this type of data by describing the student's understanding and skill level in the context of a continuum, thus outlining a target for future learning and fulfilling a goal of formative assessment.

The growing prominence of assessment in both education policy and practice makes it essential that its role be to make positive contributions to education—meaning that it supports genuine student learning—and that the assessments in use are well suited to fulfill all their intended purposes. Although the difference between formative and summative assessment is chiefly one of how the data is used, it would be a mistake to assume that all assessment data is appropriate for use in more than one way. Assessments are beholden to a number of educational demands: they must be practical to implement in a consistent way across large numbers of students; they must be aligned with a variety of local, state, and national standards; and above all, they must be able to provide meaningful data that accurately describes students' knowledge, skills, and even their thinking. Assessment experts do not believe this can be understated: "it is arguably as strong a moral imperative upon educators to satisfy themselves that the technologies being used are 'safe' and beneficial as there is on a nuclear scientist or a biologist working on genetically modified crops" (Broadfoot & Black, 2004). Poorly designed or enacted assessments have the potential to be detrimental to teaching and learning, not only because of measurement error: assessments send a message about what is important or valued in instruction, so when there is a lack of alignment between those goals and the nature of the assessments themselves, it undermines reform efforts in other areas such as inquiry-based curriculum projects. Additionally, without the tools to measure progress in such reform-minded areas, it is impossible to know if the larger goals of science education are being met. This speaks to the concept of construct validity, which will be discussed later in this chapter.

2.2 Relationship Between Cognition and Assessment

In 2001, the National Research Council published *Knowing What Students Know,* an exhaustive report on the applications of cognitive science and educational measurement research to assessment design. In *Knowing What Students Know,* the committee on foundations of assessment presents a framework for a coherent assessment system called the assessment triangle. The assessment triangle, shown in Figure 1, is a theoretical model of an assessment system—not just an assessment tool, but a process by which evidence of knowledge or learning can be gathered and interpreted, and by which the instruments used to gather that evidence can be developed and refined. The assessment triangle makes explicit connections between cognition and the tools and methodologies used by assessment systems. This is an important necessity in educational measurement; if we value any particular style or pattern of cognition as a learning outcome, we must find a way to measure it.

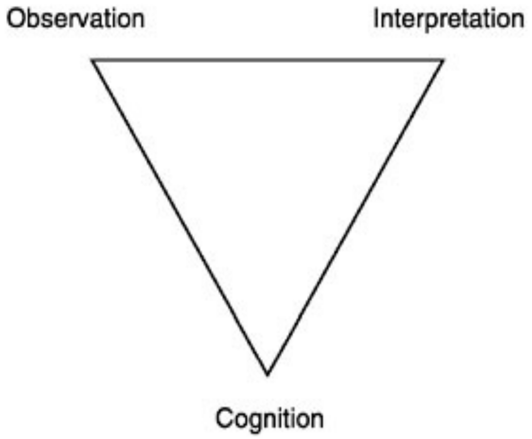


Figure 1: The Assessment Triangle. From *Knowing What Students Know: The Science and Design of Educational Assessment* (NRC, 2001).

Because it serves as the basic theoretical model for assessment design in this study, the components of the assessment triangle are important to understand. Each labeled corner represents a different component or aspect of a functional assessment system. The *cognition* corner represents a theory of learning for the specific content or skill area being assessed. A useful model of cognition will "identify performances that differentiate beginning and expert learners in the domain… [ideally,] a model of learning will also provide a developmental perspective, laying out one or more typical progressions from novice levels towards competence and then expertise, identifying milestones or landmark performances along the way" (NRC, 2001, pp. 181-182). In other words, the model of cognition describes successive levels of achievement that an assessment should be able to measure. The authors of the assessment triangle model caution that the intended scope and purpose of an assessment tool will necessarily dictate the most appropriate use, grain size, and focus of the cognitive model that guides its design. For example, when designing large-scale summative assessments, "a coarser-grained model of learning that focuses on the development of central conceptual structures in the subject domain may suffice" (NRC, 2001, p. 184).

The purpose of grounding an assessment in a cognitive model is to ensure that you are measuring the things you intend to measure. The cognitive model describes typical (and perhaps some atypical) "milestones" that learners hit as they progress from novice or beginning understandings to more complete, advanced, and expert understandings. Therefore in order to successfully measure where a learner is, the assessment must prompt or provide room for responses at those particular milestones. The assumption we allow ourselves to make is that, when an assessment is based on a research-derived cognitive model of learning in a particular domain, the learner's responses to that assessment will give us useful information about where the learner is now, as well as what the next stage(s) in the development of her understanding may be.

Recent research on science learning progressions supports this assumption. A learning progression describes a possible development of understanding between two anchor points, one at the novice end and one at a more expert end (NRC, 2007). It can delineate typical, expected levels of understanding or even misunderstanding that learners often move through as they move towards the more sophisticated anchor point. Wiser, Smith, and Doubler (2012) refer to these intermediate levels as "stepping stones," noting that they are valuable for guiding curricular targets that will "allow students to bridge successfully between the two anchors" (p. 9). If these stepping-stones are useful points for teaching, they are also worth assessing. An assessment designed to pinpoint which stepping stone a learner currently occupies will provide useful information about how to direct her learning. This need is explicitly stated in the recommendations made by the Committee on Developing Assessments of Science Proficiency in K-12, for the purposes of supporting the NGSS. They conclude, "the *Next Generation Science Standards* require that assessment tasks be designed so that they can accurately locate students along a sequence of progressively more complex understandings of a core idea" (NRC, 2014, p. 38).

The counter-scenario is this: when an assessment is based not on a data-derived model, but on a guess, a hypothesis, or merely a checklist of topics, it cannot provide evidence of the learner's understanding relative to a progression or known schema. If we want to measure the nature of understanding, or the sophistication of a learner's knowledge, it is vital that we base our assessments on appropriate and relevant cognitive models. Without them, we are likely to make either incomplete or off-target measurements. The cognitive model chosen defines for us the type of thinking and understanding we will be able to measure.

It is important to remember that the quality of the cognitive model alone will not guarantee a valid assessment; it is but one of three interrelated parts. The other two, which make up the remainder of the assessment triangle framework, factor heavily in the assessment design process and are discussed more thoroughly in chapter 4. The *observation* corner of the assessment triangle represents the evidence of student learning, understanding, or skill level gathered via an assessment tool. The *interpretation* corner of the assessment triangle represents the way in which we make sense of the output produced by students in response to the assessment. In most assessment systems, this is the way we generate and represent scores for individual students. If any one of these three corners of the assessment triangle is missing, ineffective, or misaligned with the other corners, the entire assessment process loses validity.

2.3 Validity

What does it mean for an assessment to be valid? While a cognitive model helps provide bounds for what we wish to measure, and may also guide the design of the assessment tool used to do so, a cognitive model alone cannot guarantee construct validity, no matter its level of quality or detail. Linn and Gronlund (2000) define construct validity as "the process of determining the extent to which performance on an assessment can be interpreted in terms of one or more constructs." (p. 83) Simply put, this means that construct validation involves determining whether or not the thing (knowledge, skill, behavior, understanding) we intend to measure is in fact the thing being measured by an assessment.

Much as our understanding and use of assessments has changed over time, so has our concept of validity, which is no longer constrained by the notion that assessments must use closed-response items for purposes of statistical analysis. As assessments become more diverse and open-ended, "the initial meaning of 'measuring what it purports to measure' in relation to traditional multiple choice and pen-and-paper tests has been expanded as the notion of validity has been developed with respect to the quality of alternative assessments, such as performance assessment" (Moss 1992, in Bell, 2007, p. 989). Regardless of the nature of the assessment items or the specific statistical indicators used, "the validity of an assessment can be evaluated in terms of the extent to which the assessment relates to the ascribed educational values, learning theories, and teaching theories as well as to the realization of the desired assessment theory" (Cumming & Maxwell 1999, p. 193).

In her review of issues surrounding assessment in science education, Beverly Bell points out that there are many aspects of validity deriving from the intentions behind a given assessment, from the educational theory underpinning its content and structure to the purposes for which it will be used. These different aspects of a more broadly viewed validity include "consequences, equity, fairness, cultural validity, trustworthiness, appropriateness, manageability, fidelity, and authenticity" (Bell, 2007, p. 990). While all these considerations are important, they do not eclipse nor reduce the need to ensure that assessments are valid measures of the constructs they address and the interpretations about student understanding that are suggested by the assessment results. Therefore, assessment design efforts focus on construct validity as perhaps the most fundamental aspect of validity when developing new assessment tools.

The Assessment Triangle, discussed in the preceding pages, provides a useful frame for looking at validity, and the ways in which validity may be compromised. In the following section, I will examine the ways in which assessments may fail to provide validity, and how cognitively-based performance assessments represent potential solutions to these problems.

2.4 Cognitively-Based Assessment Solution

What does it mean for an assessment to be cognitively based? Simply put, cognitively-based assessments must focus on a model of cognition as the foundation for establishing validity. A cognitively-based assessment is one that takes modes, progressions, or patterns of learning—particularly those that have been established through empirical research—into account. The NRC's assessment triangle therefore makes an implicit argument that *all* assessments should be cognitively based assessments. Why is this so important? In *Constructing Measures* (2005), Mark Wilson describes an important difference between the type of knowledge measured by traditional assessments and the "knowing" that we hope to achieve through education, and therefore must be able to measure:

> To cognitive psychologists, knowing is not merely the accumulation of factual information and routine procedures. Knowing means being able to combine knowledge, skills, and procedures in ways that are useful for interpreting new situations and solving problems. Thus, assessment of cognitive constructs should not over emphasize basic information and skills – these should be seen as resources for more meaningful activities. (Wilson, 2005, p. 183)

We should use the cognitive psychologists' definition, and understanding, of *knowing* in order to design assessments that will give students the opportunity to demonstrate their "knowledge, skills, and [facility with] procedures" that are valued by the scientific community. Without the cognitive model as a foundation, we are setting ourselves up to measure the wrong thing, or at the very least, to measure only part of our target.

One example from the nascent field of cognitively-based assessment is the assessment of college readiness in mathematics designed by researchers at the UC Berkeley BEAR Center. The study by Wilmot, Schoenfeld, Wilson, Champney, and Zabner (2011) looked at students' understanding of mathematical functions in secondary school (grades 6 – 12), and was designed to "quantify and measure student thinking on a developmental trajectory," using existing research in mathematical reasoning as the basis for the construct map (Wilmot & Champney, 2008). The authors of the assessment found that they were able to capture evidence of reasoning at different levels on a theoretical hierarchy via a performance-based assessment.

In contrast to these examples of successful cognitively-based assessments, there are myriad ways in which assessments may fail to fulfill their intended purpose—in essence, failing to achieve validity. We can examine these failures through the framework of the assessment triangle, looking at one corner of the triangle at a time. The cognition corner tells us that high-quality assessments must be based on a useful cognitive model. Most assessments are not based on a research- or even observation-based model of cognition. Typical science assessments, particularly in a summative high-stakes context, emphasize factual knowledge or basic, decontextualized skill, and "do not require students to demonstrate knowledge of the integration between scientific practices and conceptual understanding" (NRC, 2014, p. 15).

Merely by taking into account what is known about how students learn science, a cognitively-based assessment is taking an important step towards construct validity that many traditional assessments cannot. While traditional assessments are "driven by the assumptions of … decontextualization of knowledge", "the study of active cognitive processes" has led to a greater need for assessments that are appropriately situated in a meaningful context (Klassen, 2006). Unfortunately, this is not yet standard practice in assessment design: Li, Klahr, and Siler

(2006) found that alignment between science assessments and science standards is superficial, and does not adequately support alignment of curriculum to the intent of the content standards. In a 2004 report, O'Neil, Sireci, and Huff examined state-mandated assessments from consecutive years and found that while "content area representation was fairly consistent across years... important cognitive distinctions among test items… were not captured in the test specifications" (p. 129). In other words, the tests were measuring different cognitive skills from year to year, despite purporting to measure specific cognitive skill areas that were described by the test specifications.

Why is this so important? According to Mark Wilson, cognitively-based assessments need to measure more than knowledge. Because of this, "traditional tests, which usually record how many items examinees answer correctly or incorrectly, fall short. What is needed are data about how they reach those answers and/or how well they understand the underlying concepts" (Wilson, 2005, p. 183). Some efforts have been made to gather these data using traditional testing methods, pointing to the importance of a high-quality observational tool—in other words, to good item design. For example, in 2011 the American Association for the Advancement of Science (AAAS) launched a science assessment database of ordered multiple-choice questions, researched and designed by the AAAS's Project 2061 reform initiative to provide meaningful data about students' thinking. The goal of these types of assessment items is that, "when an item is well designed, students should choose the correct answer only when they know the targeted idea and should choose an incorrect answer only when they do not know the idea." (AAAS, 2007, p 3) This online database of assessment items created by Project 2061 attempts to address a longstanding problem with multiple choice items; namely, that distractors or other aspects of a question may confound the data by preventing students from choosing a correct answer even when they have the knowledge or skills to do so (Briggs et al., 2006). (Or vice versa: a student may select the correct answer even when she does not know that it is correct.) Project 2061's approach to solving this problem is to create ordered multiple-choice items that take a cognitive model of student knowledge into account. However, it is impossible to ignore the fact that random chance will always play into test results when multiple-choice items are included.

Finally, assessments may fail due to bad interpretation. For example, O'Reilly and McNamara (2007) studied the correlation between cognitive abilities and student achievement on content-based science tests. They found "significant gender differences," and that "reading skill helped the learner compensate for deficits in science knowledge… and had a larger effect on achievement scores for higher knowledge than lower knowledge students." Interpretations of these test results that make claims about students' level of science content knowledge fail to take reading level into account. This is an example of how scores do not always reflect the type of understanding or skill that they claim to represent, thereby causing a failure of construct validity, because the construct intended to be measured is not necessarily what the scores capture.

International assessments that address STEM learning, including PISA (OECD, 2019) and TIMMS (Mullis & Martin, 2017), are being implemented in multiple countries to facilitate international comparisons of student learning and achievement in a variety of areas over time. The results of such assessments are used for different purposes, both internally by individual nations and cross-culturally in the global endeavor to advance science literacy and practice (e.g., Bybee, McCrae, & Laurie, 2006). Critiques of international assessment echo those of domestic assessments, such as the failure to take the differential impact of position effects into account when comparing item performance across demographic groups (Nagy et al., 2018). While both PISA and TIMMS have some grounding in cognitive models, scholars have raised questions about whether the overall design of these assessments is appropriate for making claims about student learning or making valid comparisons across nations and demographic groups over time (Mazzeo & von Davier, 2009).

2.5 Performance Assessment Solution

Performance assessment, in particular, has the potential to address these breakdowns through a stronger intrinsic alignment to science standards and the types of scientific practices that are valued by the scientific education community. A Project 2061 study found that the loss of alignment comes primarily from the use of traditional assessments (Stern & Ahlgren, 2002). Project 2061 analyzed widely used middle school curriculum materials and associated assessments that were included with the published curriculum materials. They found that the curricula themselves were strongly aligned to benchmarks and national science standards, but that assessments did not address many of these standards, particularly those that concerned science practices.

Unlike traditional assessments, performance assessment requires students to *do* something. They therefore represent a different method of gathering assessment data than the multiple-choice or written tests often used for summative purposes. In the case of science performance assessments, students must engage in the same types of practices, activities, problem solving, and skills that they are required to employ as they learn science. This goes beyond mere "hands-on" or lab-based learning; well-designed performance assessment can address the many standards of scientific practice described in both national and state science standard documents in ways that traditional paper-and-pencil tests cannot. This is particularly true in the case of standards like the *Next Generation Science Standards* (NGSS) that are written in terms of performance, describing actions students will be able to do, rather than ideas they will hold or questions they will be able to answer (NGSS Lead States, 2013). This means performance assessment naturally provides a superior alignment between the *cognitive model* that describes science learning and the *measurement tool* that is being used to capture the knowledge or skills that have come out of that learning.

It does not, however, mean that performance assessments are automatically valid as a result of their design. Even when they require students to engage in open-ended tasks, they may still fail to assess complex cognitive processes or knowledge (Baxter and Glaser, 1998). Care must still be taken when creating performance assessments to ensure that they provide opportunities for students to demonstrate understanding across a range of levels, from low to high.

2.6 The Need for Valid Performance Assessments in Science

Unfortunately, in science, as in other subjects, there is a dearth of validated performance assessments appropriate for use on a large scale. As the previous section outlined, this lack of performance assessment tools is more significant in science precisely because science is so well suited to performance assessment tasks. The nature of science, and the goals of science education, make it particularly vital that we measure performance-based achievement in science subjects.

The goals of science assessment include measuring scientific thinking and practice, including “knowing and understanding scientific facts, concepts, principles, laws, and theories… the ability to reason scientifically… [and] the ability to communicate effectively about science” (NRC, 1996, pp. 79 – 82), all of which extend well beyond mere content knowledge. The need for assessments that can address this type of thinking will become even more pronounced as the *Next Generation Science Standards* (NGSS) are implemented. The NGSS are divided into three major "dimensions," which together create a framework for three-dimensional learning: practices, crosscutting concepts, and disciplinary core ideas (or what we would call "content knowledge").

Practices in the NGSS encompass the knowledge and skills "that scientists employ as they investigate and build models and theories about the world" (NRC 2012, p. 30). Additionally, the NGSS are written in terms of performance. Rather than merely describing what students will understand, they define what students will be able to do. When standards describe a performance, they demand a performance-based assessment. Much as it is difficult to learn science practice without engaging in them, it is unreasonable to expect assessments to capture students' understanding of and facility with these practices if the students are not given the opportunity to perform them as part of the assessment. This is why performance assessments have the potential to capture this dimension of student understanding in a more valid way than traditional assessment can do. Performance assessment is uniquely suited to assessing the three-dimensional science learning because requires students to demonstrate their mastery of science practices through the actual use of those practices, while engaging in tasks that are grounded in disciplinary core ideas and crosscutting concepts. Indeed, the recommendations for NGSS-aligned assessment make it clear that this change in approach to science education requires an equally significant shift in approach to science assessment (NRC, 2014).

Another way in which performance assessment meets the changing needs of science education is by providing variety of integrated tasks, rather than discrete or segregated questions. The Committee on Developing Assessments of Science Proficiency in K-12 points out that "tasks that demand only declarative knowledge about practices or isolated facts would be insufficient to measure performance expectations in the NGSS … [A]ssessment tasks aligned with the NGSS performance expectations will need to have multiple components—that is, be composed of more than one kind of activity or question." (NRC, 2014, p. 89).

Until recently, designing cognitively based science performance assessment tasks has been particularly challenging due to the dearth of clear, valid ways of identifying, describing, and measuring science understanding. However, the changes in the national standards point to a larger shift in the field of science education. New research in learning progressions, upon which the NGSS are based, and a growing body of data from early iterations of large-scale performance assessments have begun to pave the way for their cognitively-based descendants. Existing learning progressions that have been established through education research serve as ways to conceptualize the cognitive model on which a performance assessment is based. The work done by Smith, Wiser, Anderson, and Krajcik (2006), for example, uses a learning progression for matter and atomic-molecular theory to create performance-based assessment items. They argue that assessments based on learning progressions are better equipped to illuminate student thinking. In turn, the data from performance assessments in early phases of development is useful in refining and creating learning progressions that can guide future assessment design. This is important because "reliable interpretations [of student performance on assessment items] require a research base" (Wiser et al. 2006, p. 95); much work remains to be done in the arena of science learning progressions. While some core ideas in science such as atomic theory and genetics (Duncan, Rogat, & Yarden, 2009) have been the subject of learning progression research in the last decade, many of the ideas taught and assessed at the high school level have not yet been formally described in terms of a learning progression. The development of cognitively-based assessments themselves have also supported the mapping of student understandings onto a progressive cognitive model, such as with the ACORNS instrument for measuring undergraduate level thinking about evolution (Opfer, Nehm, & Ha, 2012). This process can also highlight potential issues or areas for revision in performance assessment; for instance, an assessment aligned to a mathematics learning progression about area measurement concluded that the cognitive model was better described as a network of concepts than a hierarchy of understandings (Lai et al., 2017).

Researchers at SUNY Buffalo have worked to develop performance tasks that are aligned to the New York State science curricula as well as to the existing standardized assessments in selected science content areas, the New York State Regents exams. All of these tasks purport to measure a subset of the concepts from the state curriculum in that subject as well as some of the skills used by practicing scientists. A laboratory skill test in physics (Hussain, 2001) was found to have a high degree of reliability and agreement between independent scorers. The assessment aligned with the chemistry curriculum included high- and low-level inquiry performance tasks, which were found to be reliable and valid (Uchinella, 2002). A third assessment addressed the Living Environment (New York State's high school life science) curriculum, specifically targeting the content and skills covered by the mandated state labs (Wright 2002). This is an example of a performance assessment that measures the same understandings in an existing large-scale written assessment, since the New York State Regents Exam in Living Environment includes a written section (Part D) devoted to the lab content. These studies point to the need for performance-based assessments to align with the performance expectations in a state curriculum, as well as the potential for this kind of work to be strengthened by the use of cognitive models like learning progressions.

More recent work examining the relationship between learning progressions and assessment at the high school level has identified several strengths to this approach, along with potential stumbling blocks for teachers attempting to adopt assessment practices grounded in learning progression models (e.g., Schneider & Andrade, 2013). For example, the Daphne Assessment of Natural Selection (DANS) is a multiple-choice test based on the Elevate learning progression for high school learning about natural selection in biology (Furtak et al., 2011), developed via an iterative process of matching think-aloud responses to students' choices on test items. This resulted in a valid assessment for measuring student ideas relative to 13 construct maps for ideas about evolution. The Learning Progression-Based Assessment of Modern Genetics (LPA-MG) created by a team at Wright State University (Todd, Romine, & Cook Whitt, 2017) was able to reliably assess students' learning over time on 12 related constructs within a genetics learning progression. The water systems learning progression (WSLP) developed by Gunckel et al. (2014) as part of a larger environmental literacy research project was found to provide a valid framework for measuring high school students' accounts of how water can move through both natural and engineered systems. These studies show the promise of learning progressions in guiding science instruction and assessment. However, there is a gap in the field's knowledge of how this emerging practice can be applied to performance assessment. The marriage of performance assessment and learning progression-based assessment is an important next step.

2.7 Contributions

This study explores the nature of construct validity and overall coherence in cognitively based science performance assessment task(s). To do this, I examine student thinking while students are engaged in such task(s). This study can provide insights into the types of science thinking and problem solving elicited by the tasks—specifically, in the realm of geology content knowledge and practices—and therefore contribute to the growing field of science education research that describes learning progressions and ways in which students build understandings of science concepts. This is particularly important for topics that are part of the Earth & space sciences, as these subjects are historically underrepresented in science education research and literature. The study will also be able to inform the field on the use of existing learning progressions and learning science research to guide assessment design, and what connections between the fields might be most fruitful.

Relatedly, this study explores several aspects of designing and implementing science performance assessment tasks. It addresses issues around the sensitivity of scoring and the considerations that must be given to the design of assessment components—including the cognitive model or theoretical framework underlying the task, the nature of the task and student responses, and scoring procedures—to accurately reflect these sensitivities. Because science performance assessments often require students to engage in a more active mode of "doing" than performance assessments in other disciplines, studies focused on science performance assessment make unique and significant contributions to the assessment field.

These two contributions—to the development of cognitive models regarding Earth science learning, and to the development of a new generation of science performance assessments—are important on their own, but it is the combination that fills a gap in current science education research. Performance assessment has been shown to meet the needs of many science assessment charges, and current approaches to assessment design increasingly emphasize the use of cognitive models. In this study, I attempt to leverage the benefits of both strategies in a coherent approach to creating, administering, and scoring science assessment.

Chapter 3: Cognitively-Based Performance Assessment Task Design

3.1 Design Goals

One of the most fundamental and difficult challenges facing the assessment field is the impossibility of directly observing a student's thinking, understanding, or any other type of cognition. We cannot see what is going on inside a student's head; instead, we must rely on her performance to provide evidence of her thinking. Therefore, in order to have a high degree of construct validity—in other words, in order to measure precisely the thing we intend and claim to be measuring—it is necessary to design tasks with which the student can engage that gives her the opportunity to demonstrate the particular type of knowledge, understanding, skill, or thinking in which we are interested. If an assessment is to provide useful information about a student's science understanding, it must focus on the student in the process of doing science. This goes beyond merely observing what students do in the course of their everyday classroom science tasks. It is necessary that the students are performing a task designed for the specific purpose of assessment (Mislvey, 2004).

Well-designed science performance assessments are able to examine a different form of knowledge from what is demonstrated by multiple-choice tests, which more closely correlates with that used by scientists in their work, including a schematic understanding of why a phenomenon occurs (Rothman, 1995; Li & Shavelson, 2001). As previously stated, research has shown that experts in a particular content area exhibit not only a deeper and more comprehensive understanding of that content, but also different ways of thinking about the content, which are manifested in how they describe and solve problems (NRC, 2001). Therefore, it is important that high-quality science assessments are able to characterize the nature of a learner's reasoning in addition to her content knowledge. This need is underscored by the different standards that have been used to guide decisions around science curriculum design and instruction for the past decade. Both the *Next Generation Science Standards* (NGSS) and the New York State Earth Science Core Curriculum (NYS ESCC) include science inquiry practices, habits of mind, and an understanding of the connections between ideas in their descriptions of what a student should gain from her science education (NGSS Lead States, 2013; NYSED, 2001).

This work has set out to develop an example of a cognitively-based performance assessment task that would address the need for measurement of students' thinking and problem-solving skills related to specific complex science concepts, while confronting the challenges associated with the design and implementation of traditionally designed performance assessments described in the literature. When designing these cognitively-based assessment task components, the primary focus was on addressing the challenges around construct validity that are inherent in traditional performance assessments. The aim of this strategy for task design was to create a tool whose properties would ensure a transparent, reliable relationship between the students' recorded responses and the nature of their understanding within that content area. The strength of this relationship is a significant determinant of the strength of claims that can be made about students' thinking in that area.

3.2 Theoretical Basis for Task Design

Because my research questions address the relationships between different components of the assessment design and interpretation process, they require the creation of multiple, but related, products, each situated on one corner of the assessment triangle described in Chapter 2 (NRC, 2001). The first is a clear definition of the constructs, or what is being measured, in the form of the construct map, corresponding to the "cognition" corner of the assessment triangle. The second is the assessment task itself, corresponding to the "observation" corner of the assessment triangle. The third is the scoring procedure, which includes the outcome space and a standardized means of using it to generate a numerical score, corresponding to the "interpretation" corner of the assessment triangle.

This approach to designing the performance assessment task is based on the conceptual aspect of the BEAR assessment system (Wilson, 2005) and follows from the recommendations made by the NRC in *Knowing what students know* (2001), as well as embracing the philosophical approach to science performance assessments that have been developed in order to more effectively measure the types of knowledge, practices, and skills that are integral to effective learning in science (e.g. Mislevy & Baxter, 2005).

The BEAR assessment system is based around four "building blocks" that guide the development and interpretation of assessment items, based on a theory of learning and thinking in a given domain. Wilson (2005) describes this process as an iterative approach to good assessment design. In the first building block, the construct(s) to be measured are operationalized via a construct map, which for each construct defines a continuum of levels of understanding ranging from novice to expert. The construct map represents theoretical levels of understanding that may be possessed by the respondent; it is used to design task components that will give the student appropriate opportunities to demonstrate the level of her knowledge and skills. In order to create the construct map, it is necessary to identify & describe the levels of progression of student thinking or understanding that exist within a given construct, in a way that is informed by a cognitive model of learning.

The second building block is the creation of the assessment tasks themselves. The task(s) must elicit thinking within the given construct(s) from the student, such that there is the opportunity for all levels of thinking to be demonstrated over the course of the entire assessment.

The task should be aligned with the construct map to ensure that the students are thinking about the intended constructs. For performance assessment, some parts of the task may target specific constructs or construct levels, while others may be able to measure multiple constructs at once due to the complexity and richness of the performance task.

The construct map is also used to frame the task outcome space, which defines typical student responses that are expected at each level of understanding and is the basis for the third building block in the BEAR system. When dealing with performance assessment, it is not always the case that there are discrete items the way there would be on a written exam. However, it is still possible to describe a student response for each level of the construct map on individual components of the task. The outcome space comprises these descriptions. It is then used to create an overall scoring guide based on the outcome space for each assessment component. As both the task components and the outcome space are mapped to specific levels of the construct map, the student's response to the task will allow an instructor to infer the level of sophistication of that student's understanding within a given construct, and to take a first step towards identifying the ways in which the student can be supported in progressing to a more expert understanding.

The fourth building block entails the collection of quantitative data (in the form of scored student responses) that can be used to validate the construct map, task design, and scoring guide and to improve future versions of the assessment. Because my study is focusing on the first attempt at an assessment task design, the process is not yet iterative and the fourth building block of data collection is therefore used to address research questions, not to inform the content or design of any part of the construct map, assessment, or scoring guide. The data from student responses will, however, be useful when making future attempts at revising and designing new cognitively-based performance assessment tasks.

3.3 Design Principles for Performance Assessment Task

In order to accurately measure a student's level of understanding, it is necessary to observe her responses to a variety of assessment items within each construct. A greater variety and quantity of items provides, at the very least, greater reliability (Linn & Gronlund, 2000). Performance assessments typically comprise only a small number of items in a given content area, making any variation between those items a greater threat to validity. An important part of this design process, then, was to address this problem without removing the performance aspect that makes it possible to observe students engaging in authentic science practice.

The following design principles were established to guide the development of the performance assessment task:

1. *The assessment task(s) must be complex and rich enough to elicit responses across the spectrum of possible performance levels for a given construct, while simultaneously addressing a variety of related constructs.* Because each component of the assessment task can be scored relative to multiple constructs, this will enable me to make enough discrete observations of student performance on several concepts within a reasonable time period. Additionally, the complexity of the task in which the students engage is important when assessing understanding of interconnectedness of ideas (Wilson, 2005; NRC, 2001).

2. *The task must be authentic to the content area in which it is situated.* For the purposes of an Earth Science assessment targeting high school (9-12) level students, this means that the task requires the students to engage in practices similar to those undertaken by geologists, although the content may be somewhat simplified (Wiggins, 1989). The value of "authentic assessment" was supported in the National Science Education Standards as well (NRC, 1996). The NSES define authentic assessment as tasks that are similar to situations or problems that

either scientists or lay persons might deal with in the world outside the classroom. In the NGSS, this idea of authenticity is embodied in the standards themselves in two ways. First, it can be found in the focus on scientific practices as one of the three main dimensions of the standards. Second, the standards themselves are written as performances, indicating the authors' belief that science knowledge must be demonstrated through the types of tasks or activities that a practitioner would undertake. The front matter of the NGSS links this to assessment by stating, "students should be held responsible for demonstrating knowledge of content in various contexts and Scientific and Engineering Practices" (NGSS Lead States, 2013).

3. *The task must be standards-based, in that it addresses content knowledge and skills that have been identified as appropriate for high school students by a widely accepted set of standards.* In this case, the tasks were designed to align with both the Next Generation Science Standards (NGSS Lead States, 2013) and the New York State Core Curriculum in Earth Science (NYSED, 2001). The New York State Education Department in December 2016 adopted updated standards, the New York State P-12 Science Learning Standards (NYSP12LS), which are identical to the NGSS for the geology topics represented in this assessment. I chose to maintain the alignment to the NYSESCC because the transition to full implementation of the NYSP12LS is intended to take place over a ten-year period, with the first administration of the associated state level tests occurring in June 2025. Thus, the cognitively-based performance assessment in this study connected to the curriculum high school students were learning in their Earth Science classes during the time of data collection.

4. *The task components must be based on research about learning*. This is what it means for the tasks to be "cognitively based" – the design of the task takes into account what is known about learning and thinking in a given domain. This includes work around cognitive models

(Manduca & Mogk, 2006) and science learning progressions (e.g. Smith et al., 2006). An assessment that represents a more complete model of how students generate and express understanding about science will allow us to make the most useful observations and meaningful inferences about their learning (NRC, 2001).

These design principles are general and may be applied to any effort in creating a cognitively-based science performance assessment. This first attempt deals with geology-based content and skills, but is primarily being used to examine the issues around construct validity and coherence for this type of science assessment.

3.4 The Construct Map

The assessment tool itself is suite of associated items composing a single performance assessment task in Earth Science. The content areas addressed in this task fall within the broad discipline of geology, and include topographic maps, surface processes, plate tectonics, and geologic history. Following the procedure recommended by Wilson (2005), the first step in designing a cognitively based performance assessment task was defining the constructs that will be measured by the task.

Broadly, the four constructs addressed by this performance assessment task are:

- Geologic time and stratigraphy: knowledge and understanding of geologic history as represented in present-day strata;
- Surface processes (weathering and erosion): knowledge and understanding of surface processes that influence landscape shape;
- Plate tectonics: knowledge and understanding of tectonic processes that influence the local structure of a landscape;
- Topographic maps: understanding of and facility with maps that show topography

with contour lines;

Each of these four constructs is divided into four discrete performance levels, which were defined based on a combination of current learning theory in geology, (e.g. Manduca & Mogk, 2006), student responses to open-ended questions and interviews, and review from a number of Earth Science teachers and researchers in geology. Figure 2 outlines the construct map.

Although I do not claim that the cognitive distance between levels is uniform across constructs—that is, the progression between levels 2 and 3 for one construct may be more challenging than the progression between levels 2 and 3 for a different construct—each level represents a similar type or mode of thinking in each construct. The hierarchy describing the different levels of sophistication in thinking, reasoning, and understanding that I used to guide the characterization of the construct map was developed by Anderson & Krathwohl (2001), and is intended to be a more accurate and student-centered framing of thinking processes than the more "classic" taxonomies of thinking. Typically, levels 1 and level 2 describe remembering and comprehension modes of thinking, defined by Anderson and Krathwohl to mean the recall of factual information and the act of making inferences or extrapolations based upon that knowledge. Level 3 in the construct map generally describes application and analysis modes of thinking, in which the student is able to recognize patterns, make comparisons and connections, and transfer understandings from one context to another. At level 4, the most expert level, the learner's cognition around these topics is more likely to be characterized by synthesis or evaluation modes of thinking. In addition, I am using the generalized understandings expected by the NYS Earth Science Core Curriculum as a target for level 3. This is another force of standardization placed upon the construct map.

Construct	**Level 1** (more novice)	**Level 2**	**Level 3** (NYSESCC target)	**Level 4** (more expert)
Geologic Time & Stratigraphy [GTS]	• Incorrectly states number of stratigraphic layers indicates absolute age • Classifies events in geologic history as either "extremely ancient" or "less ancient"	• Always describes relative age in terms of superposition, not taking potential later events into account	• States that thicker rock layers do not necessarily represent longer time periods • Makes correlations between layers and fossils in different outcrops, can identify that they are the same / similar age. • Identifies crosscutting features /events as younger than the layers being cut	• Connects relative time to absolute time • Describes change along a variety of timescales, from short term (e.g. volcanic eruptions) to long term (e.g. climate change due to continental drift)
	• Makes few or no connections between events and the context in which they occurred	• Able to identify a series of rock formation events but can't describe the changing environmental context • Accurately identifies rock **composition** or **materials** involved in formation	• Accurately describes rock formation **processes** based on present-day appearance of an outcrop • Describes a correct, generic formation environment based on rock properties (e.g. "marine")	• Can visualize and describe a landscape in the environmental condition that it was in during the formation of rock layers • Describes a correct, **specific** formation environment based on rock properties (e.g. "continental margin offshore from a river delta")
Surface Processes (weathering & erosion) [SP]	• Makes no distinction between weathering and erosion processes • Misattributes effects of weathering and erosion processes to other causes	• Recognizes evidence of weathering and erosion processes • Can distinguish between weathering and erosion in general	• Can explain how the composition of a landscape or rock layer affects weathering processes • Recognizes evidence for past processes in current landscape features. • Identifies or describes generalized effects of erosion processes on land	• Identifies or describes specific effects of different agents of weathering and/or erosion on landscapes • Able to predict future patterns or events of surface change based on past evidence
Plate Tectonics [PT]	• Describes Earth's surface as static, not dynamic • Describes continents floating through the ocean • Describes location of tectonic plates as below the surface of the Earth; unable to make connections	• Describes static spatial relationships between plates • Correctly identifies tectonic plates as crustal rocks floating on the mantle • Does not describe destructive/ constructive nature of plate boundaries (connection to rock cycle)	• Describes causal, dynamic relationships • Correctly identifies mantle convection as driving force of plate tectonics • Describes plate boundaries in context of rock cycle processes (e.g. constructive or destructive	• Describes connection between causal relationships and visual/spatial effects that are observable at the surface (e.g. mantle upwelling at divergent boundary) • Makes connections between continual, long-term

	between plates and surface changes • Ascribes causality to purely surface causes or phenomena	• Recognizes rapid change events (e.g. earthquakes) as resulting from plate movement	effects; metamorphism)	processes and rapid change events. Describes duality of gradual or sudden effects of constant movement. • Correctly describes contribution of ridge push / slab pull to plate movement
			• Identifies long-term processes (e.g. widening Atlantic ocean) as well as rapid change events as resulting from plate movement	
Topographic Maps [TM]	• Does not relate scale model to real life • Does not connect patterns in maps to appropriate real-world archetypes • Incorrectly describes or cannot describe topographic features based on the shape of contour lines.	• Correctly interprets iconic representations (map thing looks like real world thing)	• Correctly interprets symbolic representations (e.g. landforms)	• Transfers observations of scale model to find or analyze a real-world object / location. • Uses elevation data to construct an accurate profile of a given line • Describes patterns in contour lines representing typical geologic formations (schema)
		• Interprets topology in terms of qualitative or relative relationships (e.g. can identify uphill / downhill; steep / flat) • Can identify properties at a specific location but not put it into a larger context	• Interprets topology from a particular viewpoint or perspective (e.g. can identify the general shape of a profile along a given line) • Able to make comparisons or statements about different locations relative to each other (e.g. steeper, higher, change in a particular direction, etc.)	

Yellow = performance indicators that are directly based on existing literature or standards documents
Golden = performance indicators that are extrapolated from existing literature or standards documents
Red-Orange = performance indicators that are derived from teaching experience

Figure 2: Construct Map.

3.5 Construct: Geologic Time and Stratigraphy [GTS].

This construct deals with the concepts surrounding the history of the Earth and the ways in which that history is preserved (or revealed) by the physical state of the Earth's lithosphere. According to research by Dodick and Orion (2006), there are two aspects to geologic time that a learner must grapple with. The first is relative time, or the sequential nature of Earth's history in which we can determine the order of events based on locations or spatial relationships in parts of the Earth's crust. The second is absolute time, or the assignment of a specific, numerical age to components of the Earth's crust and the associated events that created them. These ideas are both part of the concept of "deep time," the idea that the timescale of Earth's existence (and, therefore, the geologic processes that shaped it) is far greater than that of human existence, but they represent different ways of thinking about deep time. It is appropriate, therefore, to consider both of these modes of thinking about geologic history when defining the theoretical levels of this construct.

It is difficult for novice thinkers to grasp the length of geologic time precisely because of its "depth" – it spans such a greater amount of time than our human experience, in terms of both an individual's lifespan and our collective existence as a species. Beginning geology thinkers also have trouble connecting physical clues in the rock layers, such as grain size, crossbedding patterns, unconformities, and fossil evidence to changes in the depositional environment and other factors that affect the appearance of the outcrop. Based on the literature and on the anecdotal experience of Earth Science teachers, I am inferring that the ability to do this progresses from a recognition level—in which the student is able to identify discrete events (such as deposition of a specific rock layer, erosion, or faulting)—to an explanatory level, in which the student is able to picture the different environmental surroundings that would have led to the currently visible changes in the lithosphere.

3.5.a GTS Level 1

The novice level for this construct is based on Dodick & Orion's (2006) work on cognitive models for how an understanding of geologic time develops. According to their research, the learner initially possesses no real awareness of absolute time as a quantitative, measured phenomenon. Instead, the novice learner conceives of the number of stratigraphic layers as an indicator of absolute time, as if rocks were formed at a constant rate at locations around the world (this view fails to take erosion into account). In her mind, the rock layers are functionally equivalent to tree rings: there is a linear and direct relationship between the number of layers and the age of the outcrop. In other words, she has a total lack of awareness of factors that might affect either the rate of rock formation (e.g. deposition events; rising & falling sea levels) or the destruction of existing rocks. The novice learner is therefore thinking only in terms of relative time. Dodick and Orion (2006) posit that at this level, relative time is mentally chunked into two large categories: events are either extremely ancient (for example, the existence of prehistoric life including organisms like dinosaurs and trilobites) or less ancient (anything more obviously continuous with modern events, including the history of human evolution and recent ice ages).

The remaining descriptors in this construct (indicated with blue sections in the construct map) are based on an extrapolation from the literature on student learning in geology (Anderson, 2006) and Earth systems (Herbert, 2006), plus my experience with student interviews and qualitative, informal assessment tasks in a normal classroom context. This part of the construct deals with the ways in which currently visible evidence in rocks provides clues, or a narrative, for how the local environment has changed throughout geologic time. It is connected to the previously discussed concepts around strata because both the characteristics of individual rocks and how those rocks are situated in a larger environmental context are necessary to formulate a complete understanding of past events.

At the first level, the naïve learner is unable to look at the physical properties of a rock layer and see those properties as evidence of the original formation environment. Surface processes of weathering and erosion are defined as a separate construct; this one is concerned with the understanding that rocks in the present day can show you what the local environment was like at the time they were created. The novice learner does not possess this understanding, and in fact may be unaware that the differences in formation environments or deposition events can have any effect on the physical, observable properties of a rock unit.

In general, then, the first level of this construct is one in which the student has a rudimentary awareness that the rocks come from, and may contain materials from, the Earth's past. She is not able to accurately draw conclusions about that past, regarding either how long ago it took place or what was happening then.

3.5.b GTS Level 2

The major difference between this first and second level is that the student is now able to infer a sequence of events with some level of qualitative specificity. This means that the student identifies the relative age of the rock strata in a single stratigraphic column according to superposition; i.e., the oldest rock is the lowest in the column and the youngest rock is the closest to the surface. This is true in the absence of subsequent events that could have overturned the rocks, and a learner at this developing level would be unlikely to take the possibility of rocks being overturned into account.

As previously stated, part of this construct is based on Dodick & Orion (2006)'s research on students' understanding of geologic time. The difference between levels 1 and 2 in this research-based criterion is that while level 1 represents a sort of dichotomous understanding of geologic time, level 2 represents an understanding of time as a continuum.

In the second level 2 criterion, the additive knowledge represented is that the learner is able to see that there are different, individual rocks (or rock layers) each with specific different physical properties (e.g. limestone versus sandstone), and that those properties are connected, to different types of formation events. At level 2 the learner is still unable to describe the factors that influenced those differences.

3.5.c GTS Level 3

The third level of the construct map describes a proficient, but not yet expert level understanding of geologic time. This level on the construct map is also defined by the standards articulated in the New York State Earth Science Core Curriculum (NYS ESCC) (NYSED, 2001). A student who exhibits the type of understanding described at this level of the construct should be successful when dealing with tasks from a standardized assessment aligned to the NYS ESCC, as the New York State Regents Examinations are intended to be.

The major difference between level 2 and level 3 is that at this third level, the student understands that the outcrop represents a series of specific, rather than general, events. This understanding includes the awareness that events such as erosion or folding may disrupt the superposition of stratigraphic layers, and that therefore it is not always possible to determine a complete sequence of events based on a single stratigraphic column or outcrop. It also includes an understanding that different types of rocks represent different formation environments. The level 3 student understands that some rocks may form faster (due to increased rates of deposition) and that rocks may become eroded during subsequent geologic periods/events.

In the research-based criteria, the difference from level 2 is that at level 3, the student can see connections between outcrops that are separated by distance, understanding that the same sequence of rock layers and/or the same index fossils provide evidence that those segments of the separate outcrops are the same age. At level 2 the student considers only one outcrop at a time.

In the extrapolated criteria, which deal with the student's understanding of the connection between rock properties and formation environment, the difference between level 2 and level 3 comes from an increasingly refined understanding of rock-formation processes as situated in different environments. At level 3, the student can give a broad description of the formation environment—such as general geologic events, landscape features, or the type of ecosystem present—based on sediment type, grain size, or the habitat of fossil organisms.

3.5.d GTS Level 4

The overall difference between levels 3 and 4 is that the student now understands the outcrop(s) as evidence of continual, specific changes that occurred at specific times. The layers are not merely demonstrating a series of events; they are defining a timeline. This description of the more expert level understanding is, once again, drawn from the research done by Dodick and Orion (2006) on learners' understanding of geologic time. Learners of geology tend to think about geologic time in different ways. They think about "deep time" as the whole history of the earth, defined by significant global events; the "macro" time scale. On the other side of things, they think about the history of an individual outcrop in terms of the sequence of localized events that shaped the strata in that area. Dodick and Orion maintain that it is necessary to understand timescales and the earth's dynamism in order to understand that earth's history is one of (usually) slow but continuous change, with evidence of the changes being recorded in rocks and fossils. More experienced geology students are able to relate these different types of time –the "micro" and "macro" or the relative and absolute—to one another in order to reconstruct a geologic history (including events and processes). Therefore, at level 4, the student can assign specific ages to individual geologic elements such as rock formations or fossils, in terms of both an age (how long ago) and a named time period (e.g. the Late Silurian). In level 3 the student primarily understands the ages of rocks within the relative time framework; now in level 4 the student is able to apply those temporal relationships to the established geologic timeline of earth's history. (This is different from a student who may be able to memorize or use a reference guide to identify a fossil as coming from a specific time period, but cannot say based on the positions of the fossils that one is older than the other, etc.)

In the extrapolated criteria about understanding of formation environments, the level 4 student can describe the changes that must have occurred in the environment to lead to the formation of different types of rocks. This is a progression from level 3, where the student may identify characteristics of the landscape (e.g. a river delta or a deep ocean) without connecting it to a large-scale geologic event (e.g. climate change, formation of a passive margin, rising or falling sea levels, tectonic activity). At level 4, the student is able to interpret outcrops in terms of a story, with both micro-level details and macro-level context, by making inferences about changes in changes to the local environment based on differences in rock properties from one layer to the next. Although there may be more sophisticated and refined understandings, this represents a level of understanding with significant expertise relative to the standards set out by the NYS ESCC.

3.6 Construct: Surface Processes [SP]

The Surface Processes construct deals with concepts surrounding the changes that happen the rocks on earth's surface due to chemical and physical forces. This construct is derived from and bounded by the anecdotal experience of Earth Science Teachers in New York State. Most of this experience is my own; the criteria at each level of the construct map are based on students' past responses to assignments and informal interviews. I also consulted with two other Earth science teachers. However, because this particular construct is one with virtually no grounding in existing research, I anticipate that one outcome of this work will be the revision and refinement of the criteria at each level.

3.6.a SP Level 1

I believe the naïve level of understanding is one in which students don't know the difference between weathering, erosion, and deposition; all these ideas are lumped together in one mental compartment about sediments changing on Earth's surface. At this level the student knows that something is happening to the rock—that it is being "eaten away" or destroyed—but she is not able to describe exactly what, or why. This is the least sophisticated level of thinking as described by Anderson and Krathwohl (2001): the student is merely remembering a fact, and an incomplete one at that.

3.6.b SP Level 2

At the second level of this construct, the learner is able to recognize the difference and classify various processes (for example, based on the appearance of rock, she can say that it has undergone physical weathering). At this level the student is responding with characteristics or evidence from the landscape or from a rock. Therefore, the difference between level 1 and level 2 is largely in the clarity of content knowledge, plus the ability to connect that knowledge (of definitions, properties, and cause & effects) with physical appearance-based evidence. This is comprehension level thinking.

3.6.c SP Level 3

At level 3, the learner makes connections between environmental factors and surface processes. This is the level of understanding that is measured (though in a superficial sense) by the Earth Science Regents; students are expected to know that humid environments experience more chemical weathering, while arid environments experience more physical weathering, resulting in comparatively angular landscape features. As with all the constructs addressed here, Level 3 in the construct map is intended to align with the NYS ESCC.

The student at Level 3 is also able to connect the evidence they see with a sense of “sequentiality” or timescale. This means she can “see” that a running river with fast water created a certain set of landscape features; those features are no longer just evidence of physical weathering and/or erosion that resulted in the removal of sediments. In other words, the difference between level 2 and level 3 is the difference between “what” and “how and why.” To truly achieve a level 3 understanding, the student must use more than just recall-level thinking. It is not enough to memorize the fact that running water creates v-shaped valleys (which would be a similar type/mode of knowledge as level 2); the level 3 student must be able to explain the past in the context of present evidence, describing processes and change over time. This level embodies, in a sense, the geological maxim that “the present is the key to the past,” and requires application and analysis modes of understanding.

3.6.d SP Level 4

Finally, the learner reaches a level of expertise where she is able to apply this understanding in a predictive way, by describing what changes might occur in the future and specific ways in which those particular surface processes would affect the appearance of the exposed bedrock and the shape of the landscape. The difference between level 3 understanding and level 4 understanding is one of supposition or application. The student is now able to synthesize the evidence of surface processes up to this point in time to not only deduce a series of past processes and events (which is evidence of understanding at level 3 in the construct map), but also to place them in the context of a continuum that extends into the future. The level 4 student will be able to think in terms of hypotheticals and possibilities, not just evidence and inferences. This difference between levels 3 and 4 represents more sophisticated understandings and modes of knowledge. This can be characterized as synthesis thinking, because the student is creating new ideas about possible future(s) rather than reproducing the correct ideas about existing events.

3.7 Construct: Plate Tectonics [PT]

The plate tectonics construct is about the student’s understanding of the internal drivers that affect changes at Earth’s surface, most notably the movement of lithospheric plates. The novice and expert descriptions for this construct are based on Gobert’s (2005) series of studies on mental models of plate tectonics. Gobert’s work focuses on the connection between visual or spatial reasoning and the sophistication of a student’s explanations of tectonic processes.

3.7.a PT Level 1

The naïve conception of tectonic plates held by novice learners is that they are somewhere underground. Initially, this understanding does not include the fact that the plates are moving. Typical novice-level ideas about tectonic plate movement make tectonic plates somewhat analogous to rafts floating on water, as if there were a lot of space between the boundaries, with no sense that the plates are effectively interlocking and covering the entire surface of the earth. This level is based around research on preconceptions (e.g. Gobert, 2005) that frequently exist before any formal instruction or learning about plate tectonics has taken place (though any individual learner may not necessarily hold all of these ideas). This conception of the plates as discrete chunks moving (or sitting statically, perhaps) somewhere beneath the earth's crust obscures any obvious relationship between that movement and the effects that are evident on the surface, such as plate boundary features (e.g. mountains, trenches, or ridges) or events (e.g. earthquakes, volcanism).

3.7.b PT Level 2

This level is extrapolated from the novice/expert dichotomy described in Gobert's research, along with my anecdotal experience as a teacher (based on both student work and conversations with students). At this second level, the learner has developed an incomplete, but no longer overwhelmingly incorrect, conception of the physical nature of the tectonic plates as well as the events that may result from interactions at plate boundaries. This correctness is the major difference between levels 1 and 2, and represents a move from a conception based mostly around intuition and incidental awareness of these ideas to one based around exposure to scientific representations of the topic. The level 2 student can accurately describe the nature of the plates and their location that composes the outer layer of the Earth (this is comprehension level thinking). She can identify plates as existing on either side of a boundary where they touch, but may have an unclear conception about the continuity of the plates between boundaries. She may or may not clearly identify the tectonic plates as being definitionally equivalent to the lithosphere.

Due perhaps to the timescales involved, as well as to their intermittent nature, I believe that most students understand discrete and short-term events like earthquakes before they understand continuous, long-term events like mountain building.

Once the learner has an idea of plates as part of the crust that moves, she can attribute physical effects at the surface (such as earthquakes) to that movement. This understanding may exist even if the learner continues to have an incorrect mental image of the nature of the plates (for example, still imagining them being somewhere under the crust), because her understanding of the interactions at the plate boundaries allows her to think productively about the surface effects. At this comprehension-based level of thinking, the student does not connect the events at the boundaries to the processes associated with the rock cycle, but understands these two related concepts independent from one another.

3.7.c PT Level 3

At first, these surface effects are understood in terms of short-term events. A more sophisticated understanding includes the ongoing cause-and-effect dynamic between tectonic motions and crustal events or features. Although a student at level 3 does not necessarily describe a specific type of convection cell or the difference between, for example, the various types of faults characteristic of different boundary motions, she is able to explain that plate movements are caused by convection currents within the mantle, and that these movements lead to the destructive and constructive effects on the lithosphere at boundaries between plates. She is able to specify differences between convergent and divergent boundaries, and identify the connection between plate boundaries and tectonic events such as earthquakes and volcanic activity. This level of understanding is the same as that described by the NYS ESCC. The difference between level 2 and level 3 of this construct is that this mode of thinking requires analysis and some synthesis thinking modes, rather than the less sophisticated comprehension mode of thinking that characterizes level 2.

3.7.d PT Level 4

Level 4 of this construct is once again derived directly from the literature, primarily from Gobert's (2005) research on the differences between novice and expert thinking about plate tectonics. As the student begins to understand it in terms of a system or a continuous process, she is able to fully understand the causal nature of the plate actions (not just the effects, but how the process really works and what observables it creates). The difference between levels 3 and 4 is that at this level, the student is making productive, meaningful connections between invisible or hidden processes beneath Earth's surface and the specific, varied effects that are observed at Earth's surface. The student at level 4 can describe not only the existence but also the nature of the convective mechanism driving plate tectonics. This level is characterized by synthesis and evaluative thinking.

For example, this student can describe the relationship between ongoing processes at a subduction zone—such as the growth of mountains and formation of trenches—and more "acute" events like the eruption of a volcano. She can also explain why some effects are common to all boundary types (e.g. earthquakes) while some are boundary-specific. In the context of this assessment task, the student might point to evidence of folding and faulting as the continues and abrupt effects of the same convergent movement, driven by downwelling currents in the asthenosphere. In contrast, at level 3, this student would identify the features as a result of compression or moving together, but would not identify the nature of the mantle convection cell or distinguish between fast and slow processes.

3.8 Construct: Topographic Maps [TM]

This construct is about a skill, rather than about pure content knowledge. Students in Earth Science are often required to use maps that represent features of the earth via topographic contour lines, isolines that represent some other form of numerical data, color- or pattern-coding, or other means of graphic symbolism. In order to do this effectively, a student must be able to derive or understand the relationship between the map and the information about real Earth features that it represents. This often involves spatial thinking: the student might need to visualize what a three-dimensional surface looks like based on a two-dimensional representation, or she may need to visualize a ground-based perspective based on the map's birds-eye-view.

Kastens and Ishikawa (2006) describe this as a spatial thinking task whose development is mediated by geoscience expertise. Experts develop schemata that allow them to recognize and categorize meaningful patterns based on characteristic shapes captured in the contour lines of a topographic map. In contrast, novices do not have the experience that allows them to "see" the shapes embodied by the contour lines. In the process of generating these mental schemata, learners develop intermediate skills and thinking patterns that allow them to make sense of the topology first by more of a "brute force" methodology, then with gradually increasing sophistication. In order to picture the contours of a landscape as seen by an observer on the group, the student must be able to envision different frames of reference. Typically, learners are able to use relative frames of reference (such as determining where two points are relative to each other) before they are able to use absolute frames of reference (such as identifying a quantitative difference between two points in distance, direction, or elevation) (Kastens & Ishikawa, 2006; Rapp & Uttal, 2006). By creating a task that elicits responses along this continuum, we are able to characterize a student's thinking according to its level of sophistication, rather than simply coding it as correct or incorrect.

3.8.a TM Level 1

Although the research by Kastens & Ishikawa (2006) upon which this part of the construct map is based frames the novice/expert dichotomy for this skill largely in terms of a deficit at the novice level, it has been reframed here in terms of what novice learners tend to do, rather than in terms of what they fail to do. The purpose of this reframing is to facilitate better correlations between the construct map, the assessment task, and the outcome space. Since assessments are intended as observational tools, they are better suited to measuring or recording what students *do* than what they do *not* do.

At the first level, the skill is characterized by naïve thinking. The student looking at the map does not know what the contour lines mean. She may be able to discern some basic ideas from the self-contained information in the map itself, (e.g. labels, titles, or scales), but this is less related to the spatial thinking component of the skill. This level of thinking is what you would expect from a student with little experience reading or looking at topographic maps. Therefore, she tends to make statements or draw conclusions that are incorrect.

3.8.b TM Level 2

At level 2, the student is beginning to understand the topographic representations in the context of the map itself. This means she can identify features of specific locations, based on the labels on the contour lines or perhaps the spacing of the contour lines. At this second level this skill is still greatly facilitated by iconic representations on the map, as described by Kastens & Ishikawa (2006). Based on this finding in the literature, the student might be able to say "this is at the top of a hill" or "this is a river" (correct interpretations of local topology) but not say that the river valley and the hill are related because the presence of the river valley results in a surrounding region of higher elevation (the understanding of a larger context is missing). The student at level 2 can identify hills based on the closed topographic curves, but in the absence of numbers on the uppermost line will not necessarily state which hill is higher or which side of a hill is steeper. When drawing profiles, this is manifested in profiles that have the approximate correct shapes but that do not correspond to the appropriate heights or relative elevation gradients.

The difference between Level 1 and Level 2 is that at level 2, the student is now familiar with the features and conventions of topographic maps, but understanding of those features is still in early stages of development. This means she does not yet have a schema-based understanding of the spatial features represented on topographic maps, and her facility with reading and interpreting the maps is correspondingly rudimentary. This is comprehension-level thinking.

3.8.c TM Level 3

The first item in this level of the construct map is based on the Kastens and Ishikawa (2006) research. Although it is not stated explicitly in the paper, the natural progression between level 2 (at which the student is reliant on iconic representations) and the expert level should include the ability to identify features or places on the map using symbolic representations that do not necessarily resemble the things they represent. Since this skill is required by the NYS ESCC, it is placed here in level 3.

The remaining items in this level are based on personal teaching experience and student interviews, but are still informed by the work of Kastens & Ishikawa (2006) and Rapp & Uttal (2006), along with the standards outlined in the NYS ESCC. At this point, the student can describe what the landscape looks like from the ground. This description may be graphic or written. This means that she can interpret the map in terms of an actual, continuous landscape, not just a collection of topographic features. Profile drawing may be somewhat rote but the student recognizes that the profile represents a particular physical viewpoint. This is different from level 2 because the cognitive connections between the representation (contour lines on the map) and the real world (actual topography of the landscape) are stronger, in part because of more sophisticated spatial reasoning skills.

At level 3, the student can also consider and analyze components or locations on a map in terms of a larger context. For example, she would be able to make comparisons between two or more locations to find the steepest slope based on the separation between contour lines, or to describe the relative change in the landscape as you travel along a particular path. This represents a change from level 2 where the map was seen as a collection of features, but didn't really add up to a whole picture.

In general, this growth from level 2 to level 3 is the result of an increased facility with spatial thinking including mental rotation of two-dimensional images and extrapolating a three-dimensional mental image from a two-dimensional representation. Additionally, the ability to identify landforms based on the characteristic shape or pattern in the contours is the beginning of schema-based thinking (as described by Chi et al., 1988).

3.8.d TM Level 4

At this expert level, the student is able to use the map in the way that a geologist would. In real-world practice, this would mean that she can use the map as a tool in the field, by finding locations or interpreting local geology and topography. In the context of the assessment task, it means using the map as a tool to support inferences, analysis, and hypothesizing about the geological behavior and topological effects of the represented region. This kind of understanding requires evaluation and synthesis modes of thinking.

3.9 Task Description

A performance assessment task was developed that requires students to interpret a topographic map in conjunction with diagrams of outcrops at specific locations and actual hand samples of rocks found at those locations. (See Appendix C for student response form.) All of the materials are manipulable by the student during the assessment. In order to complete the task, students are required to record their responses in a variety of formats, including graphs, diagrams, calculations, and both short and extended written descriptions. The task is designed to take approximately one hour for students to complete, working individually.

The primary physical components of the performance assessment task are a topographic map, diagrams of outcrops located on the map, and hand samples of rocks representing the bedrock in the mapped area (see Appendix B for a detailed list of assessment materials). The map will represent an area typical of slightly inland east coast, somewhere within the Appalachian orogenic belt, characteristic of most of the Hudson Valley in New York State. This is general enough that it could apply to anywhere along the east coast that had significant deposition during the Devonian period, and it is typical of the type of geological history content included on New York State Earth Science Regents exams.

The region in the map is made of sedimentary rocks, primarily of marine origin. The fossils (crinoids & other shelled creatures like *Centroceras*) in the rocks indicate that the environment at the time was a shallow equatorial sea. The grain size of the sedimentary rocks as well as the types of fossils in some layers show that the sea level rose and fell during the period when the bedrocks was being formed, a time period spanning several hundred million years during the Paleozoic era (from the late Silurian to the early Triassic periods). The different grain sizes of the sedimentary rocks indicate that some were deposited in very shallow water, where wave action and tidal changes allowed the surface may have been periodically exposed to air, creating thin layers of slightly larger sediments. Other layers were deposited when the sea level was higher, making it possible for layers of fine sediment (created by erosion on the highlands) to be deposited below the water's surface. The changes in sea level were due to both global changes and local changes caused by tectonic activity that moved the land up and down. Not all geologic epochs within this timespan are represented in the bedrock layers, due to erosion of some layers and interstitial mountain-building events.

The landscape has been shaped over time by a series of orogenies as well as local surface processes (weathering & erosion by water, mostly). The originally horizontal layers were deformed later by collisions at a tectonic plate boundary. The evidence of these collisions can be seen in the tilting, folding, and in particular the reverse thrust faults that occur in bands throughout the region. The associated fault propagation folds create tight syncline / anticline pairs beside the thrust fault. These folds suggest that the rocks were deformed fluidly for a long time before the fault formed and caused a break that offset the layers. More recently, the movement of a river through the region has caused weathering & erosion that cut a v-shaped valley into the landscape. The asymmetry of the valley walls indicates that some of the rock layers were more susceptible to weathering than others. The rounded, hilly contours of the landscape are due to the weathering patterns of a humid environment.

The end goal of the performance assessment task is for the students to describe and illustrate the changes in the landscape and local environment over a long period of time, beginning with ancient geologic history and proceeding to a predicted future environment.

Before completing this integrated task component, which requires them to apply a variety of content knowledge and skills, they will complete several smaller items. The purpose of these items is both to scaffold the students' thinking and to target specific constructs and levels within those constructs.

As Figure 3 shows, each component of the cognitively based performance assessment task is rich enough in content that it addresses multiple standards, constructs, or levels within a construct. This is in accordance with a key design principle of my strategy for making this approach to measurement successful. One of the major barriers to using performance tasks as a significant component of large-scale assessment programs is that they are more complex and time consuming than traditional written assessment items (Dunbar, Koretz, & Hoover, 1991). One way to make it possible for performance tasks to compose a larger portion of assessments is to design these tasks so that they are able to measure multiple skills or understandings simultaneously. This is distinctly different from a multiple-choice item that requires the application of multiple skills or ideas; while the latter merely confounds them—effectively reducing the construct validity of the assessment—the more open-ended nature of the performance task helps ensure that it will allow for the observation of more than one construct at a time. Note that "open-ended" does not imply that a task is unstructured. Quite the contrary: the performance task items must be carefully and intentionally designed to elicit specific types of thinking and problem solving, as demonstrated via the students' actions in response to the task itself.

Table 1: Task Item Correlation to Construct Map

Assessment Task Item & Description		Geologic Time & Stratigraphy	Surface Processes	Plate Tectonics	Topographic Maps	NYS ESCC Standards
1a	Draw profile				1 – 4	6.3; 4.2.1q
1b	Describe profile				1 – 4	6.2; 4.2.1q
1c	Draw path				2 – 4	6.3; 7.2
1d	Annotate path				2 – 4	6.2; 7.2
1e	Explain reasoning				1 – 4	6.2; 7.2
2a	Explain profile		1 – 4		1 - 3	6.2
2b	Choose rock		1 – 2			4.2.1r; 4.2.1t; 4.2.1u
2c	Explain rock		1 – 3			4.2.1r; 4.2.1t; 4.2.1u
2d	Draw no weathering		2 – 4			4.2.1s; 4.2.1t; 4.2.1u
2e	Explain differences		1 – 4		1 – 2	6.2; 4.2.1s; 4.2.1t; 4.2.1u
3a	Classify rocks	2 – 4				4.2.1w; 4.3.1c
3b	Date fossil rock	1 – 4				4.1.2i
3c	Outcrop correlation	1 – 4				4.1.2j
3d	Evidence of motion			2 – 4		4.2.1n
3e	Explain mechanism			2 – 4		4.2.1k; 4.2.1l
3f	Draw arrows			1 - 3	2 – 3	6.3; 4.2.1k
4a	First event	2 – 4	2 – 3			6.5; 4.1.2j, 4.2.1r; 4.3.1c
4b	Time of fossil	2 – 4	2 – 3			6.5; 4.1.2j; 4.2.1r; 4.3.1c
4c	Time of movement		1 – 4	1 – 4		6.5; 4.2.1n; 4.2.1p;
4d	Future landscape		1 - 4	2 – 4		6.5; 4.2.1p; 4.2.1r; 4.2.1u

The mapping of task items to the construct map reveals that the plate tectonics construct is not measured as fully or rigorously by this performance assessment task than the other constructs. While this assessment may not be sufficient to provide a complete measurement of a student's understanding of plate tectonics, it is conceived as, ideally, one in a series of performance assessments. In this model of assessment, the student would be given the opportunity to demonstrate understanding of the mechanisms driving plate tectonics in a separate task; her responses from multiple tasks could then be used together to describe her understanding in terms of the sequence articulated by the construct map.

3.10 Embodiment of Design Principles Within Performance Assessment Task

In order to demonstrate the way in which the design principles are embodied in the task design, I will describe individual components of the larger performance assessment task. The first design principle demands richness. In the initial part of the task, the student encounters the materials for the first time and is asked to make sense of the information contained within them, beginning with a topographic map. The map shows a hypothetical region typical of areas within the Hudson Valley that experienced significant Devonian deposition. Topographic features of the map include a river valley with asymmetrical walls indicating that some of the rock layers were more susceptible to weathering than others, and rounded, hilly contours of a landscape weathered by a humid environment. Additional materials include a number of sedimentary rock samples, one of which is rounded due to weathering in moving water.

The student is asked to construct a topographic profile illustrating the shape of the landscape when viewed from a certain point on the map. Then she is asked to explain the effect the river's presence on the local environment, first by identifying the salient features of rocks transported within the river, then by drawing a second profile that shows how the landscape would look from the same viewpoint if the river had not formed there. The student is asked to respond to written prompts in addition to creating graphic representations, including one that asks for an explanation of the difference between the two profiles.

This example demonstrates one of the key differences between a performance-based approach and a more traditional approach to cognitively based assessment. Because the nature of performance assessment precludes the inclusion of numerous pithy questions in favor of more extended hands-on work, it becomes necessary to design tasks that are capable of measuring student understanding in more than one construct at a time. Adherence to this design principle achieves several goals: first, it allows the ability to measure multiple constructs within a reasonable time period; second, it allows us to measure the same construct within a variety of contexts; and third, it enables us to observe the student's use of connections between different topics during the problem-solving process. In this example, I intended to score the students' responses to a single prompt with regard to both the Topographic Maps construct and the Surface Processes construct.

The second design principle states that tasks must be authentic. The use of topographic maps to represent landscape features is well established within both the scientific and the professional community. Geoscientists also use their understanding of Earth processes to make conjectures and suppositions; this practice is embodied in the oft-repeated motto of geologic history that states “the present is the key to the past.” Evidence of current and ongoing processes is assumed to be indicative of processes that occurred in the distant past. These types of tasks are often included on standardized tests in geology content areas, but their used is restricted to closed- or limited-response items. By keeping in mind the ways in which Earth scientists might use these types of materials—that is, the types of questions they would be considering, and the products they would create in the process of thinking through those questions—we are able to create similar tasks that are at an appropriate level for measuring student understanding.

I have also ensured that the task in which the students must engage are aligned with the New York State core curriculum in Earth Science, making it useful for assessing the understanding that is intended to be measured by the state’s standardized tests. This is in accordance with the third design principle. Constructing profiles, recognizing features of physical weathering caused by abrasion at the bottom of a running stream, and distinguishing parts of a landscape shaped by river erosion are all things that a high school Earth science student in New York State are expected to be able to do. However, while the standardized test associated with this curriculum produces a binary measure—the only possibilities are that the student is completely correct or completely incorrect on their ability to perform each of these activities —the cognitively-based performance assessment task is able to measure the degree of sophistication present in the student’s thinking around these problems by eliciting responses that may be placed on the continuum defined by the construct map.

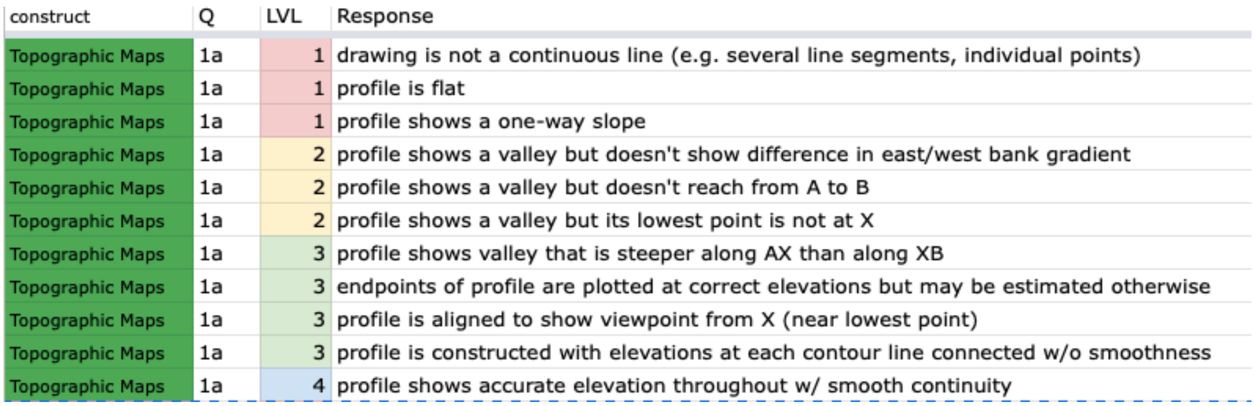
The fourth design principal states that tasks must be based on research about learning and thinking. In one part of the task, for instance, the student must use the features of the topographic map to visualize what the landscape would look like to an observer standing at a particular location. Kastens and Ishikawa (2006) describe this as a spatial thinking task whose development is mediated by geoscience expertise. Experts develop schemata that allow them to recognize and categorize meaningful patterns based on characteristic shapes captured in the contour lines of a topographic map. In contrast, novices do not have the experience that allows them to "see" the shapes embodied by the contour lines. In the process of generating these mental schemata, learners develop intermediate skills and thinking patterns that allow them to make sense of the topology first by more of a "brute force" methodology, then with gradually increasing sophistication. In order to picture the contours of a landscape as seen by an observer on the group, the student must be able to envision different frames of reference. Typically, learners are able to use relative frames of reference (such as determining where two points are relative to each other) before they are able to use absolute frames of reference (such as identifying a quantitative difference between two points in distance, direction, or elevation) (Kastens & Ishikawa, 2006; Rapp & Uttal, 2006). By creating a task that elicits responses along this continuum, we are able to characterize a student's thinking according to its level of sophistication, rather than simply coding it as correct or incorrect. This is an example of the use of research to inform task design, and is a cornerstone of cognitively based assessment.

3.11 Outcome Space & Scoring Procedure

In order to create a scoring procedure that can be consistently and objectively implemented by different readers, it is necessary to describe the characteristics of potential student responses at every level of each construct. This description, or collection of descriptions, is known as the outcome space (Wilson, 2005).

The outcome space is, then, an operationalized version of the construct map, made specific for the particular items in the performance task. The construct map is used to identify and characterize the possible responses a student might have at each level of understanding. This means that the definition of the outcome space makes up the bulk of the scoring procedure. Then, this outcome space becomes a tool for assigning a level to the student's response to each item. The scores from multiple items can be used to triangulate a more holistic score describing the student's overall level of understanding of each construct.

Each opportunity that the student has to produce some sort or response is treated as a different item. Figure 3 shows the outcome space for the first item requiring on the assessment task that requires the student to construct a topographic profile (see Appendix E for the complete outcome space for all items on this performance assessment). Note that there are multiple possible responses at each level of the construct map; although each of these represents a slightly different outcome, all responses coded at level 2 represent the same approximate level of expertise and sophistication. Scoring each item according to the outcome space generates a number of scores for every construct. This collection of scores is then used to characterize the student's overall level of understanding within each construct. The score report also indicates the consistency of the student's response, i.e., whether the student consistently scored at level 3 for a given construct or if her responses ranged between levels 2 and 4 depending on the item. Because this is the first iteration of this particular performance assessment task, much of the outcome space was initially based upon my experience as a teacher: I used this experience to predict what many common responses would be, and then used the construct map to assign those responses to a level. During this first iteration of this performance assessment task, the outcome space was modified and added to, based on the student responses observed from the initial rounds of data collection. This is similar to the norming process required by teacher teams who score the New York State Regents Examinations. All additions to the outcome space were agreed upon by a team of scorers, and guided by the level descriptions in the construct map.

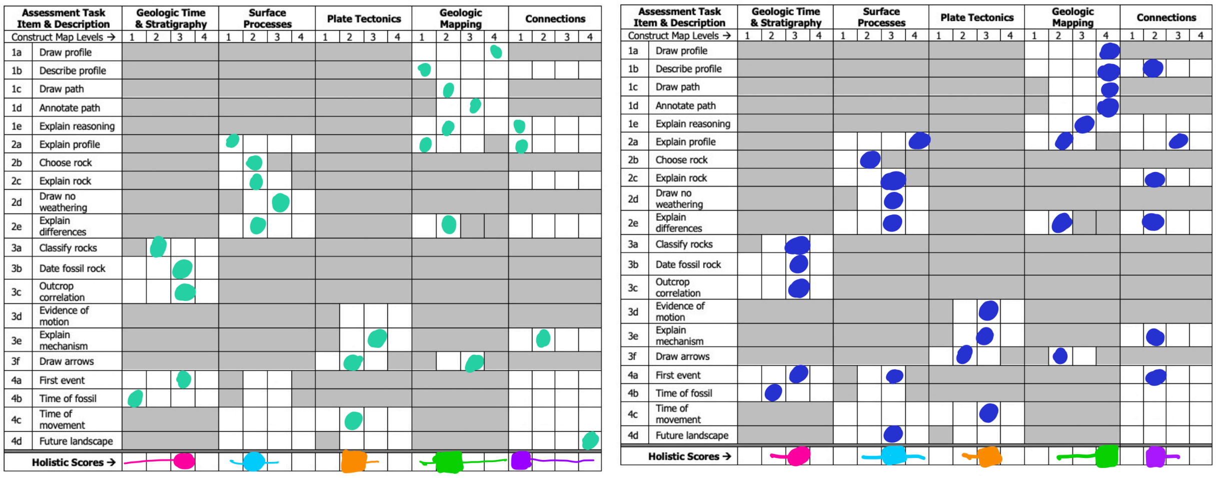


construct	Q	LVL	Response
Topographic Maps	1a	1	drawing is not a continuous line (e.g. several line segments, individual points)
Topographic Maps	1a	1	profile is flat
Topographic Maps	1a	1	profile shows a one-way slope
Topographic Maps	1a	2	profile shows a valley but doesn't show difference in east/west bank gradient
Topographic Maps	1a	2	profile shows a valley but doesn't reach from A to B
Topographic Maps	1a	2	profile shows a valley but its lowest point is not at X
Topographic Maps	1a	3	profile shows valley that is steeper along AX than along XB
Topographic Maps	1a	3	endpoints of profile are plotted at correct elevations but may be estimated otherwise
Topographic Maps	1a	3	profile is aligned to show viewpoint from X (near lowest point)
Topographic Maps	1a	3	profile is constructed with elevations at each contour line connected w/o smoothness
Topographic Maps	1a	4	profile shows accurate elevation throughout w/ smooth continuity

Figure 3: Outcome Space for Sample Assessment Item.

As this example shows, the outcome space allows for different characterizations of the responses at each level because it is based on the construct map. We don't expect to see the exact same wrong or incomplete answers—or even the exact same correct or complete answers—from each student. There is a range of responses that could be produced by students whose thinking falls within a given level of expertise according to the continuum on the construct map. By describing a panoply of possible responses, the outcome space allows us to make more meaningful inferences about the thinking behind students' responses than a typical right-wrong dichotomy would.

The scoring procedure provides instructions for any third-party scorer with sufficient Earth Science expertise to interpret a student's responses and assign the student a score in each of the five constructs, using the outcome space. This protocol and the associated score reporting tool are shown in Appendix D. Recall that the purpose of the cognitively-based performance assessment task is to characterize the student's thinking, skills, and understanding in this content realm, in addition to the goal of providing a simplified score that may be used for normed comparisons between students, schools, and so on. The score report for an individual student will indicate her subscore for each construct on an item-by-item basis, as well as an overall score within each of the five constructs. The score will be numerical—the student's responses will be characterized as level 1, 2, 3, or 4 for each construct—but will also indicate the degree to which those responses are internally consistent. In this way, it will be possible to see the difference between a student whose responses consistently fall within the criteria defined for level 3 and a student whose responses range in sophistication from levels 2 to 4. Figure 4 shows two examples of scores with different consistency profiles, generated by teacher scoring teams during the course of this study.



Assessment Task Item & Description		Geologic Time & Stratigraphy				Surface Processes				Plate Tectonics				Geologic Mapping				Connections			
Construct Map Levels →		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1a	Draw profile																●				
1b	Describe profile													●							
1c	Draw path														●						
1d	Annotate path															●					
1e	Explain reasoning														●			●			
2a	Explain profile					●								●				●			
2b	Choose rock						●														
2c	Explain rock						●														
2d	Draw no weathering							●													
2e	Explain differences						●								●						
3a	Classify rocks		●																		
3b	Date fossil rock			●																	
3c	Outcrop correlation			●																	
3d	Evidence of motion																				
3e	Explain mechanism											●							●		
3f	Draw arrows										●					●					
4a	First event			●																	
4b	Time of fossil	●																			
4c	Time of movement										●										
4d	Future landscape																				●
Holistic Scores →				●			●				●				●			●			

Assessment Task Item & Description		Geologic Time & Stratigraphy				Surface Processes				Plate Tectonics				Geologic Mapping				Connections			
Construct Map Levels →		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1a	Draw profile																●				
1b	Describe profile																●		●		
1c	Draw path																●				
1d	Annotate path																●				
1e	Explain reasoning															●					
2a	Explain profile								●						●					●	
2b	Choose rock						●														
2c	Explain rock							●											●		
2d	Draw no weathering							●													
2e	Explain differences							●							●				●		
3a	Classify rocks			●																	
3b	Date fossil rock			●																	
3c	Outcrop correlation			●																	
3d	Evidence of motion											●									
3e	Explain mechanism											●							●		
3f	Draw arrows										●				●						
4a	First event			●				●											●		
4b	Time of fossil		●																		
4c	Time of movement											●									
4d	Future landscape							●													
Holistic Scores →				●				●				●					●		●		

Figure 4: Score Reports Generated by the Scoring Procedure.

Chapter 4: Methods

The four research questions that this work intends to address, in order to examine coherence between different facets of assessment, revolve around the concept of construct validity. I am framing validity in terms of comparisons between the constructs, students' thinking while they are engaged with the performance assessment task, the written responses they produce, and the scored interpretation of those responses. The four research questions represent pairwise comparisons between these different parts of the assessment triangle, as described in the literature section (NRC, 2001). Student thinking was captured via an audiotaped and annotated think-aloud protocol; all other data sources are the product of the administration and scoring of the performance-based assessment task as described in the previous chapter.

4.1 Subjects and Data Collection Process

The human subjects for this study were twenty-two current or recent (within one year) students of Regents level Earth Science. It was important that the students had recently studied Regents level Earth Science because the NYS Earth Science Core Curriculum is what I used to bound the scope and sequence of the assessment task (as described in the section on design principles in chapter 3), as well as to inform the development of the outcome space with respect to its alignment with the construct. They were all in grades 9 – 12 and attended three different public schools in Brooklyn, New York. In accordance with IRB requirements of the New York City Department of Education, students and their parents provided written consent to via letters sent home. My goal was to have student participants who represented a range of achievement, skill, and experience levels. In addition, I intentionally recruited subjects from different demographic groups, to the extent possible within the schools where they were enrolled. Of the twenty-two subjects, seventeen identified as female, four as male, and one as nonbinary. Thirteen of the students were Black, eight were Latine, two were white, and one was Asian (these numbers total to greater than 22 because two students identified as multiracial). Ten students indicated a hispanic ethnicity. Four were students with disabilities (SWD), and two were English language learners (ELL). Although these demographic groups are too small for a meaningful statistical comparison, I believe it is important that the students represent the diversity of those served by public schools in New York City. Assessment research, policy, and practice have often ignored considerations of social justice and equity (e.g., Bell, 2007). At minimum, we should demand that new research in this area is conducted with diverse racial and gender identities.

Because I use quantitative statistical tests as part of my analysis procedure, I needed enough data points to give these results statistical power and significance. Conversely, because both the collection and coding of think-aloud data is highly time consuming, I attempted to set a reasonable target number. Wilson (2003, 2005) recommends 50 subjects for the pilot testing phase of assessment item development. Although I do not have this many subjects, I was able to collect data points in excess of this number for each construct. Each construct is measured with at least four different items on the assessment task—some with as many as nine—so this translates to potential range of approximately 80 – 200 data points for each construct in the students' scores. Coding for think-alouds and the written responses also included at least this many data points, and often many more.

Students who participated in the study were asked to complete the entire cognitively-based performance assessment task outlined in Chapter 3. In the course of completing the task, each student filled out the associated response sheets, with written or otherwise recorded-on-paper responses to each part of the assessment task. Although many of the responses required students to write, there are also graphical responses, such as a constructed topographic profile or an illustration; see Chapter 3 and Appendix C for a more thorough description. While the students worked, they were be audiotaped responding to a concurrent think-aloud protocol designed to help them accurately describe their thinking around the different parts of the assessment task. This protocol uses the approach to getting at thinking via verbal reports as described by Ericsson and Simon (1993). The think-aloud protocol, included in full in Appendix F, prompts students to fully articulate their thoughts but was designed to avoid guiding or influencing the students' problems solving in any way. As such it is quite open-ended. This protocol is based on examples used in studies of math assessments designed according to the BEAR assessment system (e.g. Wilmot, 2009). In order to allow for the observation and coding of behaviors that the students engaged in during their work (for example, moving rocks or rotating the topographic map), I kept written notes that were later added to the transcribed think-aloud data.

Although the task was designed to take approximately one hour, I did not impose any time limits upon the subjects. They were able to complete the assessment in as much or as little time as they desired. Additionally, I did not require students to respond to all items. Some of them chose to leave items blank when they felt they did not have sufficient knowledge or skills to respond. This was an infrequent occurrence and did not significantly reduce the amount of data available for my analysis.

4.2 Data Reduction and Analysis

I followed these steps to prepare and analyze the collected data.

1. Score student written responses in collaboration with a team of Earth Science teachers, in accordance with pre-defined scoring procedure.
    a. Calibrate scoring procedure using 20% of student response documents.

    b. Establish inter-rater reliability using a different 20% of student response documents.

2. Code audiotaped think-alouds
    a. According to construct-based categories and levels
    b. Identifying student reasoning strategies within each construct
    c. Additional emergent codes to flag and characterize:
        i. Nature of student struggles and their responses to difficulty
        ii. Students' physical manipuations of performance assessment mateirals
        iii. Connections between constructs
3. Code student written responses
    a. According to construct-based categories and levels
    b. Using more open coding to capture full breadth of responses
4. Examine strength of relationships via regression analysis.

These steps are explained in more detail in the following sections.

4.2.a Data Reduction

The first step in data preparation was to score the student response booklets, using the scoring procedure centered around the outcome space (see Appendix D for scoring instructions). As seen in the scoring matrix, this procedure resulted in up to 30 item-based scores for each student as well as five holistic scores representing the five major constructs measured by the performance assessment.

Since the response booklets are sources of two different types of data—the scores generated from the prescribed scoring procedure, and the emergent codes categorizing the responses—it was important to treat them carefully so as not to influence either data component.

Therefore, the scoring procedure was applied first, with the response booklets being scored blindly (i.e. student identifying information was removed before the scoring took place) by an independent scoring committee of Earth Science teachers from the New York City department of education. The scoring procedure was conducted in accordance with the same general protocols used during the grading for New York State Regents Examinations. I chose this approach because I wanted to replicate aspects of the assessment system that influence the determination of student scores in our current high-stakes assessments in New York. Since the Regents are scored by committees of teachers local to each district, so should the assessment in this study.

The scoring committee consisted of six individuals licensed to teach high school Earth Science in New York State who volunteered to assist with this portion of my research project. As a group, we undertook a norming process wherein we calibrated the scoring procedure to ensure that the outcome space and score-generating algorithm would be applied in a uniform manner. To do this, we looked at responses from five different student booklets covering all five constructs, and compared the responses to the outcome space for those items. We discussed the alignment between student responses and outcome space, and came to consensus about the most accurate score(s) for each student response. The scoring procedure allows for a single response to be scored at more than one level, in the instance that students demonstrate thinking patterns characteristic of multiple levels on the construct map. This represents a major difference from the “right or wrong” model used in science Regents scoring, so this point was carefully addressed during the norming process.

Following the norming process, teachers worked in teams of three to score student booklets. Each assessment booklet was scored by one teacher using the scoring instructions and form in Appendix D, then reviewed by a second scorer. If the teachers disagreed how any item should be scored, it was referred to the third scorer. Holistic scores were verified in the same way. I then entered scores for each item and construct into a data spreadsheet in the form of integers (1, 2, 3, or 4). If items were unscored due to insufficient information in the student response, I left the data cell blank. I did not assign a score of zero to any items.

To facilitate a correlation analysis and a comparison between the different forms of data, each of the four data components (construct map categories on assessment items, think-aloud responses, written responses, and response booklet scores) needed to be captured with quantitative values. The different parts of the assessment task were effectively coded already because each one has been designated an indicator of one of more constructs. In other words, this is the only data source that is not subject-dependent—it is the same for each student. Both these designations as well as the score output had two aspects: a construct and a level (1, 2, 3, or 4). Therefore, the main coding scheme for think-alouds and written responses used the same structure.

4.2.a.1 Coding Written Responses. The students' written responses were coded in NVivo 12 research data analysis software according to the same multi-leveled coding scheme that indicated construct categories and levels within them. The purpose of coding the written responses, in addition to scoring them according to a pre-set procedure, is to capture the true breadth and variety of the students' output without being limited by the outcome space. I coded the written responses separately from the teachers who conducted the scoring procedure, and without referring to the outcome space, in order to avoid creating artificially high agreement between the different data sources.

When coding the response booklets, I coded each item in turn, first looking for responses that reflected aspects or categories of the construct map. In these instances, I did not limit the constructs to those intentionally targeted by each item. Responses that did not correspond to any description in the construct map were pooled separately, then revisited in terms of the broad definition of the construct. In this way, codes for the written responses were able to capture additional facets of a construct that were not defined by the outcome space. These codes provide a characterization of what students are saying about the construct beyond the specific criteria prescribed by the assessment system ahead of time.

To establish inter-rater reliability, a second coder who had not been involved in scoring the assessments coded a subset of items comprising a minimum of 20 instantiations of each construct. This second coder was also a licensed Earth Science teacher in a New York City public high school. Initially, our level of agreement ranged from 75% to 90%. For constructs with agreements of 80% and lower, we discussed and came to agreement about interpretation of the construct map with respect to the student written responses. I recoded selected items using this revised correspondence to the construct map. After recoding, inter-rater reliability for the coding of student written responses was established for 85% - 95% across constructs. More detailed notes about the inter-rater reliability are included in Appendix G.

4.2.a.2 Coding Think-alouds. The data from think-alouds was in the form of digital audio recordings. These were transcribed and imported to NVivo 12. In accordance with New York City Department of Education IRB requirements, only the transcripts were used for coding and analysis, rather than the original data containing student voice recordings. When my notes from the task administration showed that the student had engaged in significant nonverbal behaviors, such as manipulating the physical assessment task materials or gesturing with their hands, I annotated the transcripts to include this information. The coding for the think-alouds had two main purposes. The primary goal was to capture the full breadth of the student's understanding as articulated by the student's verbal explanation. These codes were done in the same scheme as the scores, with each code indicating a construct and a level. During this phase of coding, I was looking for instances where the students' think-aloud performance aligned with a description in a particular level of the construct map. This level of coding identified the student's understanding of content, and the level of sophistication they demonstrated in that understanding. Within these codes as major nodes, I created additional subcodes to describe some of the most common patterns in student responses.

The secondary goal was to capture any demonstrated evidence of the students' thinking or problem-solving strategies. This was manifested in both their verbal reports about what they were thinking and their physical actions. These codes were emergent and descriptive in nature. I developed additional categories based on patterns in student responses. The codes in this phase identified the following:

- strategies used to interpret or understand tasks (such as rephrasing or starting with familiar ideas)
- attention (e.g. re-reading instructions or looking at specific materials)
- any actions involving the task materials (e.g. picking up rock samples, rotating maps, reorganizing outcrop diagrams)

I used the same procedure to establish inter-rater reliability for the think-aloud codes as for the codes on written responses. In this case, two of the constructs had inter-rater reliability below 80%. The revised mapping to the cognitive model and recoding led to a final inter-rater reliability of 85% - 95% across constructs for think-aloud codes.

These multiple phases of data reduction resulted in several data points for each student, both quantitative (levels within each construct) and qualitative (the additional emergent codes for think-alouds and written responses). All the quantitative data points were transferred to a comprehensive data spreadsheet for statistical analysis purposes.

Because the scoring procedure generates holistic scores for each construct from the pattern of scores on individual items, I needed to create a holistic value for the think-aloud codes and written response codes. I used a similar approach to the algorithm in the scoring procedure. For every student, I tallied all the codes designating construct-specific levels in order to identify the most frequently coded level for each construct. In most cases, this was sufficient to arrive at a holistic value. When there was a tie between two levels, I used tallies at adjacent levels within that construct to weight the result, and selected the tied level with the higher adjacent weight. If this tiebreaker procedure was inconclusive or unavailable, I selected the higher of the two tied levels. I applied this same method to the think-aloud code data and the written response code data. In this way, I arrived at a holistic think-aloud value and a holistic written response value for each construct, per student. These holistic values differed from the holistic scores in an important way: they were generated by a set of data points bounded only by the levels in the construct map. On any item, codes could be assigned to any construct and any level (1 – 4). This is different from how the scoring procedure limited scoring on each item to certain constructs, and certain levels within those constructs. In theory, this could have resulted in holistic values derived from a much larger number of data points than the holistic scores. In practice, I observed that the number of codes for each construct was similar to the number of scores, for individual students. However, it was the case that codes were assigned beyond the prescribed levels in the scoring procedure, for almost every subject.

4.2.b Data Analysis

The goal of my analysis is to examine the degree of agreement or matching between the different aspects of assessment design. This is different from the construct-validation approach that has been described by Wilson and others. Instead, I used the tools of construct modeling (Wilson, 2005; Linn and Gronlund, 2000) in order to examine the assessment system and process, from task design to score reporting, with the goal of describing its overall coherence.

The four research questions in this study each represent a relationship between two facets of assessment. Although the "assessment triangle" framework (NRC, 2001) discussed in the previous chapters includes only three major components—cognition, observation, and interpretation—the four different data sources each correspond to part of the triangle. Three of the data sources represent the vertices: the construct map is the model of cognition; the students' written responses are the observation of their understanding; and the scores are the interpretation. It is no coincidence that these three sources of data are also the parts of this assessment design that would continue to exist more or less in the same form if this approach to assessment were implemented on a larger scale. The think-aloud data, in contrast, is unique to this stage in the assessment development process, although it is also a form of observation. A summary of the data sources and analysis methods is shown in Table 2 below.

Table 2: Methods Summary. Data sources and analysis plan for each research question.

Research Questions: ***To what extent …***	**Data Source(s)**	**Analysis**
1: *does the performance assessment task elicit the intended construct in student thinking?*	• Construct Map • Audiotaped Think-alouds of students completing performance task	Frequency comparison of think-aloud codes to item designations of construct and level
2: *do the students' recorded responses to the assessment task correlate with their thinking during the task?*	• Audiotaped think-alouds of students completing performance task • Students' written responses to performance task	Regression analysis between thinkaloud codes and written response codes
3: *does the scoring procedure correlate with the students' recorded responses?*	• Students' written responses to performance task • Assessment scores generated via scoring protocol	Regression analysis between written response codes and assessment scores
4: *do scores represent the range of student thinking?*	• Audio-taped thinkalouds of students completing performance task • Assessment scores generated via scoring protocol	Regresssion analysis between thinkaloud codes and assessment scores

Because each type of data is an indicator for a different component of the assessment process, each of the research questions was addressed with a correlative comparison of two variables. A summary of these pairings is shown in Figure 5 below. To address Research Question 1, I needed to compare the predetermined construct and level designations for each performance assessment task component to the range of student thinking represented by the think-aloud codes. To quantify this range, I used the COUNTIF function in google sheets to tally the number of codes at for each construct and level on every task component. I then summed these values across components to find a total count of instances that each construct was represented in the think-aloud codes, by level, over the entire administration of this performance assessment.

Research Questions: ***To what extent …***	**Construct Map**	**Think-aloud Codes**	**Written Response Codes**	**Scores**
1: *does the performance assessment task elicit the intended construct in student thinking?*	Assigned to assessment task component	Construct-based, then emergent re: content & strategies		
2: *do the students' recorded responses to the assessment task correlate with their thinking during the task?*		Construct-based, then emergent re: content & strategies	Construct-based; then open coding	
3: *does the scoring procedure correlate with the students' recorded responses?*			Construct-based; then open coding	According to scoring protocol
4: *do scores represent the range of student thinking?*		Construct-based, then emergent re: content & strategies		According to scoring protocol

Figure 5: Analysis Pairs and Coding Methods

For Research Questions two, three, and four, I needed to compare pairs of codes and/or scores. Beginning with an assumption of a linear relationship between variables, I used the LINEST function in google sheets to complete a simple regression an analysis and calculate the coefficient of determination and standard error, for each of three different comparison modalities by construct. These three modalities were item-wise comparisons, in which I looked at the relationships between codes and scores on every instance in which they occurred; item average comparisons, in which I looked at the relationships between averages across items; and holistic comparisons, in which I looked at the relationship between holistic values and/or scores for each student. I used the TTEST function to perform a two-tailed t-test on each comparison, setting statistical significance level of $p \leq 0.05$. The results from these pairwise comparisons are described in detail in Chapter 5, illustrated with examples and excerpts of student responses.

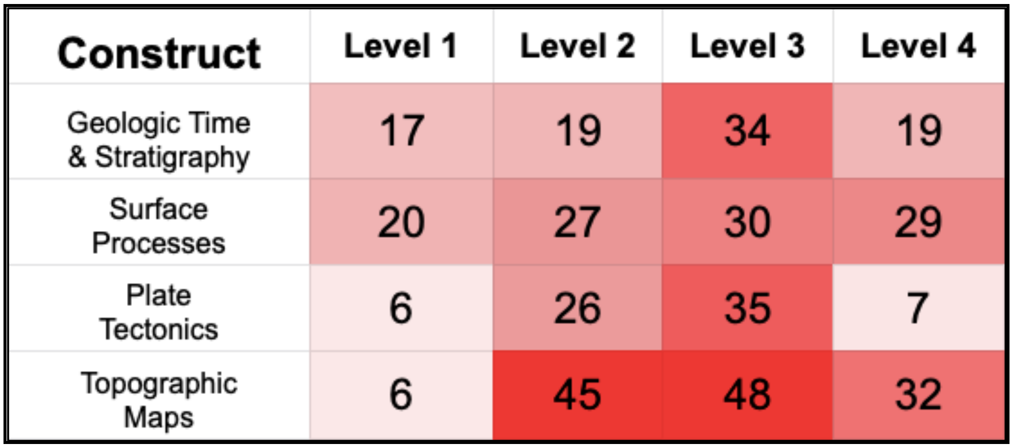
I used the linear regression analysis here not as a way to generalize the results of this assessment beyond the population included in this study, or beyond the bounds of the four-point scale of the construct map. Instead, the linear fit was intended as a way to operationalize coherence. The residuals or outliers can be interpreted as indicators of flaws in the assessment system, or areas to interrogate issues that threaten coherence.

Chapter 5: Findings

This study is attempting to answer a big-picture question: *How do cognitively-based performance assessments promote coherence between students' understanding, responses, and scoring?* I will look at this through the lens of different comparisons between data sources to ascertain an overall view of coherence of the assessment design and process.

5.1 Research Question 1: Eliciting Intended Constructs

To what extent does the performance assessment task elicit the intended construct(s) in student thinking? In order to answer this question about the assessment as a whole, I tallied the codes for each construct and level that emerged from the student think-aloud data for all 22 subjects. Each of these codes indicates an instance when the thinking described aloud by a student when engaged in the assessment task aligned with one or more criteria at a specific construct and level in the construct map. This is therefore a count of how many times the assessment task as a whole elicited the intended constructs in student thinking across all four levels of the construct map. These tallies are shown in Figure 6.

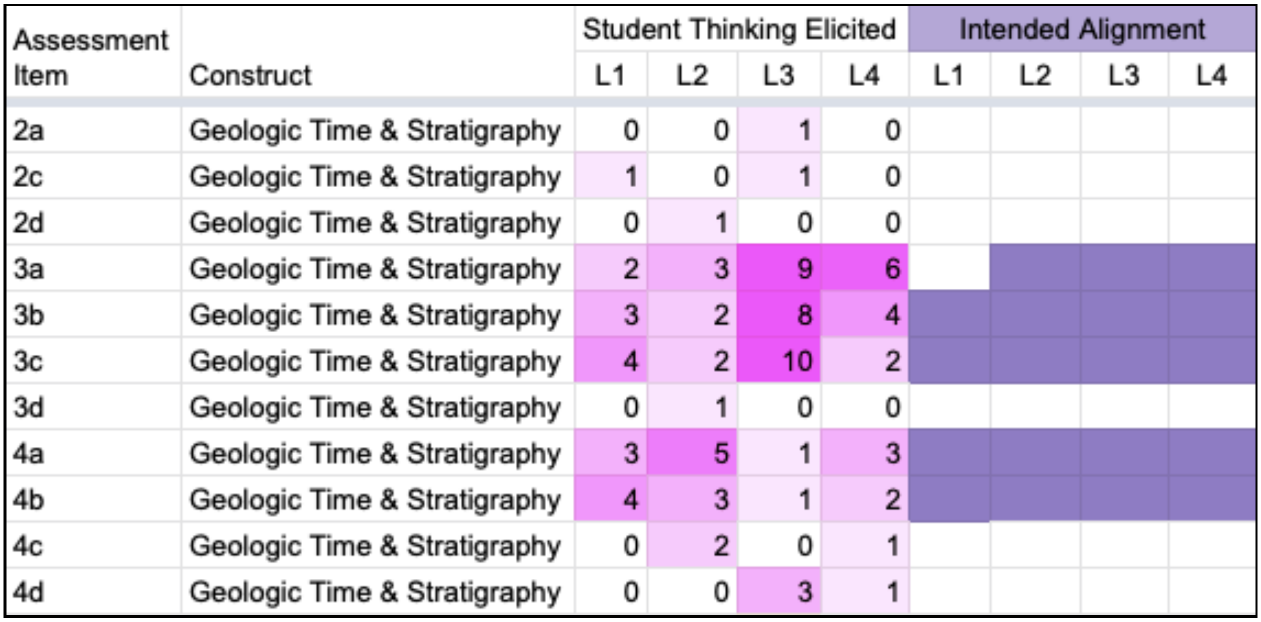


Construct	Level 1	Level 2	Level 3	Level 4
Geologic Time & Stratigraphy	17	19	34	19
Surface Processes	20	27	30	29
Plate Tectonics	6	26	35	7
Topographic Maps	6	45	48	32

Figure 6: Student Thinking Elicited by Construct. Heat map showing the total number of codes for each construct and level in student think-aloud data.

Overall, this performance assessment did elicit student thinking about all four constructs across the range of levels defined by the construct map. The frequency of thinking and distribution across levels varied by construct. The topographic maps construct was coded for 131 times, representing 131 discrete instances of student thinking aligned with this construct from the 22 subjects in my study. This was the most-coded-for construct in all think-aloud data. There were 106 codes for thinking about surface processes, and 89 codes for thinking about geologic time and stratigraphy. The plate tectonics construct had 74 codes, indicating that students had fewer instances of thinking about plate tectonics as they completed the performance assessment tasks than about the other three constructs. It is also notable that codes for the plate tectonics construct were more tightly clustered in levels 1 and 2, with the lowest and highest levels showing up less than 10 times each in student think-aloud data. In contrast, the codes for both the geologic time & stratigraphy and surface processes were more evenly distributed across all four levels. This disparity may be partly explained by true variations in student thinking; e.g., it is likely that the majority of subjects understand plate tectonics in ways that align most closely with the middle levels on the construct map.

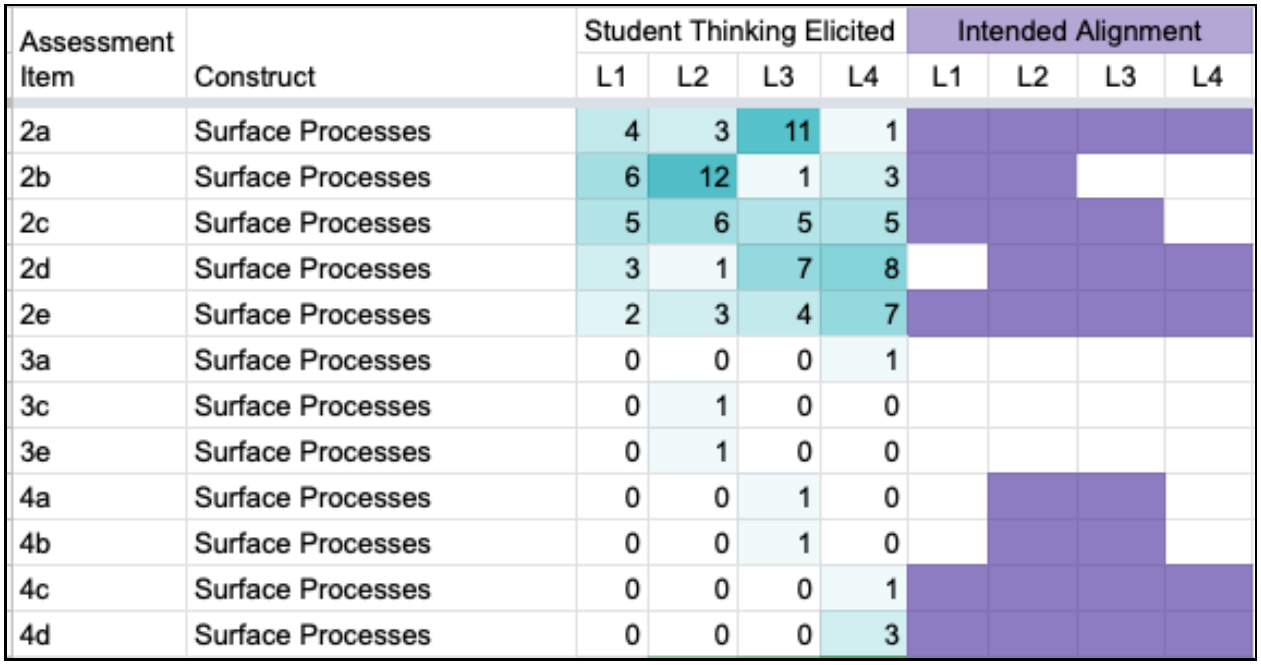
An item-by-item analysis reveals additional patterns. Figures 7, 8, 9, and 10 show the disaggregated number of codes at each level for single assessment items. These numbers are shaded to show frequency, with darker shading representing greater instances of coding for the construct at that level. The "intended alignment" columns on the right show the constructs and levels assigned to each item by the assessment scoring procedure. In other words, these columns show the range of thinking I anticipated each item would elicit.



Assessment Item	Construct	Student Thinking Elicited				Intended Alignment			
		L1	L2	L3	L4	L1	L2	L3	L4
2a	Geologic Time & Stratigraphy	0	0	1	0				
2c	Geologic Time & Stratigraphy	1	0	1	0				
2d	Geologic Time & Stratigraphy	0	1	0	0				
3a	Geologic Time & Stratigraphy	2	3	9	6		▨	▨	▨
3b	Geologic Time & Stratigraphy	3	2	8	4	▨	▨	▨	▨
3c	Geologic Time & Stratigraphy	4	2	10	2	▨	▨	▨	▨
3d	Geologic Time & Stratigraphy	0	1	0	0				
4a	Geologic Time & Stratigraphy	3	5	1	3	▨	▨	▨	▨
4b	Geologic Time & Stratigraphy	4	3	1	2	▨	▨	▨	▨
4c	Geologic Time & Stratigraphy	0	2	0	1				
4d	Geologic Time & Stratigraphy	0	0	3	1				

Figure 7: GTS Thinking Elicited. Item-by-item comparison of think-aloud codes and construct map alignment for the geologic time and stratigraphy construct.

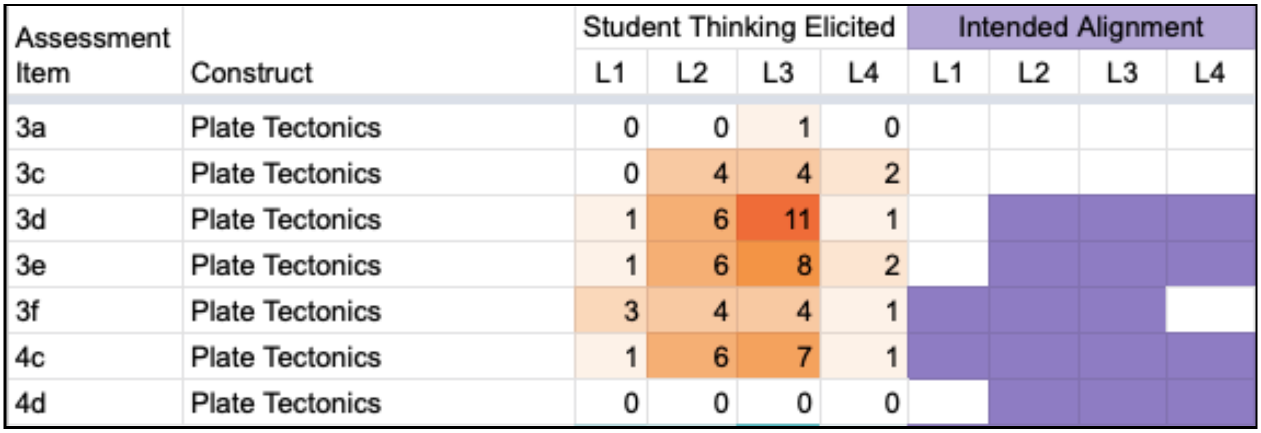
Geologic time and stratigraphy was the construct most likely to come up in student thinking outside of the places where it was intentionally elicited by the performance assessment. However, instances of these codes occurred in small numbers on each of these unintended items. Based on the lack of a pattern, it seems unlikely that there is an underlying assessment design reason for these "extra" instances of student thinking about this construct.



Assessment Item	Construct	Student Thinking Elicited				Intended Alignment			
		L1	L2	L3	L4	L1	L2	L3	L4
2a	Surface Processes	4	3	11	1	■	■	■	■
2b	Surface Processes	6	12	1	3	■	■		
2c	Surface Processes	5	6	5	5	■	■	■	
2d	Surface Processes	3	1	7	8		■	■	■
2e	Surface Processes	2	3	4	7	■	■	■	■
3a	Surface Processes	0	0	0	1				
3c	Surface Processes	0	1	0	0				
3e	Surface Processes	0	1	0	0				
4a	Surface Processes	0	0	1	0		■	■	
4b	Surface Processes	0	0	1	0		■	■	
4c	Surface Processes	0	0	0	1	■	■	■	■
4d	Surface Processes	0	0	0	3	■	■	■	■

Figure 8: SP Thinking Elicited. Item-by-item comparison of think-aloud codes and construct map alignment for the surface processes construct.

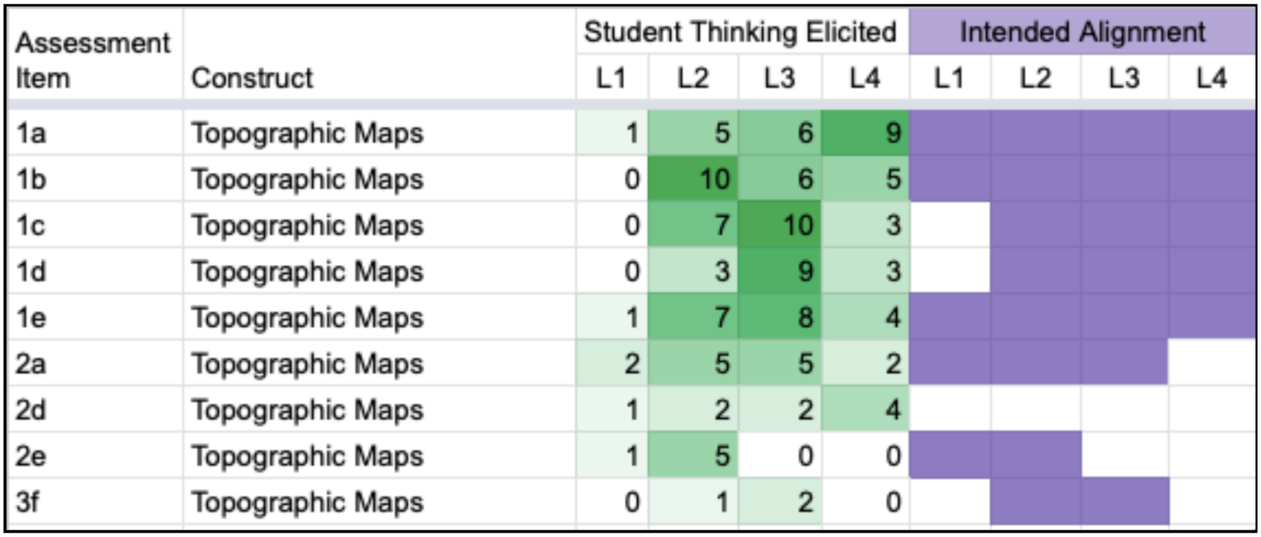
Part 2 of the assessment successfully elicited thinking about the surface processes construct, but across a slightly greater range of levels than anticipated on each item. One major reason for this was when students anticipated the ideas prompted by subsequent items as they thought about an initial question. This speaks to my intention to create a performance assessment that integrated ideas from different constructs and levels across multiple items. It is interesting to note that the distribution pattern across levels is different on items 2a and 2e, both of which I had anticipated would be able to elicit student thinking across the range of ideas on the construct map. This implies that the concentration of codes for thinking about surface processes at level 3 is an effect of the assessment design for item 2a. It may allow for level 4 thinking, but was not as likely to elicit it as items 2d and 2e, for example, which specifically prompted students to make a prediction (one of the possible elements of level 4 thinking on the surface processes construct map). This points to the importance of prompting for higher level thinking with intention, rather than assuming it will happen automatically when students are capable.



Assessment Item	Construct	Student Thinking Elicited				Intended Alignment			
		L1	L2	L3	L4	L1	L2	L3	L4
3a	Plate Tectonics	0	0	1	0				
3c	Plate Tectonics	0	4	4	2				
3d	Plate Tectonics	1	6	11	1		🟪	🟪	🟪
3e	Plate Tectonics	1	6	8	2		🟪	🟪	🟪
3f	Plate Tectonics	3	4	4	1	🟪	🟪	🟪	
4c	Plate Tectonics	1	6	7	1	🟪	🟪	🟪	🟪
4d	Plate Tectonics	0	0	0	0		🟪	🟪	🟪

Figure 9: PT Thinking Elicited. Item-by-item comparison of think-aloud codes and construct map alignment for the plate tectonics construct.

This comparison reveals an assessment design flaw. The plate tectonics construct was already underrepresented compared to the other three constructs, but one of the items (4d) intended to measure it did not prompt any student thinking about plate tectonics ideas. This item, which asked students to predict the future evolution of the landscape, was very open-ended and did not specify a timescale or mechanism for changes.



Assessment Item	Construct	Student Thinking Elicited				Intended Alignment			
		L1	L2	L3	L4	L1	L2	L3	L4
1a	Topographic Maps	1	5	6	9	🟪	🟪	🟪	🟪
1b	Topographic Maps	0	10	6	5	🟪	🟪	🟪	🟪
1c	Topographic Maps	0	7	10	3		🟪	🟪	🟪
1d	Topographic Maps	0	3	9	3		🟪	🟪	🟪
1e	Topographic Maps	1	7	8	4	🟪	🟪	🟪	🟪
2a	Topographic Maps	2	5	5	2	🟪	🟪	🟪	
2d	Topographic Maps	1	2	2	4				
2e	Topographic Maps	1	5	0	0	🟪	🟪		
3f	Topographic Maps	0	1	2	0		🟪	🟪	

Figure 10: TM Thinking Elicited. Item-by-item comparison of think-aloud codes and construct map alignment for the topographic maps construct.

The performance assessment tasks elicited student thinking in strong agreement with the intended alignment to the topographic maps construct. There was a single item (2d) where students' self-reported thinking included ideas about topographic maps that I did not anticipate in the assessment design phase. In this part of the task, students used supposition to sketch a landscape profile in the absence of running water. Some used quantitative features of the existing map to guide their decisions about how to draw the new profile. Others thought about what the contour lines would look like on a topographic map of this reimagined landscape.

5.1.a Summary

Across the entire performance assessment, there were instances where student thinking went beyond the bounds of what the items were intended to elicit, as well as items that failed to elicit much student thinking in the intended construct. The assessment as a whole succeeded in eliciting thinking about all four levels of each construct defined by the construct map. In part 4 of the assessment, student thinking was concentrated on the geologic time & stratigraphy construct and did not include as many connections to other constructs as I had hoped these "big picture" items would elicit.

5.2 Descriptive Data: Results of Performance Assessment and Coding

The results of this assessment process are in the form of a holistic score for each construct, per student. In the following bar charts, the distribution of holistic scores is juxtaposed with holistic values for think-aloud and written codes. These descriptive data provide overall context for the assessment outcomes. It is crucial to note that these charts do not provide information about how the observation of individual students changed from think-aloud to written responses to scores. The total number of students may be different (ranging from 19 to 22) depending on the construct and data source. For these reasons, this presentation of the data, shown in Figures 11, 12, 13, and 14 on the following pages, is most useful for identifying overall trends and patterns.

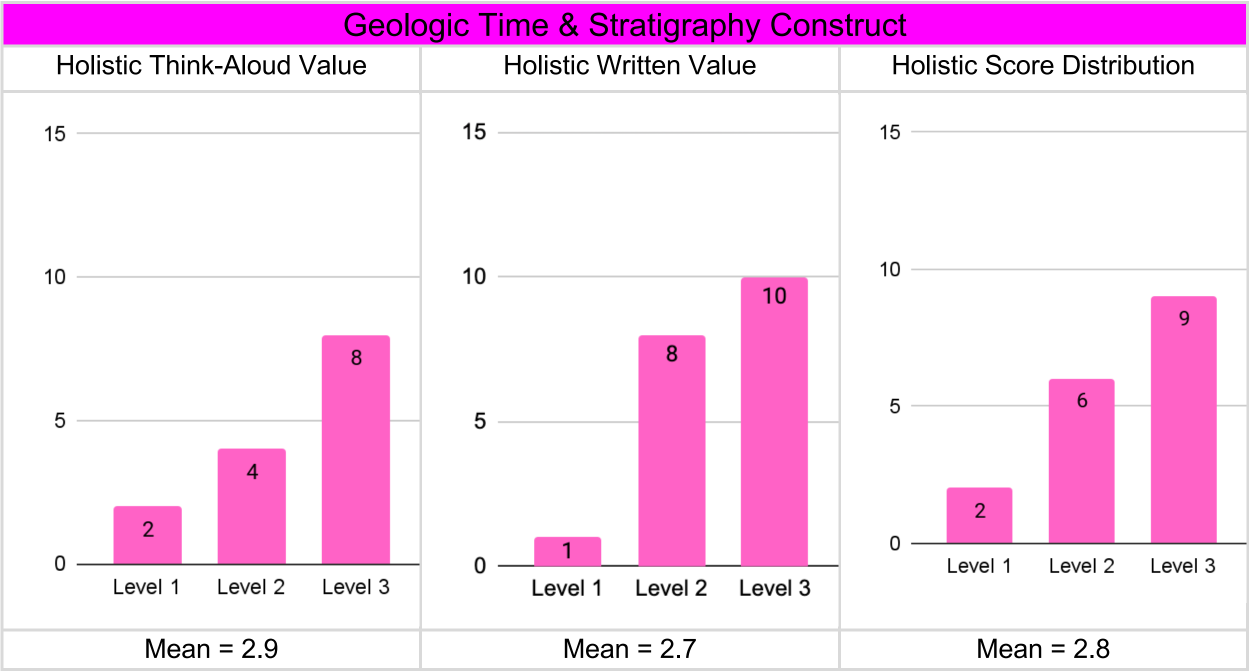


Figure 11: GTS Descriptive Data.

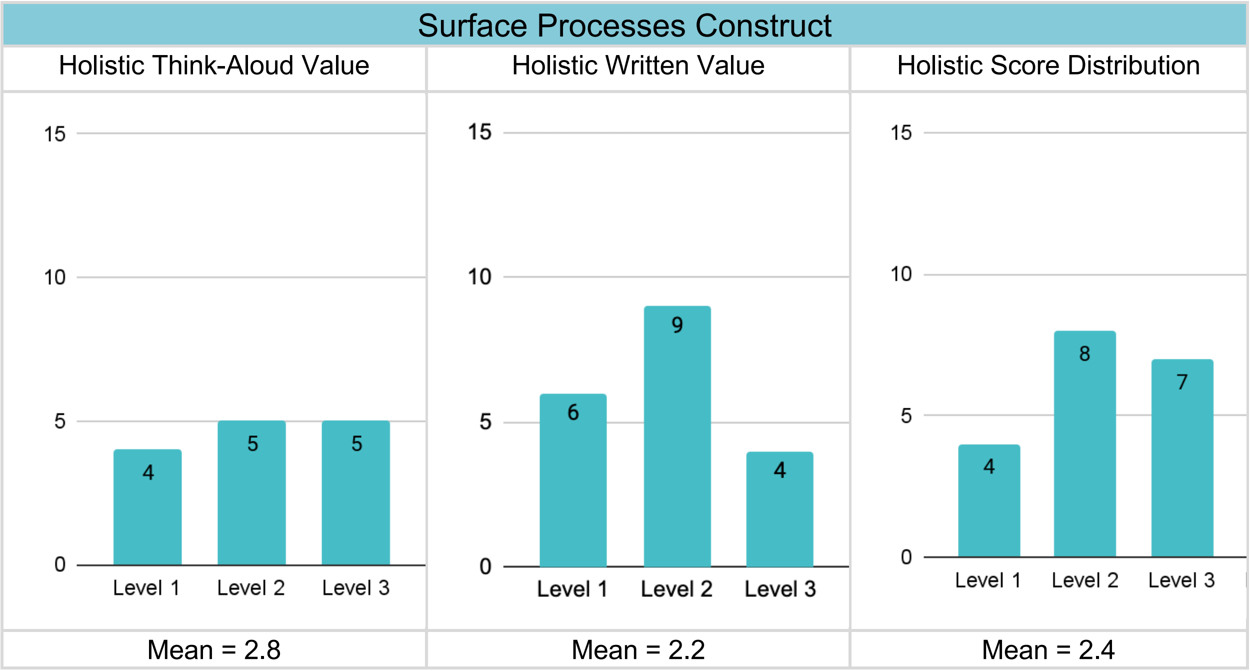


Figure 12: SP Descriptive Data.

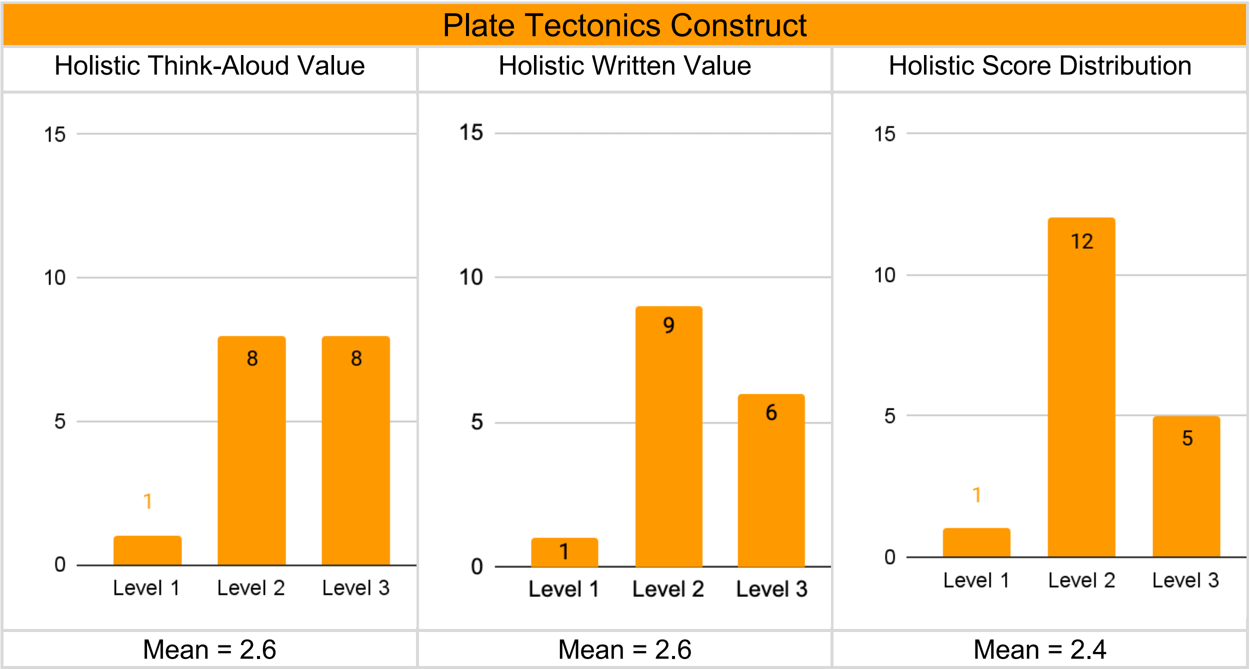


Figure 13: PT Descriptive Data.

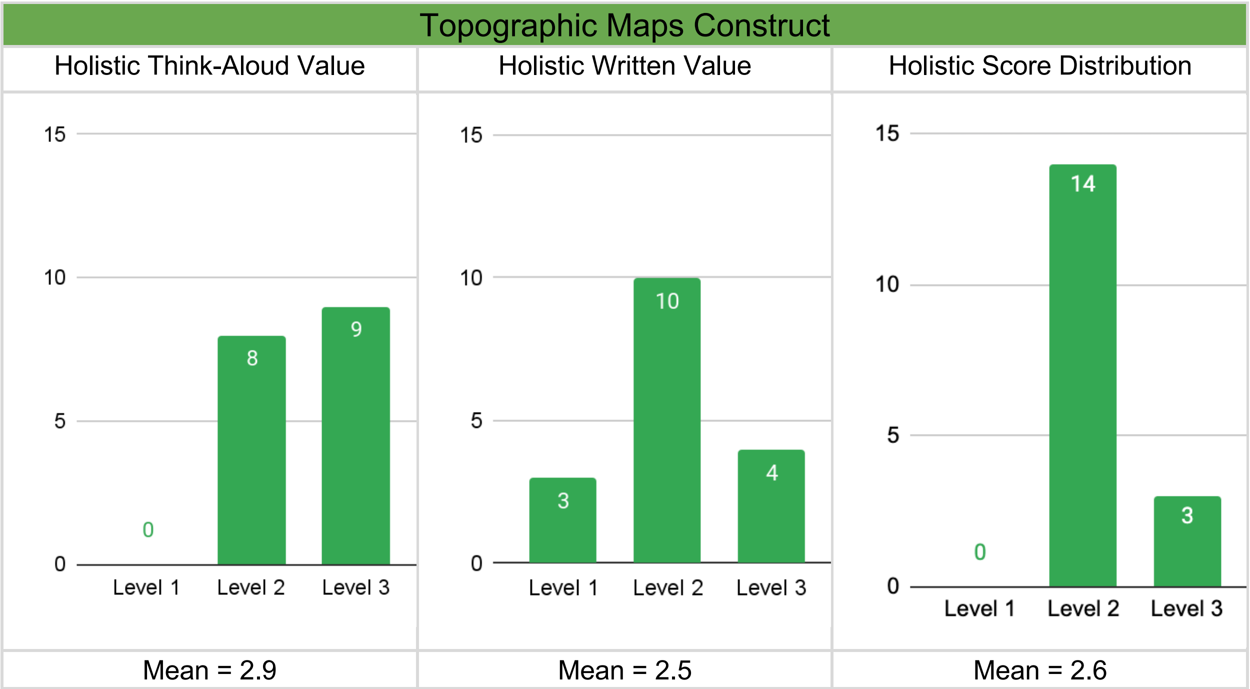


Figure 14: TM Descriptive Data.

In general, these graphs show a roughly bell-shaped distribution of student thinking, responses, and scores across the four levels. With one exception, in the think-aloud data for the surface processes construct, student thinking and performance was clustered towards the center of the construct map continuum. Codes and scores were more frequent at levels 2 and 3 than they were at levels 1 and 4.

For every single construct, the mean holistic think-aloud code is higher than the mean holistic score. These disparities are small – none of the differences indicate a reduction to an average lower level. However, they indicate that across constructs, there is some loss of information about higher level thinking in the holistic score output.

A second trend apparent in the results is the tendency for level 2 to be overrepresented in the holistic scores in comparison to the original distribution of think-aloud levels. This is especially apparent in the topographic maps construct, but also appears to have occurred in the surface processes and plate tectonics constructs.

5.3 Findings from Research Questions 2, 3, and 4

In this section, I will investigate the overall coherence of this cognitively-based performance assessment on a construct-by-construct basis. Within each construct, I will present results of each research question, via pairwise comparisons between data sources. The goal of this analysis is to illustrate the quality of the match between an "original" data source and one that is the result of further reduction or interpretation. Each comparison is done in two ways: first, via linear regression analysis to characterize how closely the data fits to an ideal scenario; second, with a descriptive analysis of the ways in which the assessment tools captured or missed information in the original data.

Note that the p-values on these comparisons, as determined by two-tailed t-tests, do not provide predictive or inferential power about how these particular assessment items would work with any population of students. Rather, they serve as a descriptive tool to identify strengths and weaknesses in the assessment system as used with the student population in this study. The t-test determines whether there is a difference in the means of compared data sources within the same construct; and the p-value describes the probability that these differences are due to chance.

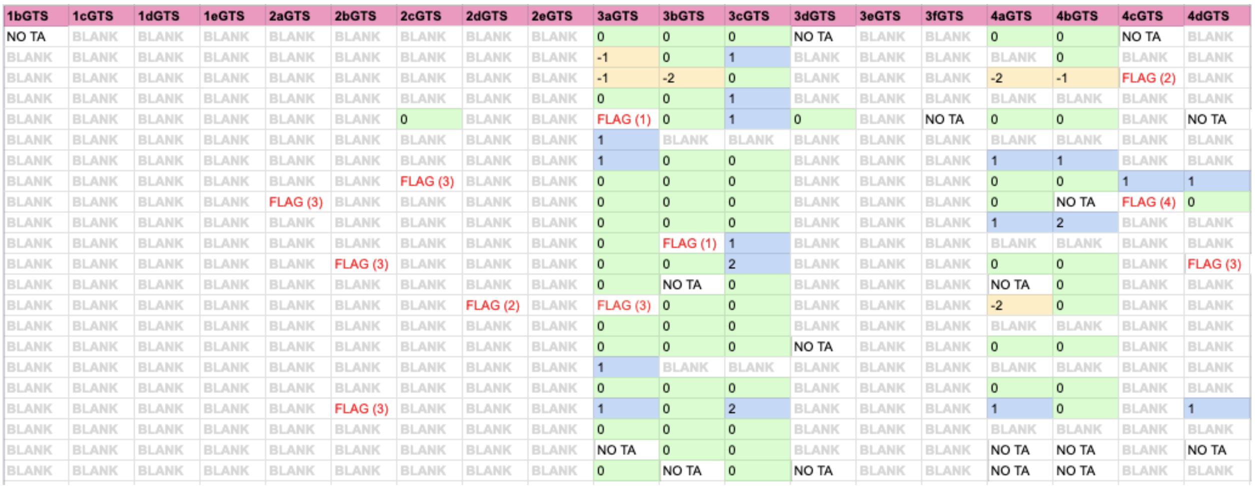
5.3.a Geologic Time & Stratigraphy Construct

For all three research questions, the holistic values provided a better match than data pairs from student responses to individual items. A summary of the coherence data for the Geologic Time and Stratigraphy construct is shown in Table 3, below.

Table 3: GTS Coherence Data.

Pair	Comparison Method	R^2	Standard error	P value	Perfect Match	Flags
WC vs TA (RQ2)	Individual Responses (87)	0.61	0.647	0.045	66%	11
	Holistic Values (20)	0.74	0.454	0.021	80%	-
Score vs WC (RQ3)	Individual Responses (90)	0.89	0.34	0.034	73%	10
	Holistic Values (22)	0.84	0.31	0.58	86%	-
Score vs TA (RQ4)	Individual Responses (82)	0.61	0.65	0.18	78%	16
	Holistic Values (20)	0.81	0.38	0.042	80%	-

5.3.a.1 GTS RQ2 Comparison. This comparison pair describes the alignment between think-aloud codes and codes on students' written responses for the geologic time & stratigraphy construct. There were 87 data pairs with codes for both the think-aloud and the written response for this construct across the assessment for all subjects. Of those 87 pairs, 57 had identical values for both think-aloud and written codes, meaning that the students' recorded responses aligned with the same construct level as their thinking 66% of the time. When the data were collapsed into holistic values, the agreement improved. For the 20 students who had holistic values for think-aloud codes, 16 of them (80%) had matching holistic written codes on this construct.



1bGTS	1cGTS	1dGTS	1eGTS	2aGTS	2bGTS	2cGTS	2dGTS	2eGTS	3aGTS	3bGTS	3cGTS	3dGTS	3eGTS	3fGTS	4aGTS	4bGTS	4cGTS	4dGTS
NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	NO TA	BLANK	BLANK	0	0	NO TA	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	-1	0	1	BLANK	BLANK	BLANK	BLANK	0	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	-1	-2	0	BLANK	BLANK	BLANK	-2	-1	FLAG (2)	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	1	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	BLANK	BLANK	FLAG (1)	0	1	0	BLANK	NO TA	0	0	BLANK	NO TA
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	1	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	1	0	0	BLANK	BLANK	BLANK	1	1	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (3)	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	0	0	1	1
BLANK	BLANK	BLANK	BLANK	FLAG (3)	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	0	NO TA	FLAG (4)	0
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	1	2	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	FLAG (1)	1	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (3)	BLANK	BLANK	BLANK	0	0	2	BLANK	BLANK	BLANK	0	0	BLANK	FLAG (3)
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	NO TA	0	BLANK	BLANK	BLANK	NO TA	0	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (2)	BLANK	FLAG (3)	0	0	BLANK	BLANK	BLANK	-2	0	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	NO TA	BLANK	BLANK	0	0	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	1	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	0	0	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (3)	BLANK	BLANK	BLANK	1	0	2	BLANK	BLANK	BLANK	1	0	BLANK	1
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO TA	0	0	BLANK	BLANK	BLANK	NO TA	NO TA	BLANK	NO TA
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	NO TA	0	NO TA	BLANK	BLANK	NO TA	NO TA	BLANK	BLANK

Figure 15: Difference Between Think-Aloud Code and Written Code for the GTS Construct.

Figure 15 shows the "location" of these data about the geologic time & stratigraphy construct within the assessment items. Each cell represents an individual student's responses to a single item. The value within the cell is calculated as the difference between the think-aloud code and written code. Therefore, a "0" means there was a perfect match between think-aloud and written codes. When the result is a positive number, it means the think-aloud code is higher than the written code.

There was no single item that stood out as having a significantly different profile from the others. Item 3c tended to elicit thinking that was coded at a higher level than the written responses students produced, while other items had a mix of higher or lower values. It appears that item 3c required more reasoning and problem solving, and that students were not given sufficient structure and opportunities to record this thinking. Most of the mismatches for this construct are a one-level difference, but six are off by two levels (this represents 20% of the disagreements and 6.9% of all student responses to the geologic time & stratigraphy construct). This is a large mismatch on a four-point scale. This is another advantage of the holistic comparison for this construct: none of the four mismatches at this level diverge from one another by more than a single level.

In addition to the "mismatch" situations, there are some instances where the geologic time & stratigraphy construct was coded in a student's think-alouds but not in their written response for the same item. These instances represent missed opportunities for capturing student thinking about the construct. They are indicated by the "FLAG" cells in the spreadsheet figure above. For this construct, this occurred once for level 4 student thinking, six times for level 3 student thinking, and twice for each of levels 2 and 1. Overall there were eleven missed opportunities for capturing student thinking about geologic time and stratigraphy across all 22 subjects. On average, it happened less than once per student. The spreadsheet reveals that most of these instances occurred on items that were not targeting the construct.

On the linear regressions, R^2 ranged from 0.61 to 0.74, meaning that the variation in student thinking accounted for up to 74% of the variation in students' recorded responses, depending on the comparison method.

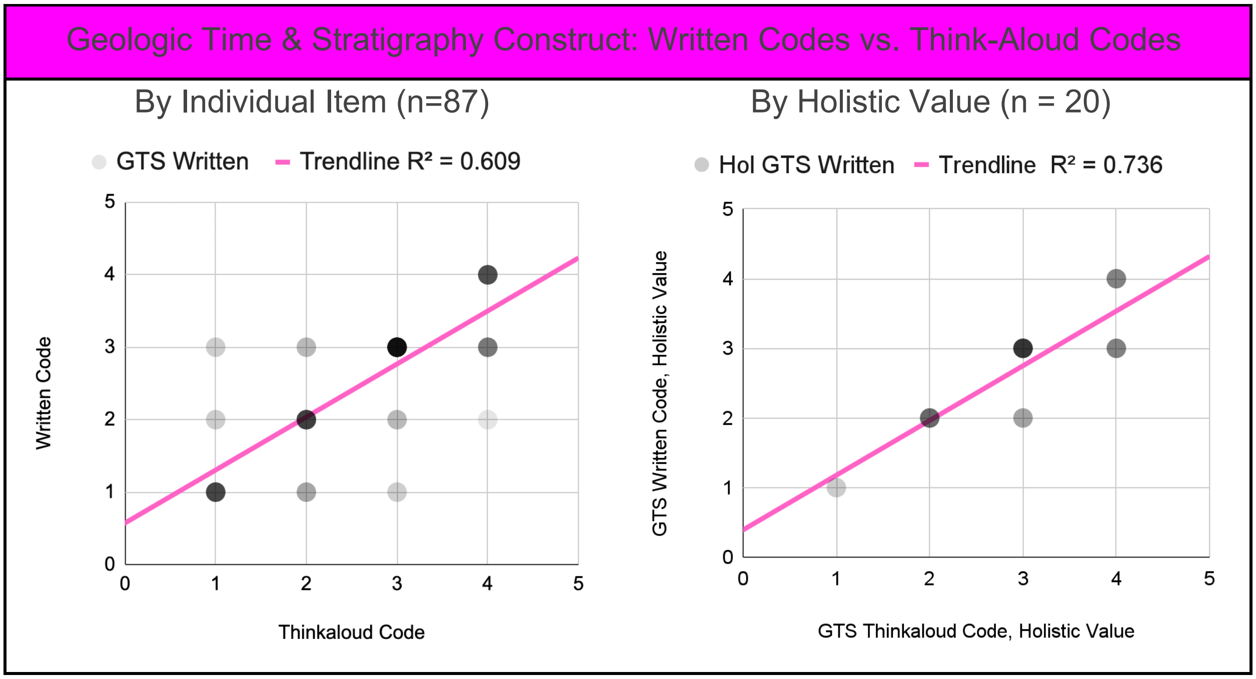


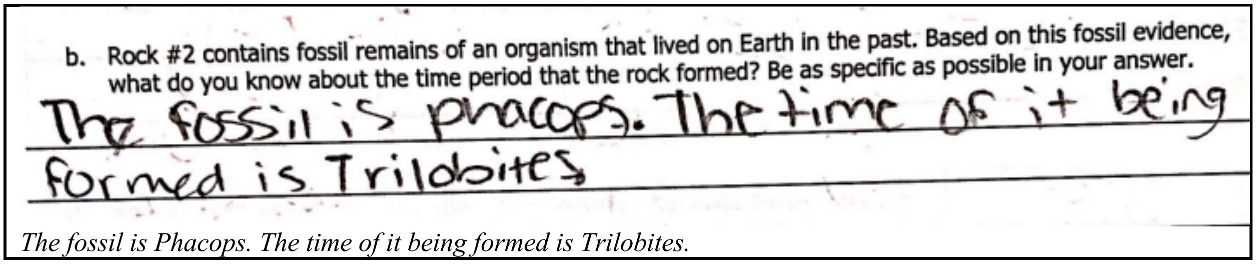
Figure 16: Linear Regressions on Individual Responses and Holistic Values Comparing GTS Think-Aloud Codes to Written Codes.

For the geologic time & stratigraphy construct, the holistic values were more closely correlated than responses to individual items, which had a relatively high standard error of 0.656. This suggests that there were assessment design flaws on specific items that lead to a mismatch between student thinking and the response ultimately recorded by the student in writing. It also suggests that the mismatch on individual items was less consistent than any mismatch that carried over to holistic codes. This is evident when looking at the scatter plots of each comparison method, shown in Figure 16. When there was a discrepancy between holistic think-aloud codes and holistic written codes, the think-aloud codes were generally coded at a higher construct level.

The following example illustrates one instance of this mismatch. This part of the assessment (3b) asked students to make inferences about a past geologic period based on fossil evidence in a physical rock sample. In this example, Christiana was thinking about the time period when a specific rock would have formed.

> "Oh I already did this one. I figured out the type of organism that it was. Now I'm going back to the geologic history chart. That's the fossil [pointing to diagram C]. I see it on this outcrop map too [pointing to outcrop D]. I think this means all these rocks with this symbols [pointing to siltstone map symbol] are from that same fossil time.
>
> That lived in old times. Like when the dinosaurs lived?
>
> Where is C at. [pointing to vertical bar in ESRT life on earth column.] That says trilobite. So that's the time? Or… hmm. That's the name. [Sliding finger to the left until it intersects with the Devonian row in the geologic time period column.] That's the time the fossil would be at. So that's the age for these rocks. This is a long time ago."

Christiana's actions and verbalization of her thinking while engaging with this part of the performance task reveals her understanding that rock layers in different outcrops can be correlated, and that index fossil species can be used to determine the age of the correlated rock layers. This understanding corresponds to level 3 in the construct map, and her think-aloud was therefore coded as a 3. However, her written response (shown in Figure 17) did not capture this understanding.



b. Rock #2 contains fossil remains of an organism that lived on Earth in the past. Based on this fossil evidence, what do you know about the time period that the rock formed? Be as specific as possible in your answer.

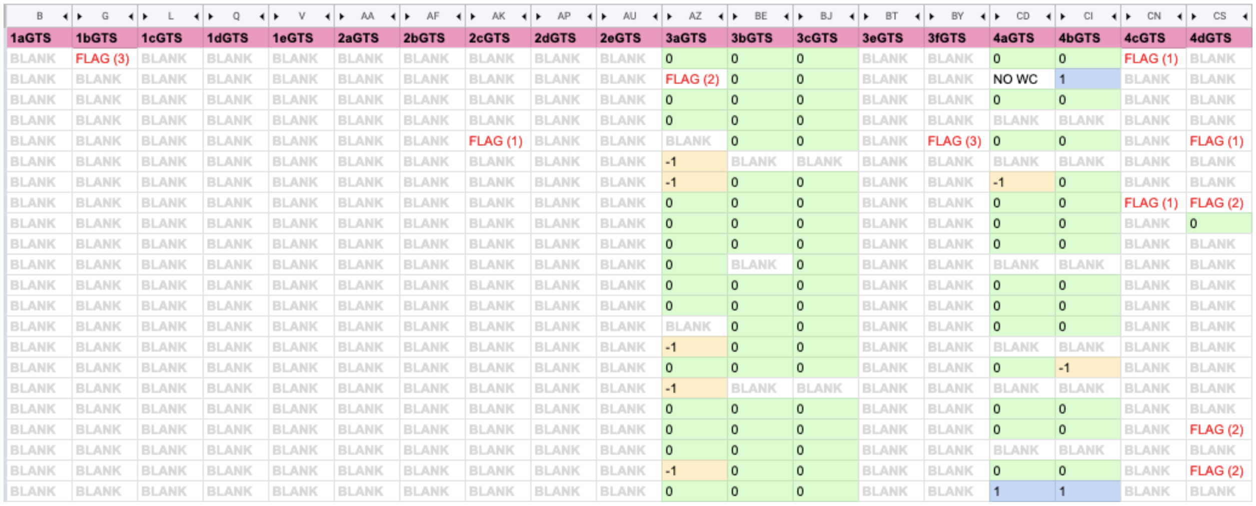
The fossil is phacops. The time of it being formed is Trildbites

The fossil is Phacops. The time of it being formed is Trilobites.

Figure 17: Christiana's Written Response, Item 3b.

Because she correctly identified a material in the rock and wrote that it formed at the time of trilobites, this student's written response to this item was coded as a level 2, lower than her think-aloud code of level 3.

5.3.a.2 GTS RQ3 Comparison. This comparison pair describes the alignment between codes on student written responses and scores that resulted from the scoring procedure. Overall, this construct had good coherence at the level of individual items. There were 90 data pairs with values for both the written response and the scoring procedure for the geologic time & stratigraphy construct across the assessment for all subjects. Of those 90 pairs, 66 had identical values, meaning the students' scores reflected the coded level of their written responses 73% of the time. Similar to the previous comparison, the holistic values had improved agreement over individual responses. Out of all 22 students, 19 of them (86%) had holistic scores on this construct that matched the holistic value for their written response codes.



B	G	L	Q	V	AA	AF	AK	AP	AU	AZ	BE	BJ	BT	BY	CD	CI	CN	CS
1aGTS	1bGTS	1cGTS	1dGTS	1eGTS	2aGTS	2bGTS	2cGTS	2dGTS	2eGTS	3aGTS	3bGTS	3cGTS	3eGTS	3fGTS	4aGTS	4bGTS	4cGTS	4dGTS
BLANK	FLAG (3)	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	0	0	FLAG (1)	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (2)	0	0	BLANK	BLANK	NO WC	1	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	0	0	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (1)	BLANK	BLANK	BLANK	0	0	BLANK	FLAG (3)	0	0	BLANK	FLAG (1)
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	-1	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	-1	0	0	BLANK	BLANK	-1	0	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	0	0	FLAG (1)	FLAG (2)
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	0	0	BLANK	0
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	0	0	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	BLANK	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	0	0	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	0	0	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	BLANK	BLANK	0	0	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	-1	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	0	-1	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	-1	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	0	0	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	0	0	BLANK	FLAG (2)
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	-1	0	0	BLANK	BLANK	0	0	BLANK	FLAG (2)
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	1	1	BLANK	BLANK

Figure 18: Difference Between Written Code and Score for the GTS Construct.

The comparison of written codes to scores by item shown in Figure 18 shows that two items, 3b and 3c, had perfect agreement between these two data sources for all 22 students. All the mismatches, which compose 27% of the total comparisons for this construct, are a one-level difference. The remaining items targeting this construct show a slight tendency for the score to be higher than the written code, resulting in a negative difference between the two. This is mainly apparent on item 3a. On this item, it was somewhat common for students to be "overscored" to level 4 compared to their written code of level 3. These differences all occurred at this level combination. On item 3a, students were classifying rock samples and inferring the formation environment of each one. This pattern points to the need to revisit the relationship between construct map, item design, and outcome space at the higher levels, since there were several instances where the "overscore" ended up matching the think-aloud code.

There were 10 flags representing missed opportunities for scoring the geologic time and stratigraphy construct. These are situations where the student's written work showed evidence of thinking about the construct, but did not receive a score. These are not the same flags as those from the previous comparison between think-aloud and written codes. One of these flags occurred on an item that was intended to be scored for the GTS construct, indicating something that should be added to the outcome space in order to facilitate more complete scoring of this construct. The other flags were on items that were not targeting the GTS construct, and therefore did not allow for the scoring of that construct. Of the 10 flags, four were on responses coded at level 1, four were on responses coded at level 2, three were on responses coded at level 3, and zero were on responses coded at level 4. These are small numbers, but it does appear that the missed scoring opportunities were weighted towards the more novice end of the progression embodied by the construct map.

For holistic values, out of 22 comparisons, 19 were a perfect match. The scatter plot shown below in Figure 19 suggests that two of the three mismatch situations arose from the level 3/4 discrepancy on item 3a, since they are students who received a holistic score of 4 but a holistic written code value of 3. The p-value for this comparison is high, because the third mismatch is a case where the holistic score is lower than the holistic written code value.

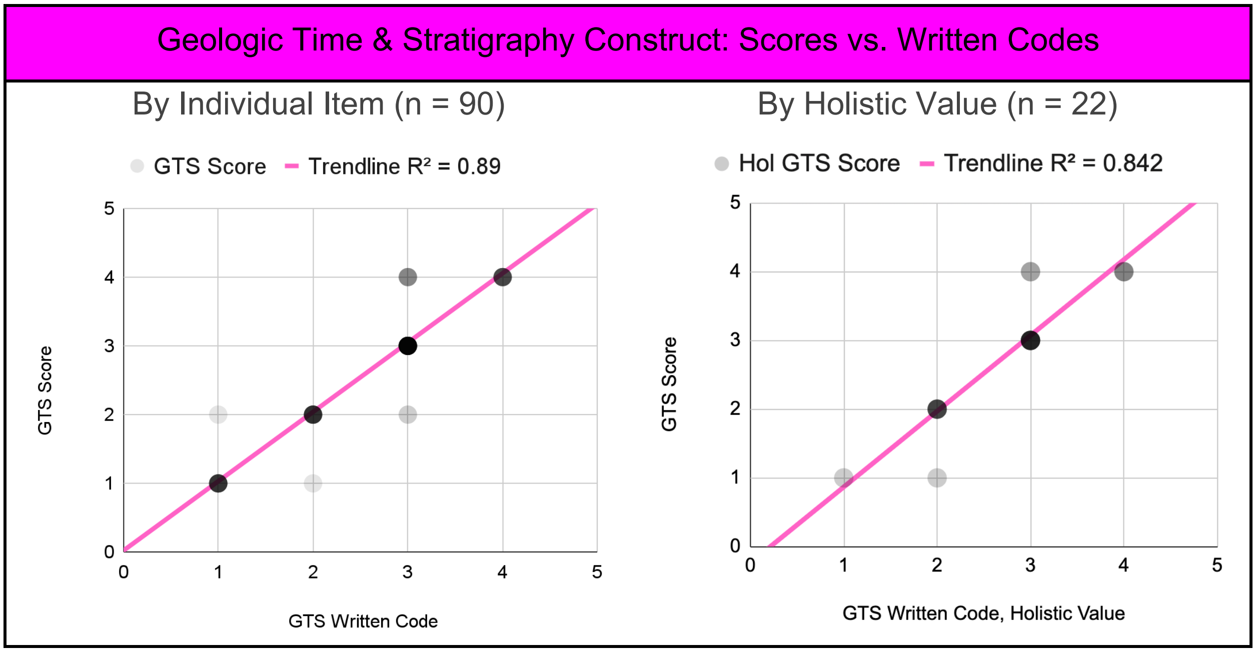
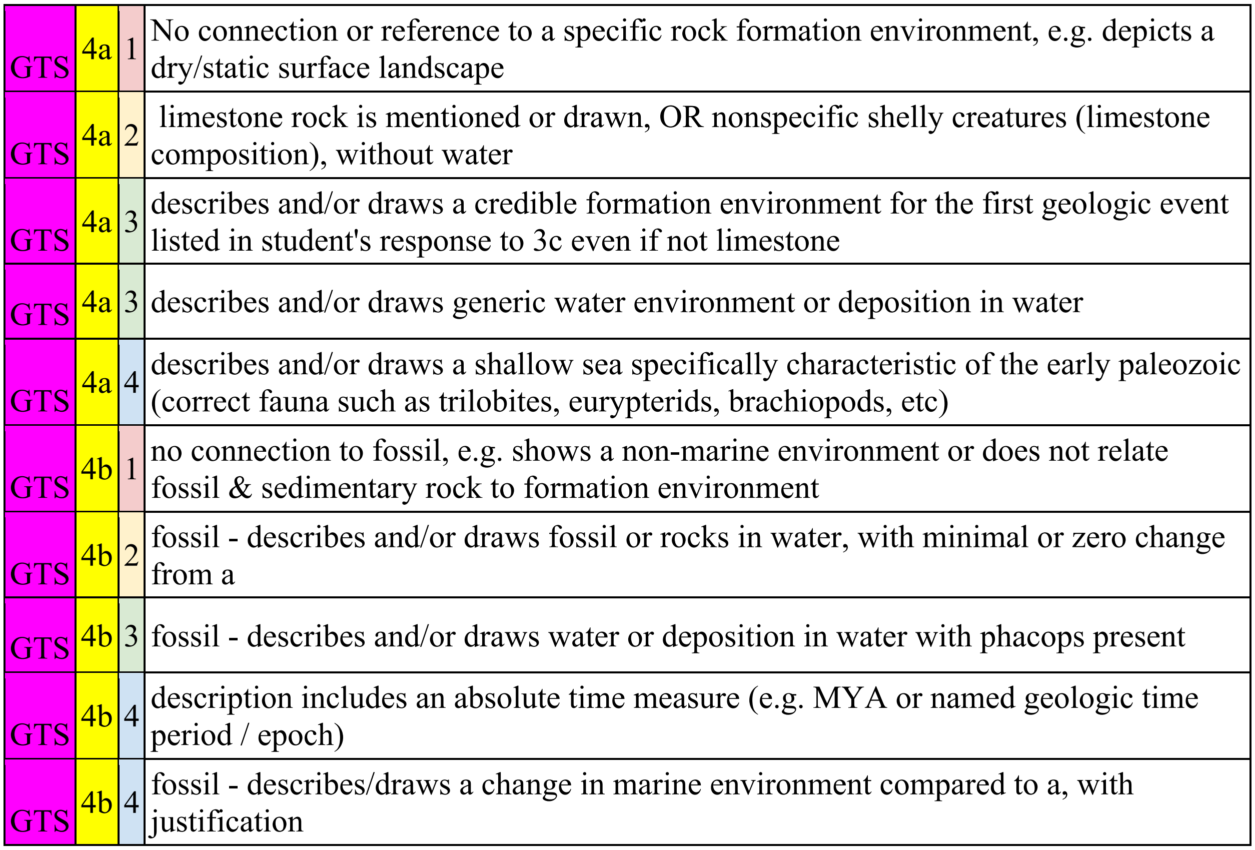


Figure 19: Linear Regressions on Individual Responses and Holistic Values Comparing GTS Think-Aloud Codes to Written Codes.

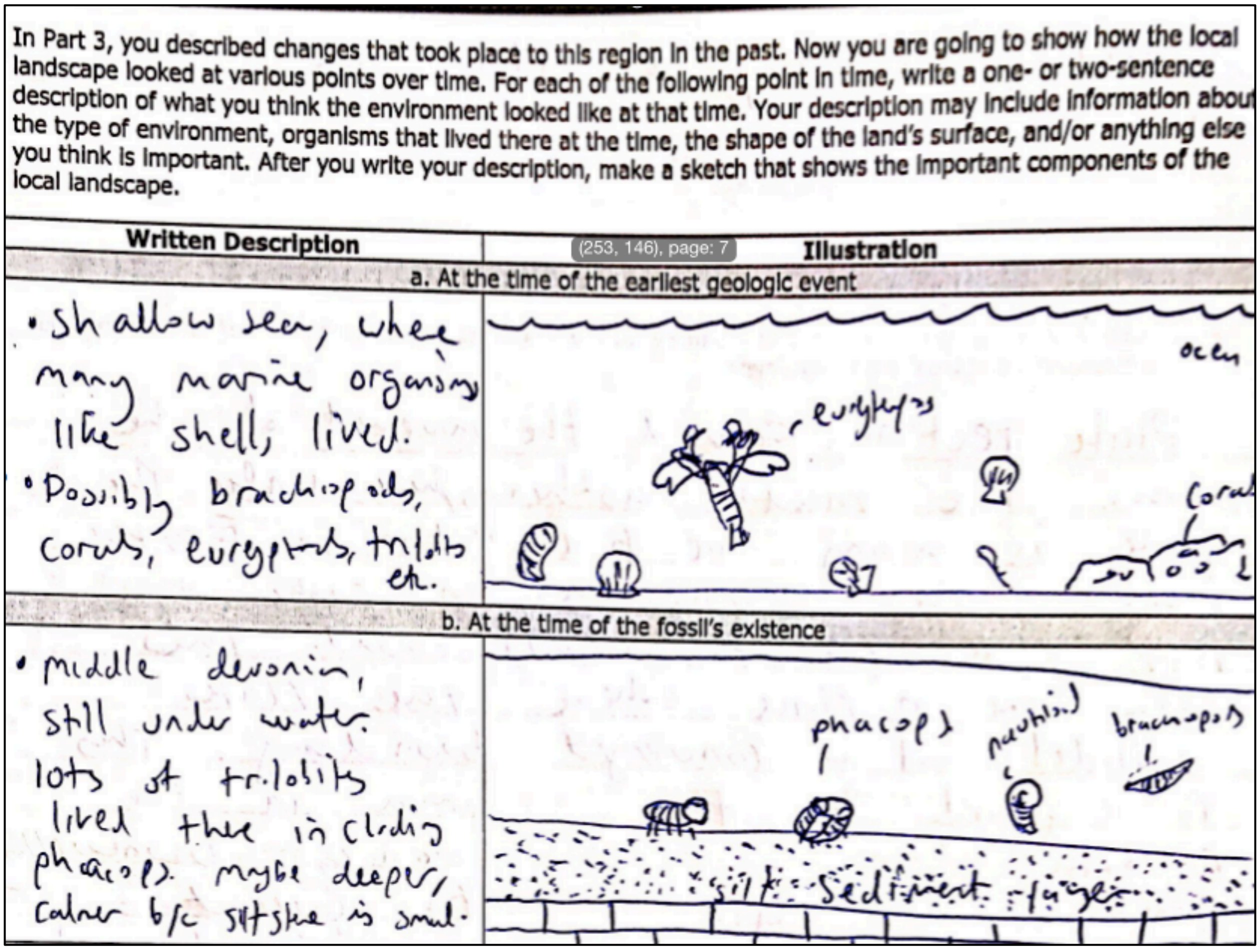
For the geologic time and stratigraphy construct regression analysis, the only significant correlation between written codes and scores was across responses to individual items. This correlation was strong, with an R^2 of 0.89 and a standard error of 0.34 ($p = 0.034$). The scatter plot shown in Figure 19 shows that the distribution of scores was very consistent across different levels, and the score assigned to an item never differed from the written code for that item by more than one level. The vast majority of scores were an exact match to the code; the outliers on this scatterplot are far less frequent than the points that fall on the line, as shown by the closeness of the trendline to the points where the two numbers are the same. This scatterplot is quite close to the ideal.

This demonstrates that, on the level of individual items, the predefined outcome space for responses to individual items led graders to assign scores that closely matched the more general criteria on the construct map. This is an example of strong coherence at an item level. The following examples show two instances of student written responses that correspond well to the outcome space (shown in Figure 20) for these items, resulting in the same score and written code.



GTS	4a	1	No connection or reference to a specific rock formation environment, e.g. depicts a dry/static surface landscape
GTS	4a	2	limestone rock is mentioned or drawn, OR nonspecific shelly creatures (limestone composition), without water
GTS	4a	3	describes and/or draws a credible formation environment for the first geologic event listed in student's response to 3c even if not limestone
GTS	4a	3	describes and/or draws generic water environment or deposition in water
GTS	4a	4	describes and/or draws a shallow sea specifically characteristic of the early paleozoic (correct fauna such as trilobites, eurypterids, brachiopods, etc)
GTS	4b	1	no connection to fossil, e.g. shows a non-marine environment or does not relate fossil & sedimentary rock to formation environment
GTS	4b	2	fossil - describes and/or draws fossil or rocks in water, with minimal or zero change from a
GTS	4b	3	fossil - describes and/or draws water or deposition in water with phacops present
GTS	4b	4	description includes an absolute time measure (e.g. MYA or named geologic time period / epoch)
GTS	4b	4	fossil - describes/draws a change in marine environment compared to a, with justification

Figure 20: GTS Outcome Space for Assessment Items 4a and 4b.

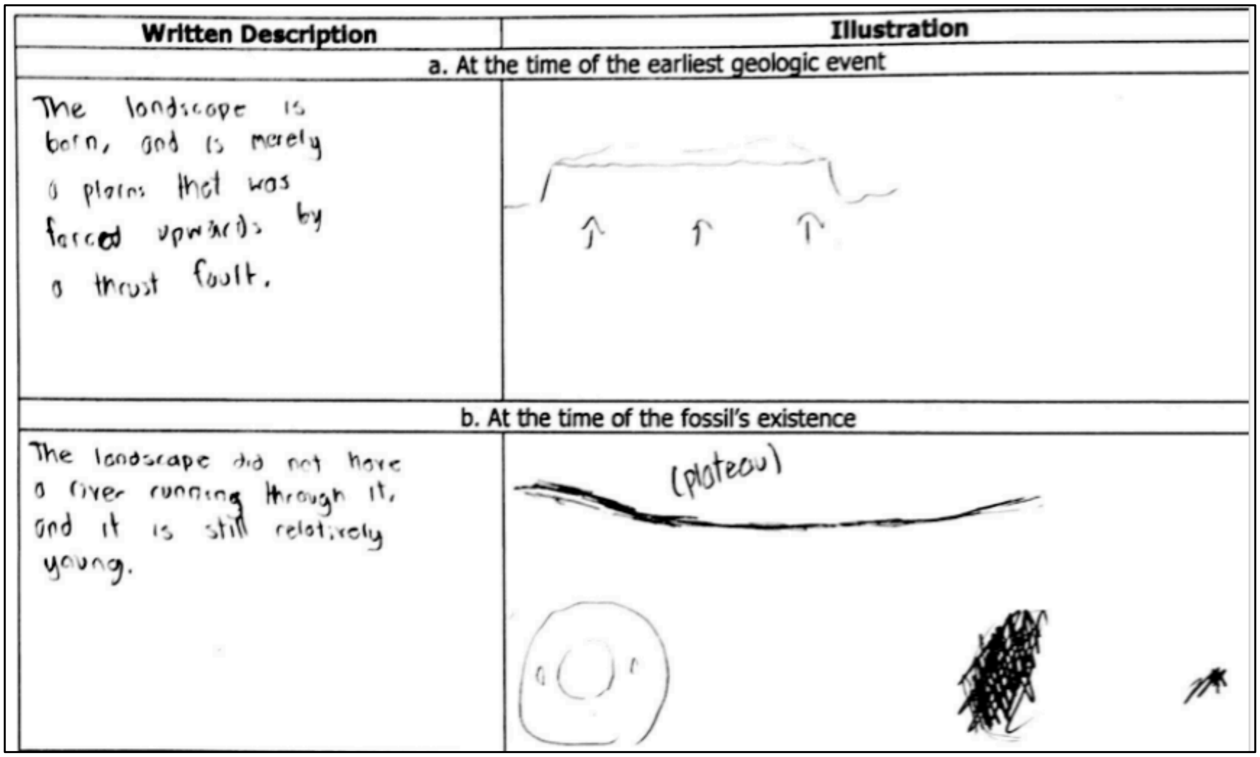


In Part 3, you described changes that took place to this region in the past. Now you are going to show how the local landscape looked at various points over time. For each of the following point in time, write a one- or two-sentence description of what you think the environment looked like at that time. Your description may include information about the type of environment, organisms that lived there at the time, the shape of the land's surface, and/or anything else you think is important. After you write your description, make a sketch that shows the important components of the local landscape.

Written Description	Illustration
a. At the time of the earliest geologic event	
• shallow sea, whee many marine organisms like shell, lived. • Possibly brachiopods, corals, eurypterds, trilobits etc.	
b. At the time of the fossil's existence	
• middle devonin, still under water. lots of trilobits lived there including phacops. maybe deeper, calmer b/c siltstone is small.	

Figure 21: Juju's Level 4 GTS Written Response.

Juju's student's responses to part 4 of the performance assessment matched the description provided in the outcome space for level 4 on the geologic time and stratigraphy construct. In part 4a, they drew and named a "shallow sea" environment that includes a Eurypterus, which lived during the Silurian time period in the early Paleozoic era. In part 4b, they used evidence from a fossiliferous rock sample and outcrop diagram. This is very similar to the descriptions of a level 4 responses in the outcome space, which are exemplars for this criterion from the construct map defining student thinking at this level: *"Can visualize and describe a landscape in the environmental condition that it was in during the formation of rock layers."*



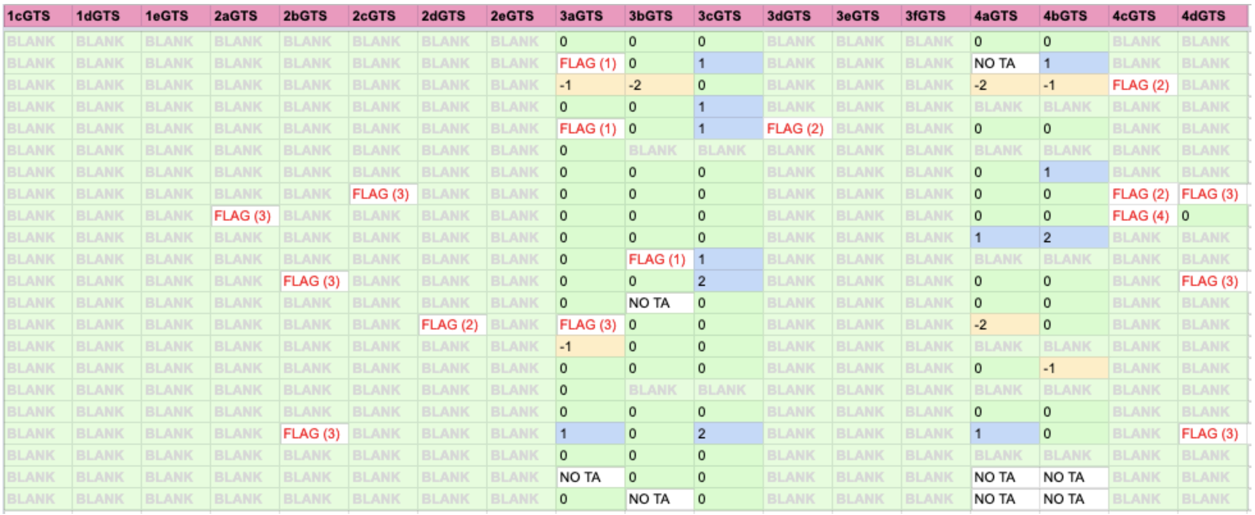
Written Description	Illustration
a. At the time of the earliest geologic event	
The landscape is born, and is merely a plains that was forced upwards by a thrust fault.	
b. At the time of the fossil's existence	
The landscape did not have a river running through it, and it is still relatively young.	(plateau)

Figure 22: August Simmons's Level 1 GTS Written Response.

August's responses to part 4 of the performance assessment matched the description provided in the outcome space for level 1 on the geologic time and stratigraphy construct. In part 4a, he drew a dry landscape described as a "plains." In part 4b, he drew a plateau and did not include any information related to the fossil or the rock layer in which it was found. The level 1 descriptions in the outcome space specify that are the types of environments a student might draw if they *"[Make] few or no connections between events and the context in which they occurred,"* in accordance with the level 1 criterion on the construct map.

These two examples show the nature of the coherence between the outcome space used to generate scores and the written responses produced by students on the performance assessment task. Close matches like this lead to high levels of correlation, promoting coherence between student responses and scoring.

5.3.a.3 GTS RQ4 Comparison. This comparison pair describes the alignment between think-aloud codes and scores. There were 82 data pairs for items with both a think-aloud code and a score for this construct across the assessment for all subjects. Of those 82 pairs, 64 had identical values for both think-aloud code and score, meaning that the scores that were generated by the application of the outcome space aligned with the same construct level as their thinking 78% of the time. When the data were collapsed into holistic values, the agreement was at a similar level. For the 20 students who had holistic values for think-aloud codes, 16 of them (80%) had matching holistic scores on this construct.



1cGTS	1dGTS	1eGTS	2aGTS	2bGTS	2cGTS	2dGTS	2eGTS	3aGTS	3bGTS	3cGTS	3dGTS	3eGTS	3fGTS	4aGTS	4bGTS	4cGTS	4dGTS
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	0	0	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (1)	0	1	BLANK	BLANK	BLANK	NO TA	1	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	-1	-2	0	BLANK	BLANK	BLANK	-2	-1	FLAG (2)	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	1	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (1)	0	1	FLAG (2)	BLANK	BLANK	0	0	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	0	1	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (3)	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	0	0	FLAG (2)	FLAG (3)
BLANK	BLANK	BLANK	FLAG (3)	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	0	0	FLAG (4)	0
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	1	2	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	FLAG (1)	1	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	FLAG (3)	BLANK	BLANK	BLANK	0	0	2	BLANK	BLANK	BLANK	0	0	BLANK	FLAG (3)
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	NO TA	0	BLANK	BLANK	BLANK	0	0	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (2)	BLANK	FLAG (3)	0	0	BLANK	BLANK	BLANK	-2	0	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	-1	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	0	-1	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	0	0	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	FLAG (3)	BLANK	BLANK	BLANK	1	0	2	BLANK	BLANK	BLANK	1	0	BLANK	FLAG (3)
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO TA	0	0	BLANK	BLANK	BLANK	NO TA	NO TA	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	NO TA	0	BLANK	BLANK	BLANK	NO TA	NO TA	BLANK	BLANK

Figure 23: Difference Between Think-aloud Code and Score for the GTS Construct.

The difference table shown in Figure 23, along with the scatter plot in Figure 24 below, explains the high p-value of 0.18 for the comparison on individual responses. There is a large spread of scores at each level of think-aloud codes, including disparities of two levels in both directions. Although these instances are less frequent than the matches–which is visually apparent from the darker scatter plot dots on the matching pairs in Figure 24–they show that the overall conformity of these data to the ideal scenario is fuzzy at the item level. Item 3b, in which students made inferences about the past environment based on the fossil present in a rock sample, had the strongest agreement between think-aloud codes and scores, contributing to the 78% perfect match overall.

One item, 3c, has a different agreement profile than the other items targeting the Geologic Time & Stratigraphy construct. On this item, all instances of disagreement resulted from a score that was lower than the think-aloud code. As previously discussed, this is the item where students had to reconstruct a sequence of geologic events based on correlated rock outcrop diagrams. The assessment missed capturing much of the higher level thinking that students did because it didn’t give them a way to record their problem-solving process. The “how do you know” column was intended to be the improvement over the typical way in which questions of this nature are asked on the NYS Earth Science Regents–in which students fill in a predetermined number of blacks or select the correct list multiple choice options–but the complexity of the think-aloud data showed that this adjustment was insufficient. I found out that there was a wide variety of ways that students actually thought about this question, and asking them to explain their logic after the fact wasn’t always a sensical way of capturing it. Additionally, the physical space available on the page placed restrictions on the amount of information kids could write down. Finally, the outcome space emphasized the accuracy of correlation without taking into account all the other ways kids would demonstrate higher level thinking on this item. This is the main reason this item was consistently underscored.

This comparison has more flags overall than the two previous comparisons. There are 16 instances of items with a think-aloud code but no score for the geologic time & stratigraphy construct. It makes sense that there is an increased number of missed opportunities for capturing information when comparing think-aloud codes to scores, because it is a combination of two problems: first, instances where students thought (and said) something but didn’t write it down;

and second, instances where the scoring procedure didn't allow for scoring of that item. For these reasons it is expected that the think-aloud to scoring comparison would consistently have the greatest number of flags for each construct.

Of the sixteen instances with a think-aloud code and no score, three were on responses coded at think-aloud level 1, four were at think-aloud level 2, eight were at think-aloud level 3, and only one was at think-aloud level 4. The flags at think-aloud level 3 are distributed across a handful of items, so this seems like a broader issue with the way the level 3 construct map criteria were operationalized within the outcome space, as opposed to an item design problem.

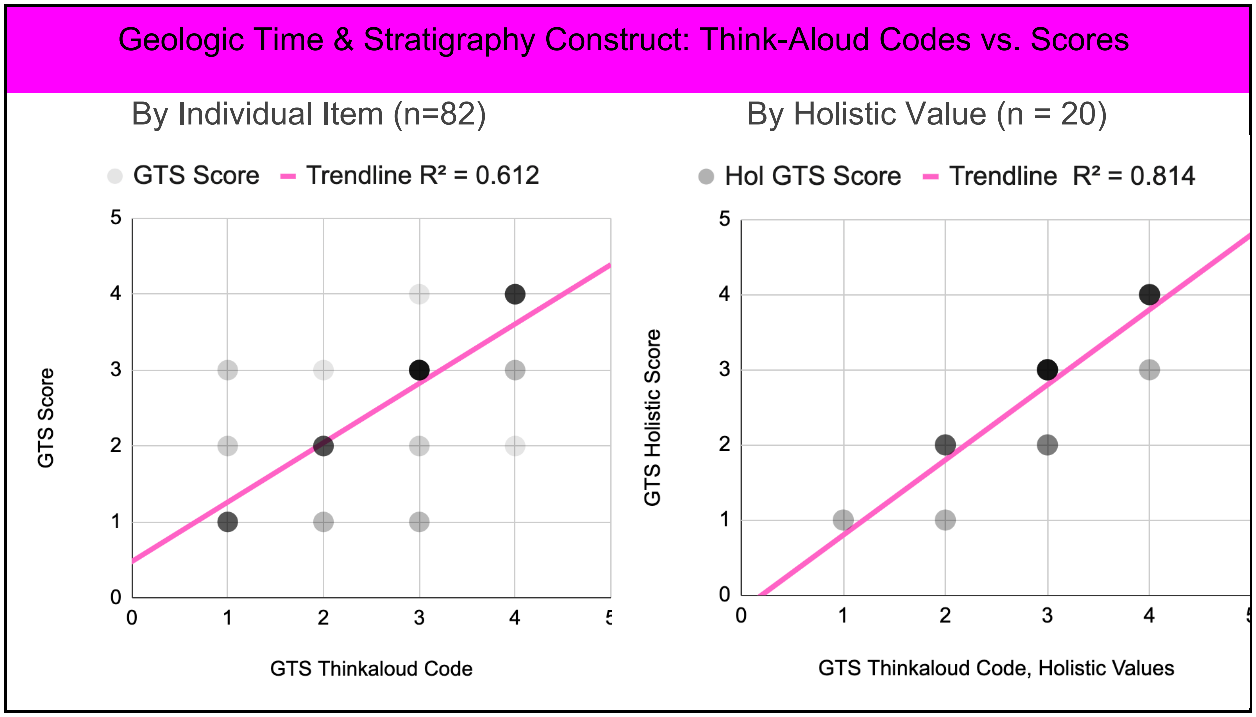


Figure 24: Linear Regressions on Individual Responses and Holistic Values Comparing GTS Think-Aloud Codes to Scores.

The holistic value comparison for the geologic time and stratigraphy construct was statistically significant with $R^2 = 0.81$ and a standard error of 0.38 ($p = 0.042$). This means that variations in student thinking accounted for about 80% of holistic scores for this construct. Figure 24 shows that agreement between holistic think-aloud codes and holistic scores was perfect at level 1. In cases of disagreement at levels 2, 3, and 4, the think-aloud code was always higher than the holistic score, suggesting that some aspects of student thinking about geologic time at higher construct levels have been lost in the assessment process. This was the case in only four instances: one student whose GTS think-aloud was holistically coded as level 4 and received a holistic GTS score of 3; two students whose think-alouds were coded as level 3 and received a holistic GTS score of 2; and one student whose GTS think-aloud was coded as level 3 and received a holistic GTS score of 1. The remaining 18 subjects (75% of 22 total subjects) all had holistic scores that exactly matched the holistic value for their think-aloud codes.

In the section presenting results for research question 3, comparing written codes to scores, I showed high- and low-level student examples for this construct. Here, I examine the same two students, this time to see the coherence across student thinking, as captured by the think-aloud, all the way to the score. First up is Juju, the student who was assigned a holistic score of 4 for the geologic time and stratigraphy construct, based in part on level 4 scores for these two items in the final part of the performance assessment:

Table 4: Juju’s Level 4 GTS Think-aloud.

4a	Okay, at the time of the earliest geologic event, I said that’s when limestone was forming. So I know it was a shallow sea, I know it was before the Devonian period, probably, or at least it was like the early Devonian. Because that fossil was Devonian so the rocks that formed before it were older. And, so, looking at that time, it was like … it could have been the Silurian, Ordovician, or even Cambrian time. But there’s no erosion between them, so probably not so far before the Devonian, I don’t really know. Let’s assume that it’s like, late Silurian, early Devonian or something. So it’s a shallow sea or ocean, and [...] there were probably organisms… oh, I know there were organisms because I know that limestone was made of shells. Um, I wonder what kind of shells lived then. I’m looking at my reference tables again. Looking at the index fossils. At that time, it looks like there’s definitely corals, and also the shell Eospirifer was there, possibly, that’s a brachiopod. So I’ll put possibly brachiopods, corals, could have been eurypterids, trilobites, etc. Okay, so I’m going to just draw water. And then, shells.[...] I’ll draw a little eurypterid… I don’t know if that looks like a eurypterid, and maybe some corals. It does say there were abundant corals. I’ll label this as “corals.” What else was alive then? [...] Um, M, oh M is Eurypterus. Oh I drew his head backwards. His little arms backwards. Sorry guy! That’s better. Okay, Eurypterus, and then E was Hexameroceras, that’s a Nautiloid. Okay, so lots of shells. Alright I think that’s good.
4b	At the time of the fossil’s existence. So I know this was the middle Devonian period. I know it was still underwater. And it was in siltstone. So I know that siltstone was really small sediments. And that’s probably like from really calm water, so maybe a little deeper. Well, I’ll make my water look calmer anyway. And I know that lots of trilobites were living there. So I know that, when I look at the graph on page 6, the stream velocity graph, I know that for silt the biggest silt particles are .006 and I know that would have a velocity -- a maximum velocity of like 0.4 centimeters per second [interpolating on graph with finger], and that’s pretty small. So I’ll just make my water look calm. And I know there was probably a layer of limestone on the bottom at this point. I’ll draw that. Because the siltstone was on top of the limestone. So there was like a limestone rock layer on the bottom of the ocean. And there was a silt bed. I’ll draw some silt sediments. It should be a pretty deep layer I think. They could’ve been from stuff on the land running into the ocean. I’m going to draw a Phacops here. I know they like to crawl on the bottom of the ocean. Um that doesn’t look too much like a Phacops. What do you look like, guy? [Picking up fossil / rock 2] You have a funny head. And little eye things. [...] At the same time Phacops was living, looking at the reference tables, lining up with C, I also see F, G, and X and Z. X is a little gastropod like a snaily kind of thing, and Z is a Mucrospirifer, maybe I’ll draw one of those. And then F and G.

This think-aloud example contains many specific instances in which Juju is *“visualizing and describing a landscape in the environmental condition it was in during the formation of rock layers,”* as specified by level 4 of the construct map for geologic time & stratigraphy. This aural report of their thinking contains significantly more details than their written response (seen in Figure 21), such as their use of sediment size in the siltstone rock to make an inference about the nature of the water in the depositional environment, or their explicit connection between the composition of the limestone rock and the ecosystem that likely existed in the prehistoric ocean where the lithification of that limestone occurred. Even though these details did not make it into their response on paper, their written work was scored at the same level. Why? Because the outcome space for these items was written at the appropriate grain size to match the way these details of their thinking were manifested in her response. The outcome space associated with the scoring procedure (see Figure 20 in a previous section) was an effective translation tool for maintaining coherence between the student's thinking about the task, their written response to the task, and the cognitive model about the construct in which this part of the task was based.

The scoring procedure was similarly successful in the case of this student who exhibited thinking that corresponded to level 1 on the construct map.

> August Simmons: When you show a local landscape, at the time of the earliest geological events. Okay. So if the fossil existence is a second thing, I imagine it's the Precambrian eon.
>
> Rabi: You're looking at page eight in the reference tables. What are you thinking about?
>
> August Simmons: I'm thinking about the time. So if it's the Fossil, which is trilobite, what'd you do from the Cambrian area era ... Period. Period. That's for part b. A, has to be before that. So that means it's the Precambrian eon. Before there was much life. I think, yeah. Land formed before creatures existed. So it was land.

The construct map criterion at level 1 for geologic time and stratigraphy is very general, stating that students at this level make "*few or no connections between events and the context in* *which they occurred.*” In this case, the student is connecting his ideas to the fossil observed in a different rock sample in order to make some very broad-stroked assumptions, but he is not making any connections to the bioclastic rock units he previously identified as being the earliest to form in this region. The outcome space (seen above in Figure 20) characterizes a response that makes “few or no connections” to the rock unit context as including “dry land.” This matches the description and illustration produced by the student.

5.3.a.4 Summary: Geologic Time & Stratigraphy. For the geologic time & stratigraphy construct overall, I found that student thinking about this construct was elicited at all four levels of the construct map across the entire assessment. Holistic values for codes and scores on this construct consistently provided a good match with the ideal correlative model. The overall data pattern shows that some higher-level thinking was lost in the assessment process. The holistic scores trended slightly lower than holistic think-aloud codes.

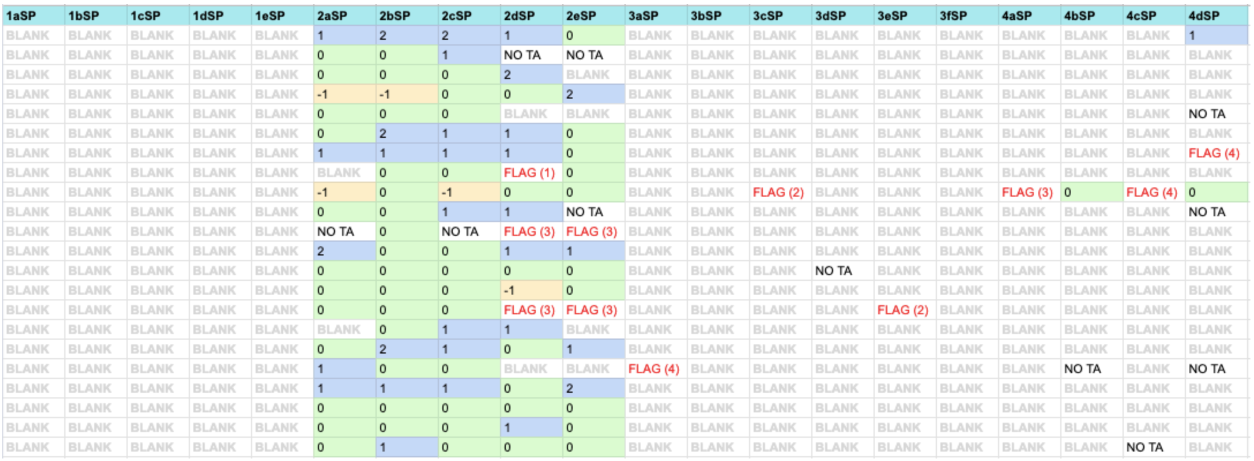
5.3.b Surface Processes Construct

Like the previous construct, the surface processes construct also has stronger correlations on holistic values than on individual responses. A summary of the coherence data for the Geologic Time and Stratigraphy construct is shown in Table 5, below.

Table 5: SP Coherence Data.

Pair	Comparison Method	R^2	Standard error	P value	Perfect Match	Flags
WC vs TA (RQ2)	Individual Responses (94)	0.57	0.71	0.000017	62%	11
	Holistic Values (22)	0.59	0.75	0.0012	59%	
Score vs WC (RQ3)	Individual Responses (96)	0.77	0.45	0.0061	83%	8
	Holistic Values (22)	0.87	0.36	0.083	86%	
Score vs TA (RQ4)	Individual Responses (95)	0.65	0.65	0.0014	68%	11
	Holistic Values (22)	0.74	0.60	0.0019	73%	

5.3.b.1 SP RQ2 Comparison. There were 94 data pairs with codes for both the think-aloud and the written response for this construct across the assessment for all subjects. Of those 94 pairs, 58 had identical values for both think-aloud and written codes, meaning that the students' recorded responses aligned with the same construct level as their thinking 62% of the time. The agreement level was similar for the holistic comparison. All 22 subjects had holistic think-aloud codes and written code values. Thirteen of them, or 59%, had a perfect match between the two.



1aSP	1bSP	1cSP	1dSP	1eSP	2aSP	2bSP	2cSP	2dSP	2eSP	3aSP	3bSP	3cSP	3dSP	3eSP	3fSP	4aSP	4bSP	4cSP	4dSP
BLANK	BLANK	BLANK	BLANK	BLANK	1	2	2	1	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	1
BLANK	BLANK	BLANK	BLANK	BLANK	0	0	1	NO TA	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	2	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	-1	-1	0	0	2	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO TA
BLANK	BLANK	BLANK	BLANK	BLANK	0	2	1	1	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	1	1	1	1	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (4)
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	FLAG (1)	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	-1	0	-1	0	0	BLANK	BLANK	FLAG (2)	BLANK	BLANK	BLANK	FLAG (3)	0	FLAG (4)	0
BLANK	BLANK	BLANK	BLANK	BLANK	0	0	1	1	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO TA
BLANK	BLANK	BLANK	BLANK	BLANK	NO TA	0	NO TA	FLAG (3)	FLAG (3)	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	2	0	0	1	1	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	0	0	BLANK	BLANK	BLANK	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	-1	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	FLAG (3)	FLAG (3)	BLANK	BLANK	BLANK	BLANK	FLAG (2)	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	1	1	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	2	1	0	1	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	1	0	0	BLANK	BLANK	FLAG (4)	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO TA	BLANK	NO TA
BLANK	BLANK	BLANK	BLANK	BLANK	1	1	1	0	2	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	1	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	1	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO TA	BLANK

Figure 25: Difference Between Think-aloud Code and Written Code for the SP Construct.

The difference table for this comparison shows why the rate of perfect matching is lower for this construct. Every item targeting the surface processes construct failed to provide a perfect match between the think-aloud and written data at least one third of the time. Moreover, there is a clear pattern in the nature of the mismatched data: in nearly all cases where there is a disagreement between the two, the think-aloud code is one or two levels higher than the written code. There does not appear to be any item that is significantly better or worse than the others at accurately capturing student thinking in their written work.

There were eleven flagged items in which the student had a think-aloud code but no written code. These flags showed up on two of the items intended to measure the Surface Processes construct, 2d and 2e. These items required students to make a prediction about an alternate reality in which the surface had not been eroded by water. These were the items most likely to elicit student thinking about surface processes that they did not record in their written response.

There were eleven missed opportunities for capturing student thinking about the surface processes construct. One of these occurred at a think-aloud level of 1. Two occurred at think-aloud level 2. Five of these flagged responses were at think-aloud level 3, and three were at think-aloud level 4. This suggests that more sophisticated thinking was less likely to be captured, or preferentially more likely to be missed.

On individual responses, R^2 was 0.57 with a standard error of 0.71 ($p < 0.00002$). There was a similar level of correlation between the think-aloud and written codes on individual item responses and on holistic values. The variation in student thinking accounted for a little less than 60% of the variation in student written responses.

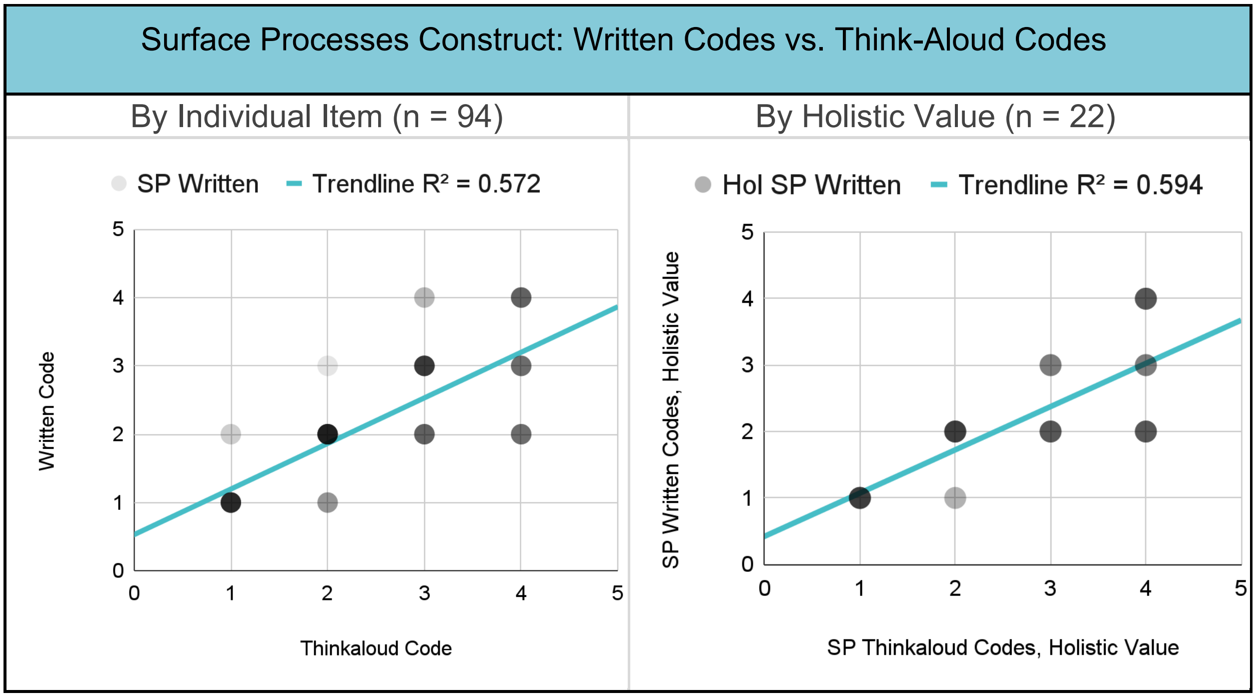
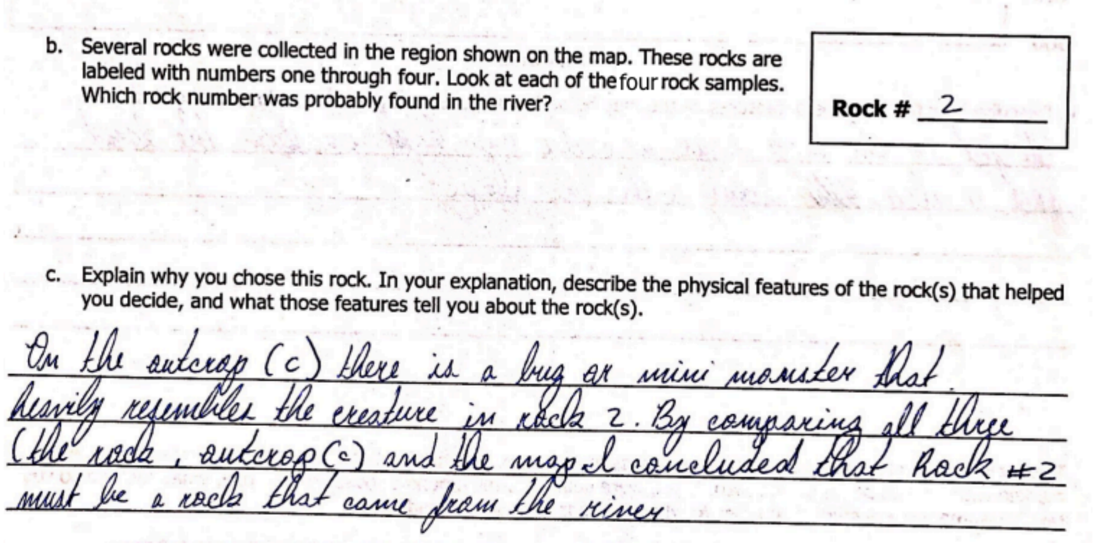


Figure 26: Linear Regressions on Individual Responses and Holistic Values Comparing SP Think-Aloud Codes to Written Codes.

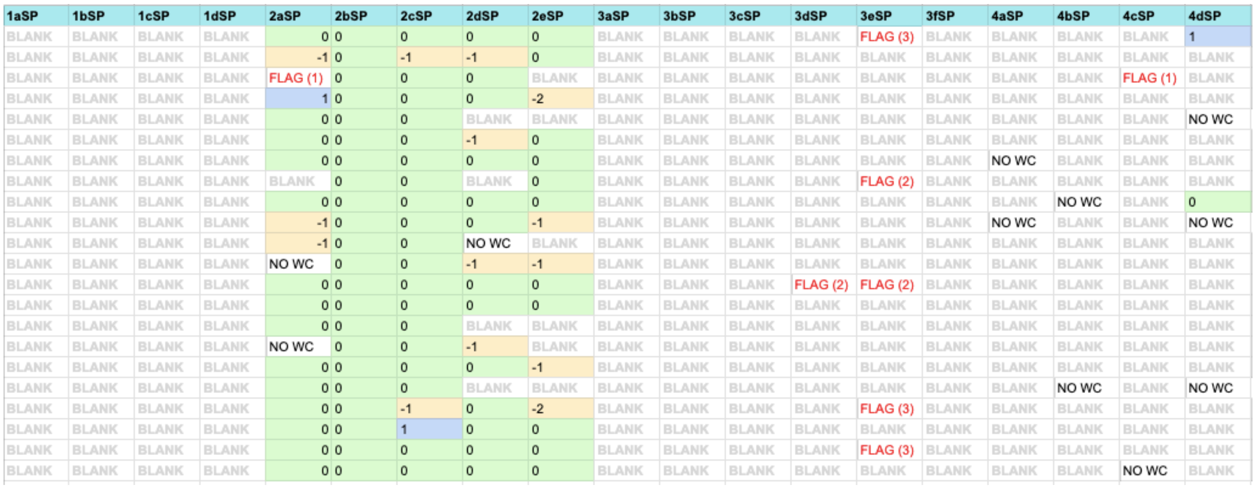
As is evident from the bar graphs for this construct in Figure 12, the similarity of this fit to the ideal linear relationship is stronger at lower levels than at levels 3 and, especially, 4. This is also visible in the scatter plots in Figure 26. For both the item-wise and holistic comparisons, the trendline intersects with a pairing of think-aloud level 4 and written code level 3. This indicates a systemic loss of a full level at the most sophisticated end of the construct map. The scatter plot for holistic values shows how the range of written output increased at higher think-aloud levels. The student response sheet should be revised to more effectively give students a reason and method to record their higher-level thinking about this construct.

The example shown in Table 6, from part 2 of the performance assessment, shows how Teddy's written answer did not reflect the entirety of her thinking. When considering which of several rock samples was likely to have been found in a moving river, she talked about the specific weathering and erosion processes that would have influenced the rock's appearance. However, since she ultimately wrote down a different (and incorrect) idea, her written response was not coded at the same level for the surface processes construct.

Table 6: Teddy's Misaligned SP Think-Aloud and Written Response.

Think-aloud	"This one, rock 4, looks like it belongs in the water. [Picking up rock 4] It resembles a pebble, but it's bigger than a pebble. It looks like it was in the water because of… just like, the marble, and how it looks. Pebbles look like this. So I assume that rock 4 is from the water. Also, the shiny parts in here. I don't know if I'm right, but aren't these like, from volcanoes? So it could have blasted and then fell in the water. Then, from being in the water and moving around in there [rotating rock in hands], its shape gets changed into this where it was worn down. [Gesturing to rounded corners on rock samples]. But also, this thing [touching fossil in rock 2] is in outcrop C. And the map shows outcrop C right next to the river. So maybe it could have fallen in."
Written Response	 b. Several rocks were collected in the region shown on the map. These rocks are labeled with numbers one through four. Look at each of the four rock samples. Which rock number was probably found in the river? Rock # 2 c. Explain why you chose this rock. In your explanation, describe the physical features of the rock(s) that helped you decide, and what those features tell you about the rock(s). On the outcrop (c) there is a bug or mini monster that heavily resembles the creature in rock 2. By comparing all three (the rock, outcrop (c) and the map I concluded that Rock #2 must be a rock that came from the river *On the outcrop (C) there is a bug or mini monster that heavily resembles the creature in rock 2. By comparing all three (the rock, outcrop (C) and the map, I concluded that rock #2 must be a rock that came from the river.*

5.3.b.2 SP RQ3 Comparison. There were 96 data pairs with values for both the written response and the scoring procedure for the Surface Processes construct across the assessment for all subjects. Of those 96 pairs, 80 had identical values, meaning the students' scores reflected the coded level of their written responses 83% of the time. The holistic values had similar agreement as individual responses. Across the data set of 22 students, 19 of them (86%) had holistic scores that exactly matched their holistic written codes.



1aSP	1bSP	1cSP	1dSP	2aSP	2bSP	2cSP	2dSP	2eSP	3aSP	3bSP	3cSP	3dSP	3eSP	3fSP	4aSP	4bSP	4cSP	4dSP
BLANK	BLANK	BLANK	BLANK	0	0	0	0	0	BLANK	BLANK	BLANK	BLANK	FLAG (3)	BLANK	BLANK	BLANK	BLANK	1
BLANK	BLANK	BLANK	BLANK	-1	0	-1	-1	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	FLAG (1)	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (1)	BLANK
BLANK	BLANK	BLANK	BLANK	1	0	0	0	-2	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO WC
BLANK	BLANK	BLANK	BLANK	0	0	0	-1	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	0	0	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO WC	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	0	BLANK	0	BLANK	BLANK	BLANK	BLANK	FLAG (2)	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	0	0	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO WC	BLANK	0
BLANK	BLANK	BLANK	BLANK	-1	0	0	0	-1	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO WC	BLANK	BLANK	NO WC
BLANK	BLANK	BLANK	BLANK	-1	0	0	NO WC	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	NO WC	0	0	-1	-1	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	0	0	0	0	0	BLANK	BLANK	BLANK	FLAG (2)	FLAG (2)	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	0	0	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	NO WC	0	0	-1	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	0	0	0	0	-1	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO WC	BLANK	NO WC
BLANK	BLANK	BLANK	BLANK	0	0	-1	0	-2	BLANK	BLANK	BLANK	BLANK	FLAG (3)	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	0	0	1	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	0	0	0	0	0	BLANK	BLANK	BLANK	BLANK	FLAG (3)	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	0	0	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO WC	BLANK

Figure 27: Difference between Written Codes and Scores for SP Construct.

This represents a strong match, and the difference table in Figure 27 shows that item 2b is perfect: every student received the same written code and score for this item. The outcome space and scoring guide for this construct worked well to translate student responses into scores that accurately reflected the progression of thinking described by the construct map. Where there is a mismatch, written codes are more likely to be lower compared to scores.

There are eight flags in this difference table indicating missed opportunities for scoring this construct based on students' written work. Two of these flags occurred on responses that were coded at level 2. Three each occurred on responses that were coded at levels 2 and 3. No written responses coded at level 4 failed to be scored. There is a concentration of flags in item

3e, suggesting that this item represents a potential opportunity to intentionally record and score thinking about surface processes. Conversely, there is only one flag in the items designed to capture thinking about surface processes. It seems that the structures and prompts provided to students on these items allowed them to record ideas about this construct overall, if not always the most sophisticated aspects of their thinking.

The linear regression analysis on Surface Processes data for this question shows that the holistic values provide the most accurate representation of student written work. On individual items, $R^2 = 0.77$ ($p = 0.0061$), while on holistic values, $R^2 = 0.87$ ($p = 0.083$). This p-value is outside the limits of significance but not by a lot. There is still a greater than 90% chance that this is a real, non-coincidental relationship. Both these scatter plots adhere closely to the ideal linear model.

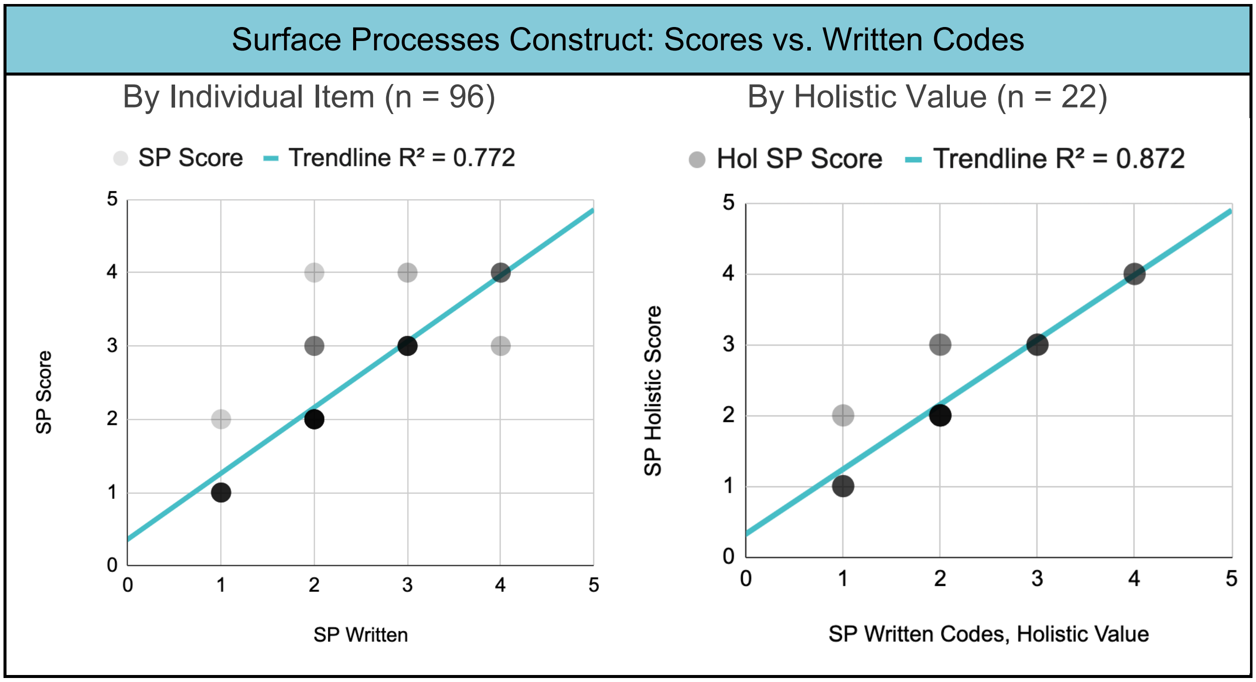
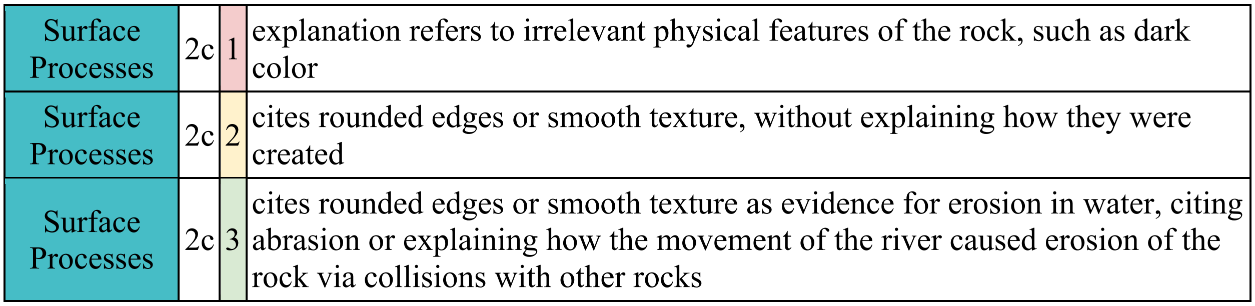


Figure 28: Linear Regressions on Individual Responses and Holistic Values Comparing SP Written Codes to Scores.

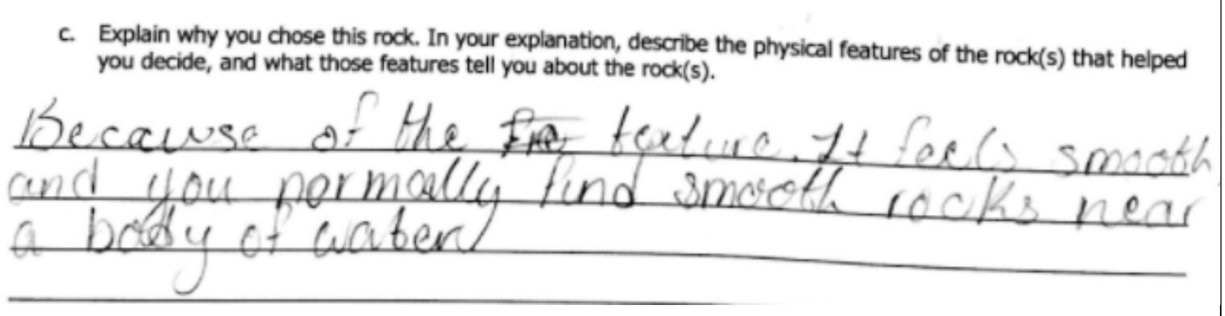
The following example from item 2c shows the coherence between a student response that aligns with level 2 on the construct map and the level 2 outcome space for surface processes. The construct map specifies that students thinking at level 2 "*recognize evidence of weathering and/or erosion processes.*" The outcome space shown in Figure 29 defines what this would look like in student responses to this item, in which they are identifying which of several rock samples was likely transported by a moving river.



Surface Processes	2c	1	explanation refers to irrelevant physical features of the rock, such as dark color
Surface Processes	2c	2	cites rounded edges or smooth texture, without explaining how they were created
Surface Processes	2c	3	cites rounded edges or smooth texture as evidence for erosion in water, citing abrasion or explaining how the movement of the river caused erosion of the rock via collisions with other rocks

Figure 29: SP Outcome Space for Item 2c.

Lucy's written response, shown in Figure 30, is a clear match for the level 2 outcome space for this item. Her written response indicates she is recognizing the physical evidence that erosion that has occurred to this cobble. The outcome space makes it clear that she should be scored at level 2 even though she has not explicitly stated what process created that physical characteristic.



c. Explain why you chose this rock. In your explanation, describe the physical features of the rock(s) that helped you decide, and what those features tell you about the rock(s).

Because of the ~~Fr~~ texture. It feels smooth and you normally find smooth rocks near a body of water.

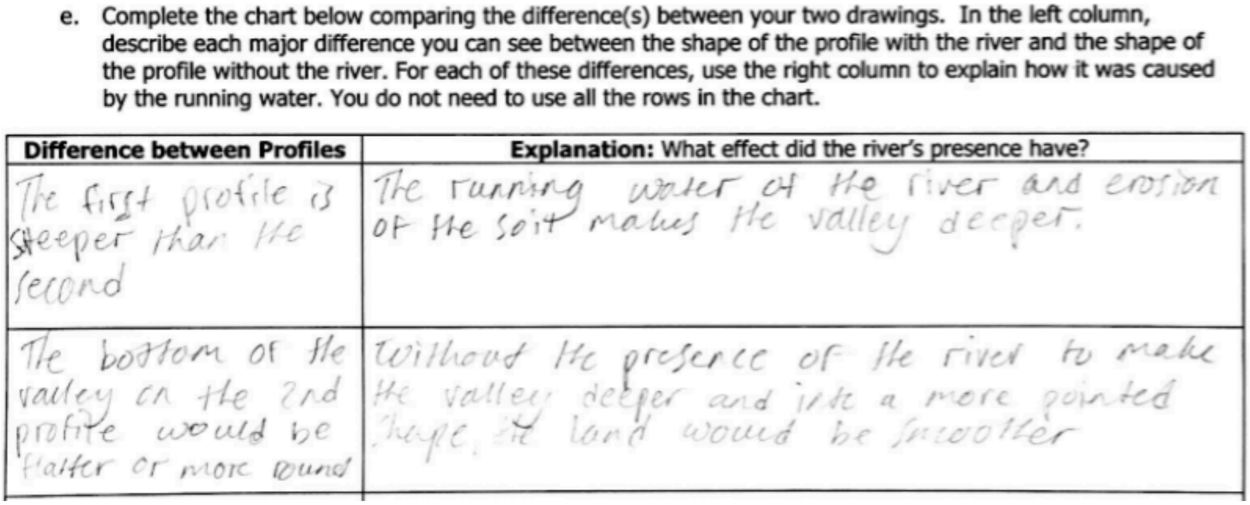
Because of the texture. It feels smooth and you normally find smooth rocks near a body of water.

Figure 30: Lucy's Level 2 SP Written Response.

On individual responses, the scores that did not match written codes were systematically higher than those written codes. Here is one example of how that happened on item 2e. In this part of the performance assessment, students had sketched an alternative profile for the map region in the imaginary scenario where there was no river running through the landscape. They then compared the riverless profile to the landscape profile that did include the river. These assessment items were intended to elicit thinking at levels 2 - 4 of the construct map for surface processes. At level 3, students *"recognize evidence for past processes in current landscape features,"* while at level 4 they are "*able to predict patterns [...] of surface change."* Figure 31 shows how these criteria were operationalized by the outcome space.

Surface Processes	2e	3	describes erosion as a result of moving water that created the valley, using a correct definition
Surface Processes	2e	4	identifies and describes change in both elevation and slope due to erosion, stating that the land would remain un-eroded (higher elevation, flatter slope) in the absence of water
Surface Processes	2e	4	identifies and describes change in both elevation and slope due to erosion, stating that the water is responsible for creating the valley's depth (lower elevation) and steeper sloped sides.

Figure 31: SP Outcome Space for Item 2c, Levels 3 and 4.



e. Complete the chart below comparing the difference(s) between your two drawings. In the left column, describe each major difference you can see between the shape of the profile with the river and the shape of the profile without the river. For each of these differences, use the right column to explain how it was caused by the running water. You do not need to use all the rows in the chart.

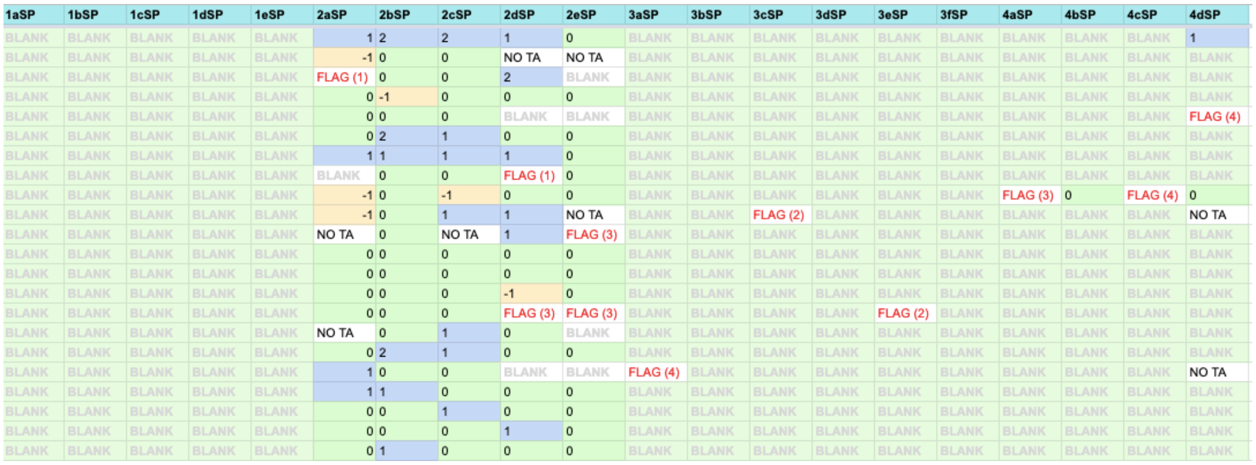
Difference between Profiles	**Explanation:** What effect did the river's presence have?
The first profile is steeper than the second	The running water of the river and erosion of the soil makes the valley deeper.
The bottom of the valley on the 2nd profile would be flatter or more round	Without the presence of the river to make the valley deeper and into a more pointed shape, the land would be smoother

Figure 32: T-Noona's Written Response to Item 2c. This response was coded at SP level 3 and scored at SP level 4.

This student's written response was coded at level 3. Her explanation that "the running water of the river and erosion of the soil makes the valley deeper" shows that she is recognizing evidence of river erosion in the present-day landscape feature of the v-shaped valley. Because her written explanation said that the land would be "smoother" in the absence of the river, this was not coded as an accurate prediction of surface change. However, the outcome space provides clearer guidelines for a score at level 4 by naming elevation and slope as the two key components. This student was assigned a score of 4 for this item on the basis of those criteria, since she discussed the depth of the valley as well as the steepness of the profile. This shows the utility of the outcome space as part of the scoring procedure in providing ready-made interpretations of the construct map in the context of the specific assessment task students are performing.

This same pattern, in which any mismatched scores are one level higher than the written code, is also seen in the holistic values, with one instance of this happening at each of written codes level 1 and level 2. However, this did not occur at written code levels 3 and 4.

5.3.b.3 SP RQ4 Comparison. There were 95 data pairs for items with both a think-aloud code and a score for the surface processes across the assessment for all subjects. Of those 95 pairs, 65 had identical values for both think-aloud code and score, meaning that the scores represented the same level of understanding as the students' thinking 68% of the time. When the data were collapsed into holistic values, the agreement improved. For the 22 students who had holistic values for think-aloud codes, 16 of them (73%) had matching holistic scores on this construct. The coefficients of determination show a similar pattern, with an R^2 of 0.65 on individual responses increasing to an R^2 of 0.74 on holistic values.



1aSP	1bSP	1cSP	1dSP	1eSP	2aSP	2bSP	2cSP	2dSP	2eSP	3aSP	3bSP	3cSP	3dSP	3eSP	3fSP	4aSP	4bSP	4cSP	4dSP
BLANK	BLANK	BLANK	BLANK	BLANK	1	2	2	1	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	1
BLANK	BLANK	BLANK	BLANK	BLANK	-1	0	0	NO TA	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (1)	0	0	2	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	-1	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (4)
BLANK	BLANK	BLANK	BLANK	BLANK	0	2	1	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	1	1	1	1	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	FLAG (1)	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	-1	0	-1	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (3)	0	FLAG (4)	0
BLANK	BLANK	BLANK	BLANK	BLANK	-1	0	1	1	NO TA	BLANK	BLANK	FLAG (2)	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO TA
BLANK	BLANK	BLANK	BLANK	BLANK	NO TA	0	NO TA	1	FLAG (3)	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	-1	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	FLAG (3)	FLAG (3)	BLANK	BLANK	BLANK	BLANK	FLAG (2)	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	NO TA	0	1	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	2	1	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	1	0	0	BLANK	BLANK	FLAG (4)	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO TA
BLANK	BLANK	BLANK	BLANK	BLANK	1	1	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	0	1	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	1	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	1	0	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK

Figure 33: Difference Between Think-alouds and Scores for SP Construct.

The difference table for the comparison of think-alouds to scores shows that no single item stood out, but 2b was the most successful overall. This has been consistent throughout the different comparisons for the surface processes construct. This is interesting because item 2b was assigned a restricted scoring range, so there was potential for large disagreements between think-aloud codes and scores. The fact that there are so many perfect matches here, even though the max on the scoring guide was a level 2, shows that this item was successful at prompting thinking of a specific nature and level within this construct. Item 2e had perfect agreement for all instances where both TA and written response were able to be coded, but also had more flags and blanks. The item could be improved by providing more specific prompts and scaffolds to students to ensure they have the opportunity to address this construct in their response.

This construct has a relatively high level of mismatch: one-third of the response pairs had different values when going from think-aloud to the final score. Disagreements were much more likely to have a positive difference between think-aloud and score, meaning the think-aloud code was usually higher when there was a disagreement between the two. In addition, there were eleven missed opportunities for scoring student thinking. Fewer than half of these flags were on items designed to observe and score the surface processes construct. Two occurred at think-aloud level 1, two at think-aloud level 2, four at think-aloud level 3, and three at think-aloud level 4.

For the surface processes construct, R^2 values were higher on holistic values than for individual item responses when correlating think-alouds and scores. Up to 74% of variation in student scores was determined by variations in their thinking about surface processes based on this linear regression, shown in Figure F below. The standard error of 0.60 for holistic values shows that there is unwanted noise in the scores.

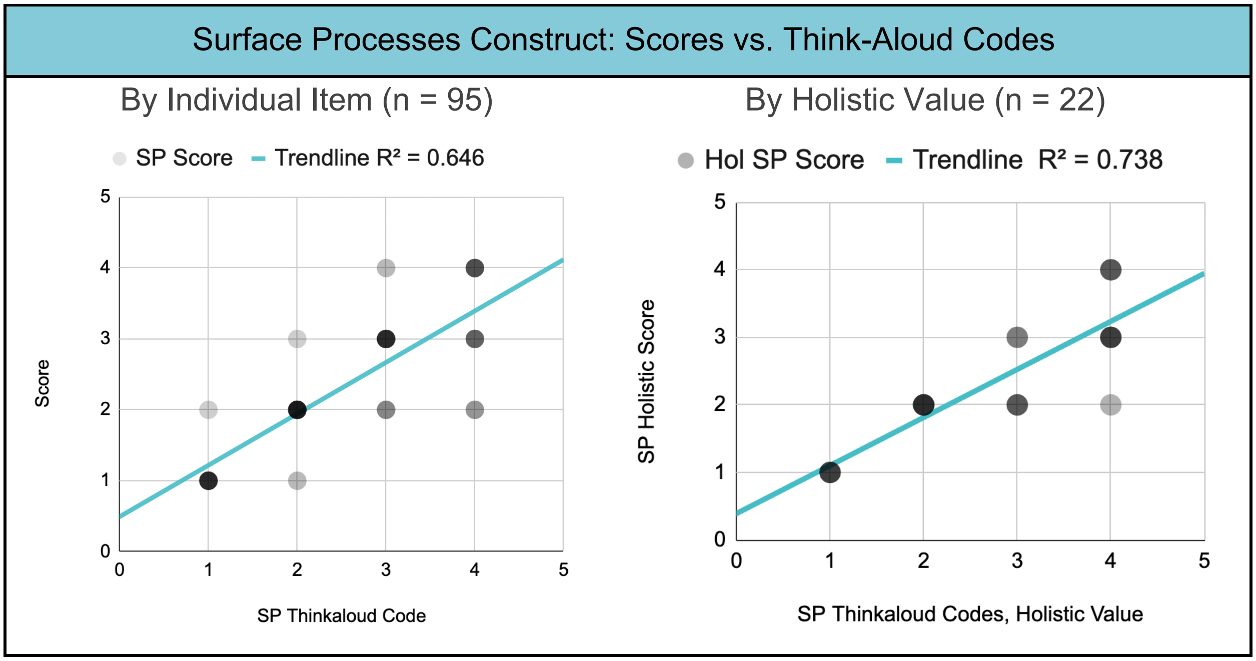


Figure 34: Linear Regressions on Individual Responses and Holistic Values Comparing SP Think-aloud Codes to Scores.

A closer look at the data reveals that the variation in scores is greater at the higher levels of student thinking. Nine students in this 22-student sample were assigned holistic think-aloud codes of level 1 or 2, reflecting that was the most common code in their think-aloud data for the surface processes construct. All nine of them received a matching holistic score via the scoring procedure for this construct. Five students were assigned holistic think-aloud scores of level 3. Of those, three received a level 2 holistic score, and the remaining two received a level 3 holistic score. Eight students were assigned holistic think-aloud codes of level 4. Of those, one received a level 2 holistic score, four received a level 3 holistic score, and three received a level 4 holistic score. It was slightly more common in this group of subjects for the students with more sophisticated thinking about this construct to be underscored than to have scores that reflected their level of thinking.

In the findings for a previous question, I showed an example of a student, Lucy, whose response was coherently coded and scored at level 2. This is her thinking aloud for that same item:

> “Now, it's asking me which rock sample would be probably from by the river. And I would say four because the way how it feels and I remember, well *I* never went to the beach, but I remember my friend telling me, found this rock and she brung it home. She brought it home to show me. I remember how like smooth it was, which was nearby the, um, the ocean. So I was figuring that it was near the ocean, and a river is made of water just like the ocean. So I figure this one [holding up rock sample 4] would be in water because of the texture of it [rubbing palm across rock surface]. And also how it looks. It seems like there's bunch of little minerals bashed together. And so that's what I'm thinking is in the river, the rocks like this is the ones in the water, so I'm going to just put [rock] 4.”

By referring back to the examples shown previously (see Figures 29 and 30), we can see how the important aspects of the student’s thinking with respect to this construct were captured on paper, and how the outcome space prompted it to be scored at a level 2. This may seem both simple and straightforward, but it is important to keep in mind the significance of the coherence extending to student thinking. This student’s thinking is clearly in alignment with the level 2 criterion in the construct map. She has not expressed ideas or understandings that imply her thinking is at a different level of complexity. The numerical value assigned to her work via the scoring procedure captures the alignment of her thinking to the cognitive model for surface processes learning. This is the goal of cognitively based performance assessment.

The following example is the think-aloud for the same student, T-Noona, who had a disparity between her written response code and her holistic score code, described previously in the surface processes section for research question 3.

> "Oh yeah, the first one was steeper, that's for sure. Than the one without the river. So would that be a difference? Yeah. The first profile is steeper. Explanation? So it would have been because the river, the water is like, making the valley deeper. I'm trying to look at the first profile. Trying to compare it with the second profile. I know in the second profile, when it dips down into a valley and then came back out I tried to make it more round. Because if the river wouldn't be there, the bottom would be more round and gentle. With the river it just keeps going down, down, down. In the first one way the river goes down, it makes kind of a sharp point at the bottom, that's the V shape that water makes. Without the river there it would be more… round… because the water wouldn't be making the valley deeper when it cuts down into the land. Round meaning not as much of a pointy V."

From this student's verbalized thinking, it is clear that she is able to make an accurate prediction or supposition about how this landscape would be affected by erosion due to running water over time. This is in line with the level 4 description on the surface processes construct map, which says students are *"able to predict patterns or events of surface change based on past evidence."* In this case, she shows that she understands how the continuous action of moving water would lead to the development of a v-shaped valley, associating the depth and angle of the valley walls with the change caused by river erosion over a period of time. She correctly surmises that the v-shape would be unlikely to form in the absence of this moving river. Although very little of the detail in her thinking was ultimately captured in writing, the outcome space anticipated the most important features that a student with this kind of thinking might include in their written response, leading to a level 4 score that matched her level 4 think-aloud code.

5.3.b.4 Summary: Surface Processes. In RQ2, comparing think-alouds to written codes, there was greater variation as think-aloud level increased. In RQ3, comparing written codes to scores, matching of holistic values was strong at levels 3 and 4 and only minor disagreement was likely at levels 1 and 2. I think this tells us that the threat to coherence for this construct lies in the ability of the assessment task to capture student thinking at higher levels. The scoring procedure assigns value to student written responses in a more coherent and consistent manner. The pattern in RQ4, comparing think-alouds to written codes, is consistent with that from RQ2. Overall for this construct, agreement between the written codes and scores was quite strong, while variation was introduced in the step of capturing student thinking via their written responses to the assessment.

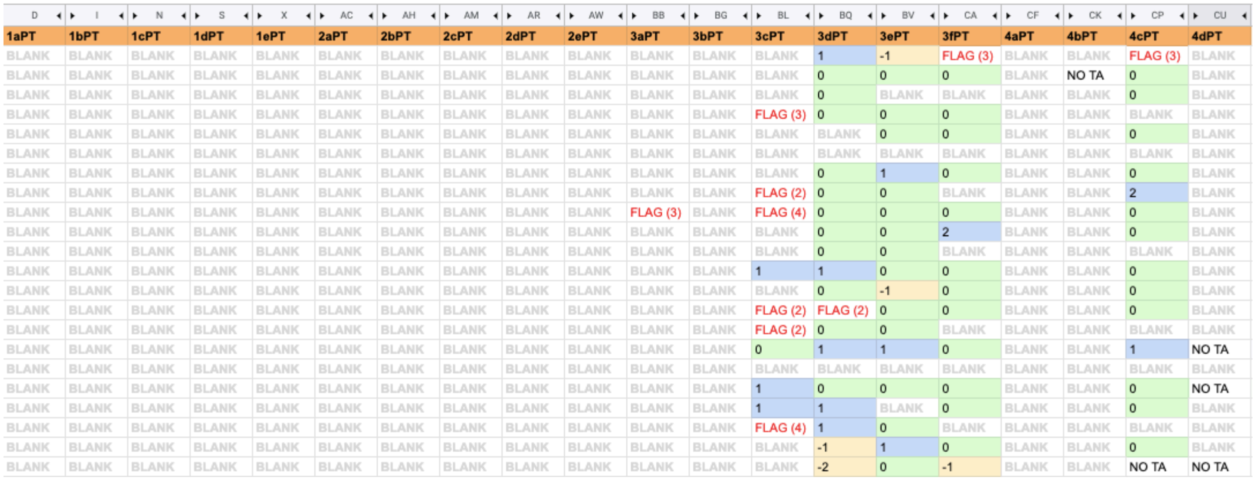
5.3.c Plate Tectonics Construct

The coherence data for the Plate Tectonics construct is summarized in Table 7.

Table 7: PT Coherence Data.

Pair	Comparison Method	R^2	Standard error	P value	Perfect Match	Flags
WC vs TA (RQ2)	Individual Responses (66)	0.46	0.57	0.16	83%	10
	Holistic Values (20)	0.66	0.50	0.33	75%	
Score vs WC (RQ3)	Individual Responses (68)	0.93	0.34	0.32	94%	5
	Holistic Values (20)	0.80	0.38	0.04	85%	
Score vs TA (RQ4)	Individual Responses (63)	0.43	0.59	0.24	75%	13
	Holistic Values (20)	0.58	0.55	0.21	70%	

5.3.c.1 PT RQ2 Comparison. There were 66 data pairs with codes for both the think-aloud and the written response for the plate tectonics construct across the assessment for all subjects. This is many fewer data points than for other constructs. Of those 66 pairs, 55 had identical values for both think-aloud and written codes, meaning that the students' recorded responses aligned with the same construct level as their thinking 83% of the time. This match is better than expected given the low R^2. Unlike on some other constructs, when the data were collapsed into holistic values, the agreement was lower. For the 20 students who had holistic values for think-aloud codes, 15 of them (75%) had matching holistic written codes on this construct. Due to the high p-values resulting from t-tests on these pairs, I cannot claim that these patterns are statistically significant.



D	I	N	S	X	AC	AH	AM	AR	AW	BB	BG	BL	BQ	BV	CA	CF	CK	CP	CU
1aPT	1bPT	1cPT	1dPT	1ePT	2aPT	2bPT	2cPT	2dPT	2ePT	3aPT	3bPT	3cPT	3dPT	3ePT	3fPT	4aPT	4bPT	4cPT	4dPT
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	1	-1	FLAG (3)	BLANK	BLANK	FLAG (3)	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	NO TA	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (3)	0	0	0	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	1	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (2)	0	0	BLANK	BLANK	BLANK	2	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (3)	BLANK	FLAG (4)	0	0	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	2	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	1	1	0	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	-1	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (2)	FLAG (2)	0	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (2)	0	0	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	1	1	0	BLANK	BLANK	1	NO TA
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	1	0	0	0	BLANK	BLANK	0	NO TA
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	1	1	BLANK	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (4)	1	0	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	-1	1	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	-2	0	-1	BLANK	BLANK	NO TA	NO TA

Figure 35: Difference Between Think-aloud Codes and Written Codes for PT Construct.

The difference table for this comparison shows that flags, or missed opportunities for capturing student thinking, are primarily coming from item 3c. This item was not intended to target the plate tectonics construct, but not only did it come up in students' reasoning (indicated by the flags), it was also expressed in some of their recorded ideas (the four instances that had paired think-aloud and written codes). The pattern of flags is consistent with the overall pattern of thinking elicited by this construct, which was heavily clustered at levels 2 and 3. The clustering is visible in the scatter plots in Figure 36 below. There were no missed opportunities to code think-aloud level 1 for plate tectonics. Levels 2 and 3 had four missed opportunities each, and there were two instances of level 4 thinking that did not get captured by a student's written response.

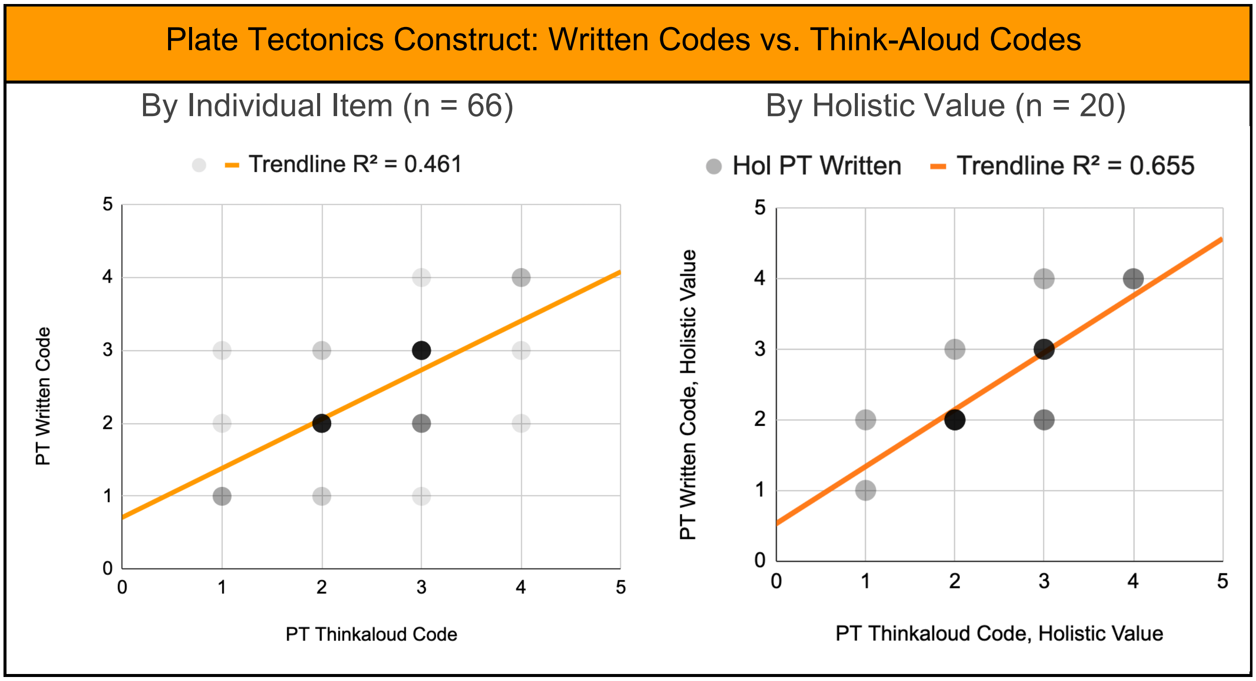
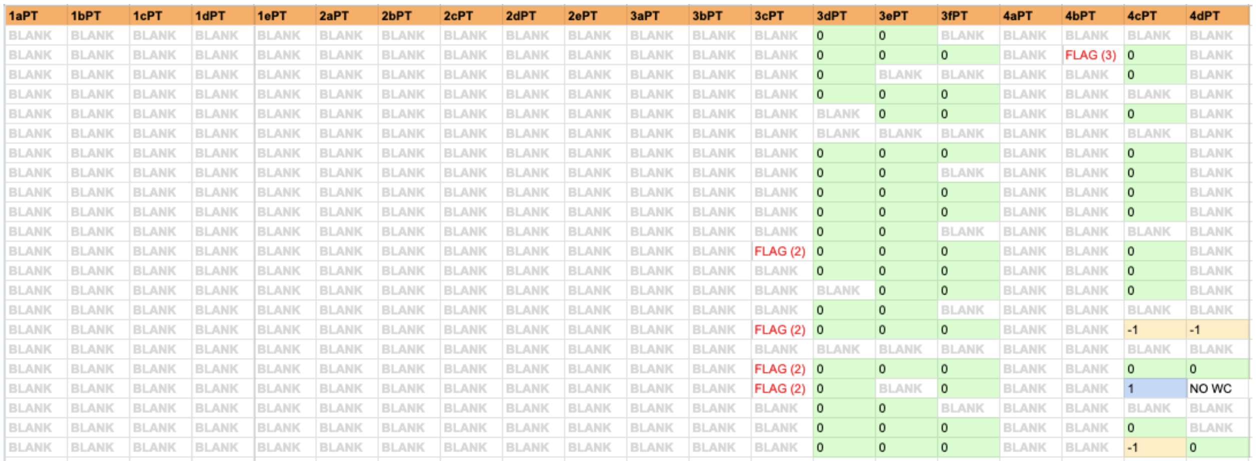


Figure 36: Linear Regressions on Individual Responses and Holistic Values Comparing PT Think-aloud Codes to Written Codes.

5.3.c.2 PT RQ3 Comparison. The plate tectonics construct had strong coherence at the level of individual items. There were 68 data pairs with values for both the written response and the scoring procedure for the construct across the assessment for all subjects. Of those 68 pairs, 64 had identical values. The students' scores reflected the coded level of their written responses 94% of the time. Once again, a high p-value shows that this correlation may be due to random chance. This alignment degenerated in the holistic comparison, where there was only an 85% perfect match. However, a clearer pattern emerged in this modality, and the p-value of 0.04 indicates that there is a statistically significant relationship between the holistic written codes and holistic scores, with 80% of the variation in scores being attributable to variation in student responses.



1aPT	1bPT	1cPT	1dPT	1ePT	2aPT	2bPT	2cPT	2dPT	2ePT	3aPT	3bPT	3cPT	3dPT	3ePT	3fPT	4aPT	4bPT	4cPT	4dPT
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	FLAG (3)	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	BLANK	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (2)	0	0	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (2)	0	0	0	BLANK	BLANK	-1	-1
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (2)	0	0	0	BLANK	BLANK	0	0
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (2)	0	BLANK	0	BLANK	BLANK	1	NO WC
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	-1	0

Figure 37: Difference Between Written Codes and Scores for PT Construct.

The difference table shown in Figure 37 shows that there was near-perfect agreement between written codes and scores on the items designed to target the plate tectonics construct. The only instances where the scores diverged from the codes was in the more open-ended section 4. Additionally, there were no flags on the items that were intended to be scored for this construct. Four out of the five missed opportunities were from item 3c, the same item discussed in regards to the previous research question.

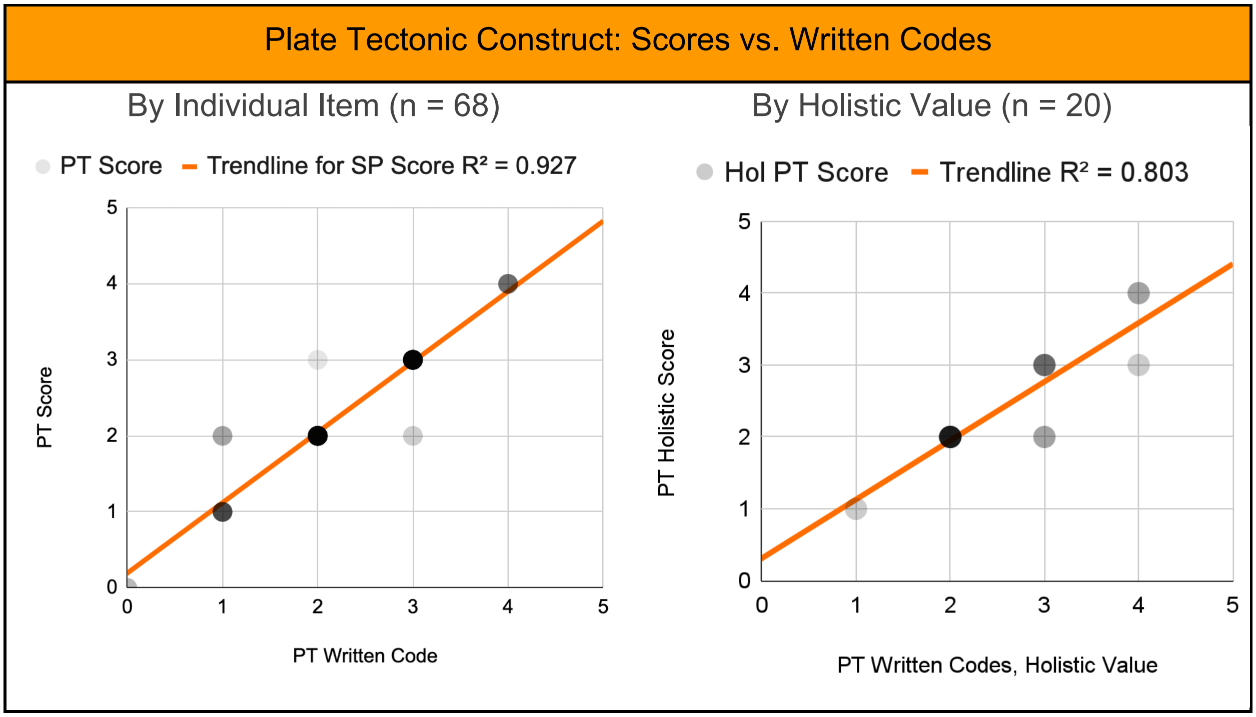
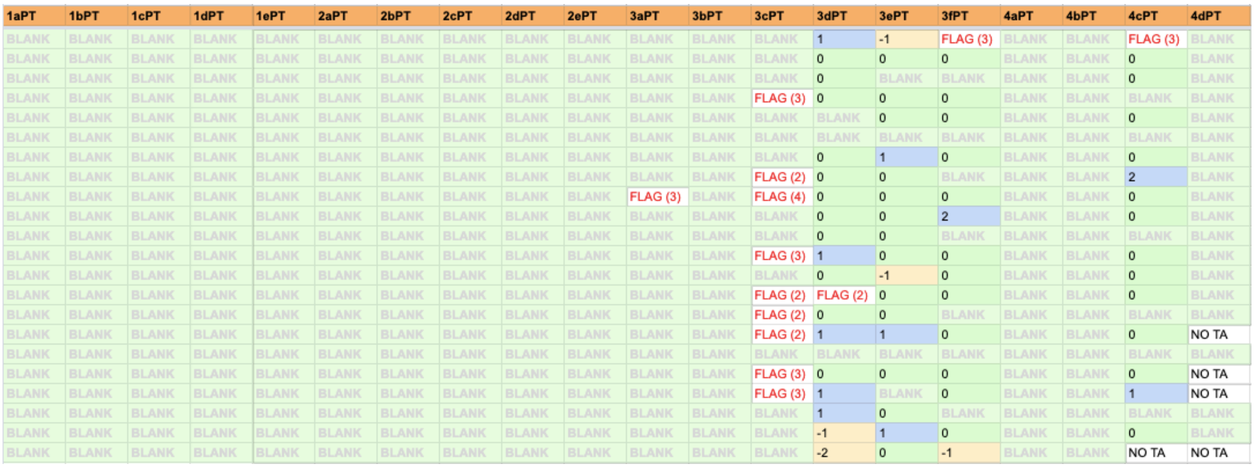


Figure 38: Linear Regressions on Individual Responses and Holistic Values Comparing PT Written Codes to Scores.

The scatter plot shown in Figure 38 shows that there was perfect agreement between written codes and holistic scores for students at level 1 and level 2 on the plate tectonics construct. Students whose written responses were coded at higher levels had holistic scores that either matched or were one point lower. This suggests a potential for underscoring via the scoring procedure and outcome space. The assessment design likely contributed to this as well because there were fewer items scored with respect to the plate tectonics construct than there were for the other three constructs**.**

5.3.c.3 PT RQ4 Comparison. This comparison pair describes the alignment between think-aloud codes and scores. There were 63 data pairs for items with both a think-aloud code and a score for this construct across the assessment for all subjects. Of those 63 pairs, 47 had matching values for both think-aloud code and score, meaning that the scores aligned with the same construct map level as their thinking 75% of the time. When the data were collapsed into holistic values, the agreement was at a similar level. For the 20 students who had holistic values for think-aloud codes, 14 of them (70%) had matching holistic scores on this construct.



1aPT	1bPT	1cPT	1dPT	1ePT	2aPT	2bPT	2cPT	2dPT	2ePT	3aPT	3bPT	3cPT	3dPT	3ePT	3fPT	4aPT	4bPT	4cPT	4dPT
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	1	-1	FLAG (3)	BLANK	BLANK	FLAG (3)	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (3)	0	0	0	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	1	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (2)	0	0	BLANK	BLANK	BLANK	2	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (3)	BLANK	FLAG (4)	0	0	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	2	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	0	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (3)	1	0	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	0	-1	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (2)	FLAG (2)	0	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (2)	0	0	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (2)	1	1	0	BLANK	BLANK	0	NO TA
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (3)	0	0	0	BLANK	BLANK	0	NO TA
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (3)	1	BLANK	0	BLANK	BLANK	1	NO TA
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	1	0	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	-1	1	0	BLANK	BLANK	0	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	-2	0	-1	BLANK	BLANK	NO TA	NO TA

Figure 39: Difference Between Think-aloud Codes and Scores for PT Construct.

This comparison has the worst "matchiness" profile of all the research questions for plate tectonics. While the majority of data pairs do agree, there is a two-point spread at every think-aloud level for individual item scores. The concentration of think-aloud codes at levels 2 and 3, which is visually obvious in Figure 40 below, show that this construct was most effective at eliciting thinking and generating matching scores in the middle of the construct map, but not at the extreme ends. The individual items shown in the difference chart all have similar profiles: most of the time there is not a difference between the think-aloud code and the score, but differences can result in either a higher or lower score with as much as a two-point disparity. With a relatively small number of data points, these instances of disagreement represent 25% of the paired measurements.

There are 13 missed opportunities for scoring student thinking on the plate tectonics construct. Nearly all of them arise from item 3c, the same one that generated flags on both research questions 1 and 2. Five of these missed opportunities are at think-aloud level 2, seven are at think-aloud level 3, and only one is at think-aloud level 4.

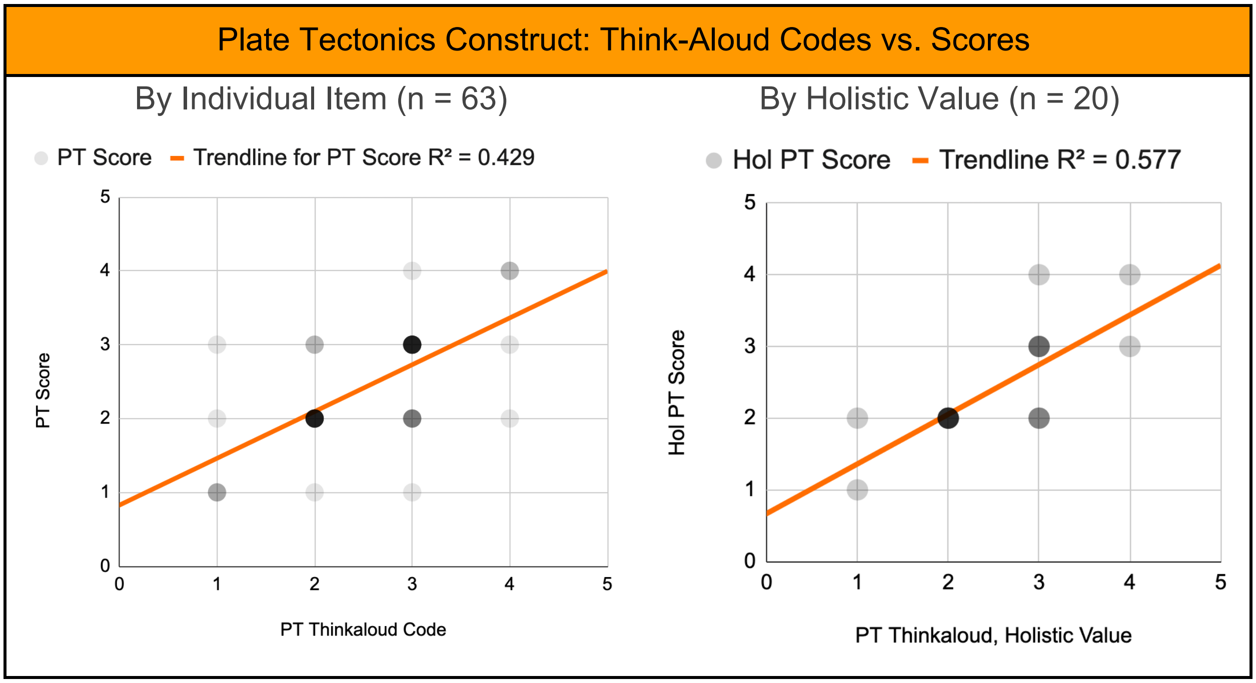


Figure 40: Linear Regressions on Individual Responses and Holistic Values Comparing PT Think-aloud Codes to Scores.

The scatter plots and trendlines for this comparison are similar for both individual items and holistic values. Both show good agreement at level 2, but decreased sensitivity to higher levels of thinking.

5.3.c.4 Summary: Plate Tectonics. The plate tectonics construct was much less successful at promoting coherence between student thinking, responses, and scores than the other constructs. Additionally, the coefficients of determination for the plate tectonics comparisons appear to be decoupled from the "matchiness" profile in a way that is distinct from the behavior of the other constructs. Based on the patterns from all four research questions, I believe there are two main reasons for this. First, the items were too few and did not prompt student thinking at level 4 of the construct. Second, one item (3c) that could have been used for this purpose was not correctly identified in the design process, and therefore was not included in the outcome space or scoring procedure for the plate tectonics construct.

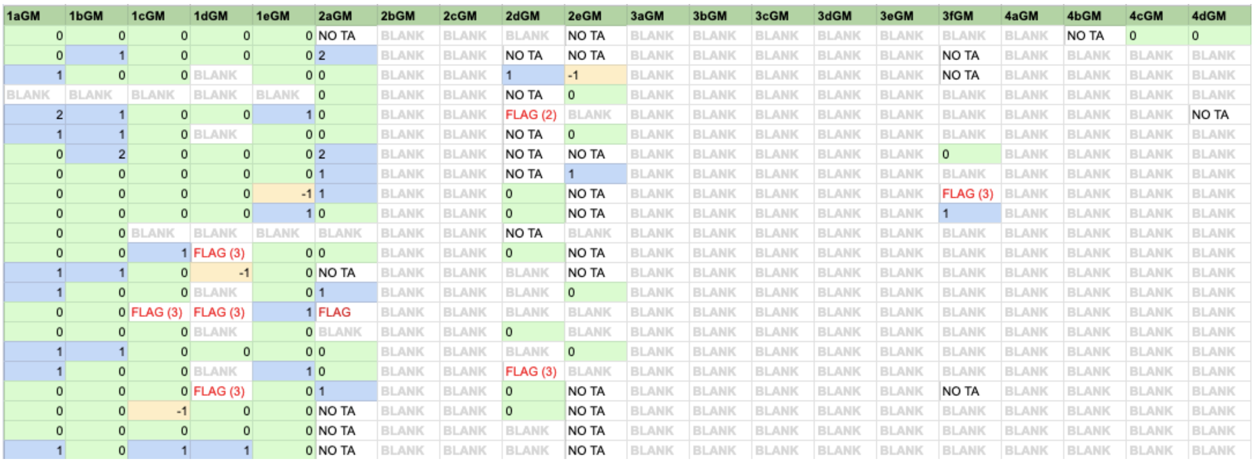
5.3.d Topographic Maps Construct

Coherence data for the Topographic Maps construct is summarized in Table 8.

Table 8: TM Coherence Data.

Pair	Comparison Method	R^2	Standard error	P value	Perfect Match	Flags
WC vs TA (RQ2)	Individual Responses (122)	0.69	0.46	0.00017	74%	7
	Holistic Values (22)	0.78	0.41	0.0051	68%	
Score vs WC (RQ3)	Individual Responses (133)	0.93	0.22	0.0077	93%	19
	Holistic Values (22)	0.86	0.38	0.58	86%	
Score vs TA (RQ4)	Individual Responses (116)	0.66	0.48	0.0030	73%	15
	Holistic Values (22)	0.75	0.433	0.011	73%	

5.3.d.1 TM RQ2 Comparison. There were 122 data pairs with codes for both the think-aloud and the written response for this construct across the assessment for all subjects. Of those 122 pairs, 90 had identical values for both think-aloud and written codes, meaning that the students' recorded responses aligned with the same construct level as their thinking 74% of the time. The agreement level was similar but slightly for the holistic comparison. All 22 subjects had holistic think-aloud codes and holistic written codes. Fourteen of them, or 64%, had a perfect match between the two.



1aGM	1bGM	1cGM	1dGM	1eGM	2aGM	2bGM	2cGM	2dGM	2eGM	3aGM	3bGM	3cGM	3dGM	3eGM	3fGM	4aGM	4bGM	4cGM	4dGM
0	0	0	0	0	NO TA	BLANK	BLANK	BLANK	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO TA	0	0
0	1	0	0	0	2	BLANK	BLANK	NO TA	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	NO TA	BLANK	BLANK	BLANK	BLANK
1	0	0	BLANK	0	0	BLANK	BLANK	1	-1	BLANK	BLANK	BLANK	BLANK	BLANK	NO TA	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	BLANK	BLANK	NO TA	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
2	1	0	0	1	0	BLANK	BLANK	FLAG (2)	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO TA
1	1	0	BLANK	0	0	BLANK	BLANK	NO TA	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	2	0	0	0	2	BLANK	BLANK	NO TA	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK
0	0	0	0	0	1	BLANK	BLANK	NO TA	1	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	0	0	-1	1	BLANK	BLANK	0	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (3)	BLANK	BLANK	BLANK	BLANK
0	0	0	0	1	0	BLANK	BLANK	0	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	1	BLANK	BLANK	BLANK	BLANK
0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	1	FLAG (3)	0	0	BLANK	BLANK	0	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
1	1	0	-1	0	NO TA	BLANK	BLANK	BLANK	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
1	0	0	BLANK	0	1	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	FLAG (3)	FLAG (3)	1	FLAG	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	0	BLANK	0	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
1	1	0	0	0	0	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
1	0	0	BLANK	1	0	BLANK	BLANK	FLAG (3)	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	0	FLAG (3)	0	1	BLANK	BLANK	0	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	NO TA	BLANK	BLANK	BLANK	BLANK
0	0	-1	0	0	NO TA	BLANK	BLANK	0	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	0	0	0	NO TA	BLANK	BLANK	BLANK	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
1	0	1	1	0	NO TA	BLANK	BLANK	BLANK	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK

Figure 41: Difference Between Think-aloud Codes and Written Codes for TM Construct.

The difference table in Figure 41 illustrates the contrast between items in part 1, which was exclusively focused on eliciting and measuring thinking about topographic maps, and items in part 2, in which topographic maps were a secondary target after surface processes. In the latter section, students frequently used topographic map skills in such a way that their written responses were able to be coded with respect to the construct, but they did not necessarily verbalize their thinking about the map or topographic profile as they were talking through their problem-solving process. Does this mean their thinking about the topographic maps was subconscious? It could also be that the act of reporting their thinking aloud prompted students to prioritize a different construct in their narrative. Nonetheless, there is less clarity about student thinking on this section of the assessment.

The greater frequency of blue cells than yellow shows that disagreements were more often due to a higher think-aloud code. This suggests that some student thinking was lost in translation to written output. There were seven missed opportunities for capturing student thinking – a relatively small number in comparison to the 122 successful opportunities. Nearly all of them occurred when the student's think-aloud was coded at level 3. There was one additional flag at think-aloud level 2, but none at the endpoint levels on the construct map.

For this construct, there was significant and relatively high correlation between student thinking and written codes on both individual items and holistic values. On holistic values, R^2 was 0.78 with a standard error of 0.41 (p = 0.0051), meaning variations in student thinking accounted for 78% of holistic values for written codes assigned to individual students. This is evidence that the design of the performance assessment items was successful at prompting students to record responses that accurately captured the nature of their thinking about this construct.

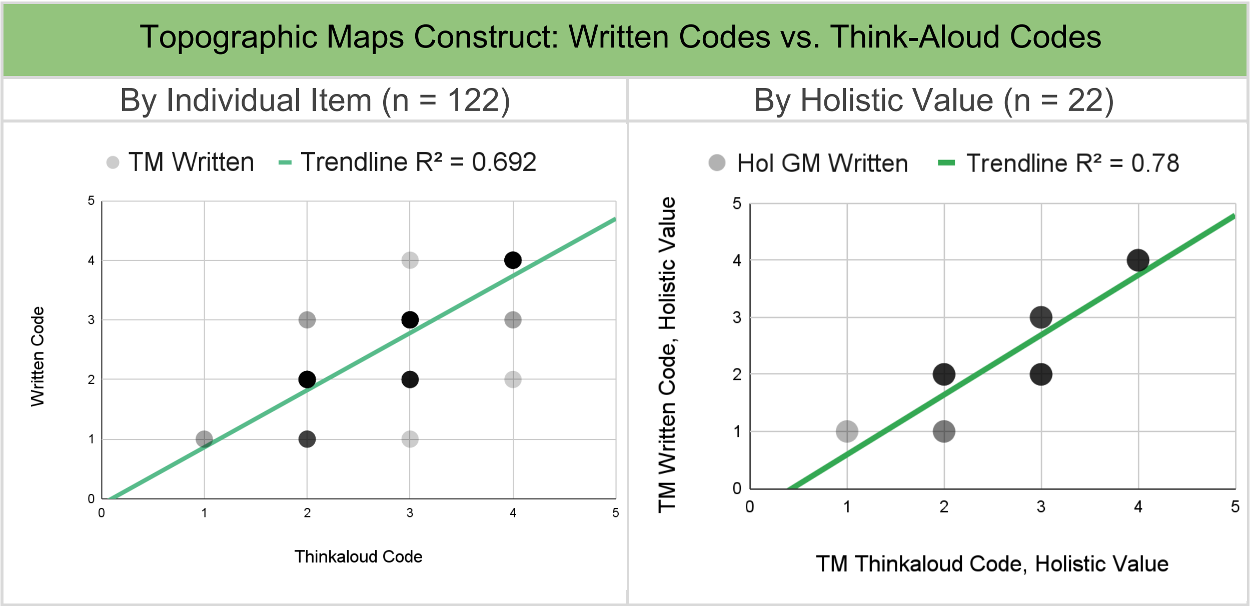


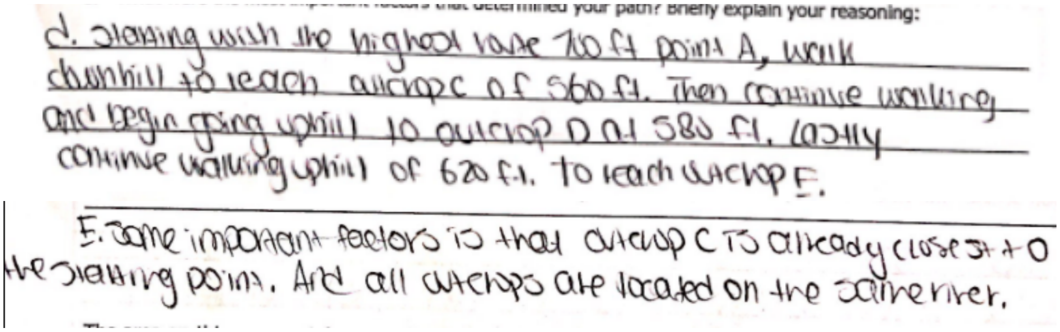
Figure 42: Linear Regressions on Individual Responses and Holistic Values Comparing TM Think-aloud Codes to Written Codes.

Figure 42 shows scatter plots for the individual response and holistic value comparisons. It is visually obvious that the holistic values have less variation at each level of thinking. When there was a discrepancy between the holistic think-aloud code value and the holistic written code value for an individual student, the written code was lower. This discrepancy occurred for 7 out of 22 subjects, five of whom exhibited thinking about topographic maps aligned with level 3 on the construct map based on their think-aloud codes, but whose written responses were assigned a level 2 holistic code. Only one student was ‘downgraded’ from level 2 to level 1 by written codes, and for students at the extremes (level 4 and level 1) there was 100% agreement between holistic think-aloud codes and holistic written codes. The scatter plot of individual responses shows the greatest amount of variation in written responses at level 3 thinking. The following response in Table 9 is from Winnie, a student who exemplifies this pattern. In this part of the assessment (1d and 1e), she was using the topographic map to plan a walking route between different locations.

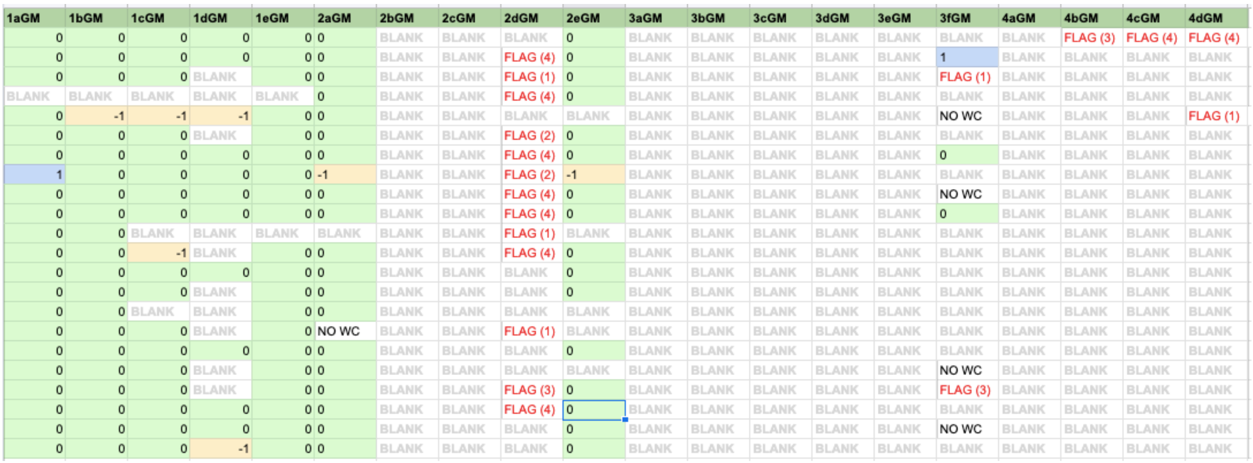
In Winnie’s thinking aloud, she made a clear comparison between different locations and routes, noting changes in elevation and steepness. As a result, her think-aloud for this part of the assessment was coded at level 3, a characterization that persisted across her think-aloud data for the topographic maps construct. In her written response, she identifies “uphill” and “downhill” along with elevation properties of specific locations. However, her written response does not include an application of the comparisons she made in her thinking, and was coded as level 2.

In addition, individual items in part 2 where the topographic maps construct was bundled with the surface processes construct failed to adequately address the topographic maps construct. Perhaps because the surface processes construct was more obviously foregrounded in the item design, many students did not verbalize ideas related to topographic maps in their thinking, or had a significant mismatch between their thinking and written response code. In every case of the latter, the think-aloud was coded at a higher level than the written response, contributing to the patterns seen in Figure 42.

Table 9: Winnie's Misaligned TM Think-Aloud and Written Response.

Think-aloud	"Here are my outcrops. [pointing to boxed locations on the map.] Oh these are the same from the pictures. So if I was standing right here [finger on point A], how would I walk. to C and D and E. Can I start at A and then go here [tracing finger down to C] and then up [tracing finger northward along river]? A is at 700 feet elevation, so it's best to not go up further. When you think about it further… when people hike, they start from the bottom. You can't start from the top. But in this case there's a road to access this high point, it says so in the question. Okay you could start from 700, go to the top one [pointing to outcrop E at top edge of map]… actually, isn't it going down? Look, when I go here, this is 700 feet at A, this is 680, this is 660 [tracing finger across the contour lines]. So since you are starting from a place high up, you could go to outcrop E and then all the way down to outcrop C and then just go home. See, outcrop C is lower than 560 in elevation. And going along the river where they are all would probably be easiest because it's not too steep here. Because the elevation is changing in a steady way. These lines are spaced out [pointing to space between contour lines between outcrop E and outcrop D] and you would go down steadily. If we went this way [pointing from C to E], the elevation is increasing. So from here, we would be going uphill. If we went that way [pointing from E to C], it's decreasing. And the other way is downhill. But if you wanted to get to the outcrops quickly maybe you should go to C first, because it's like right there."
Written Response	 determined your path? Briefly explain your reasoning: d. starting with the highest route 700 ft point A, walk downhill to reach outcrop C of 560 ft. Then continue walking and begin going uphill to outcrop D at 580 ft. lastly continue walking uphill of 620 ft. to reach outcrop E. E. Some important factors is that outcrop C is already closest to the starting point. And all outcrops are located on the same river. *d. starting with the highest route 700 ft point A, walk downhill to reach outcrop C of 560 ft. Then continue walking and begin going uphill to outcrop D and 580 ft. Lastly continue walking uphill of 620 ft. to reach outcrop E.* *e. Some important factors is that outcrop C is already closest to the starting point. And all outcrops are located on the same river.*

5.3.d.2 TM RQ3 Comparison. There were 133 data pairs with values for both the written response and the scoring procedure for the Topographic Maps construct across the assessment for all subjects. Of those 133 pairs, 124 had identical values, for a perfect agreement rate of 93%. The holistic values were only slightly less successful: across the data set of 22 students, 19 of them (86%) had holistic scores that exactly matched their holistic written codes.



1aGM	1bGM	1cGM	1dGM	1eGM	2aGM	2bGM	2cGM	2dGM	2eGM	3aGM	3bGM	3cGM	3dGM	3eGM	3fGM	4aGM	4bGM	4cGM	4dGM
0	0	0	0	0	0	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (3)	FLAG (4)	FLAG (4)
0	0	0	0	0	0	BLANK	BLANK	FLAG (4)	0	BLANK	BLANK	BLANK	BLANK	BLANK	1	BLANK	BLANK	BLANK	BLANK
0	0	0	BLANK	0	0	BLANK	BLANK	FLAG (1)	0	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (1)	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	BLANK	BLANK	FLAG (4)	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	-1	-1	-1	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO WC	BLANK	BLANK	BLANK	FLAG (1)
0	0	0	BLANK	0	0	BLANK	BLANK	FLAG (2)	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	0	0	0	0	BLANK	BLANK	FLAG (4)	0	BLANK	BLANK	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK
1	0	0	0	0	-1	BLANK	BLANK	FLAG (2)	-1	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	0	0	0	0	BLANK	BLANK	FLAG (4)	0	BLANK	BLANK	BLANK	BLANK	BLANK	NO WC	BLANK	BLANK	BLANK	BLANK
0	0	0	0	0	0	BLANK	BLANK	FLAG (4)	0	BLANK	BLANK	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK
0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (1)	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	-1	BLANK	0	0	BLANK	BLANK	FLAG (4)	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	0	0	0	0	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	0	BLANK	0	0	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	BLANK	BLANK	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	0	BLANK	0	NO WC	BLANK	BLANK	FLAG (1)	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	0	0	0	0	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	0	BLANK	0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO WC	BLANK	BLANK	BLANK	BLANK
0	0	0	BLANK	0	0	BLANK	BLANK	FLAG (3)	0	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (3)	BLANK	BLANK	BLANK	BLANK
0	0	0	0	0	0	BLANK	BLANK	FLAG (4)	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	0	0	0	0	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK	BLANK	NO WC	BLANK	BLANK	BLANK	BLANK
0	0	0	-1	0	0	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK

Figure 43: Difference Between Written Codes and Scores for TM Construct.

The difference table for this comparison seen in Figure 43 illustrates the agreement between written codes and scores. On items that targeted the topographic maps construct, there was typically only one instance of disagreement across all 22 subjects per assessment item. These infrequent mismatches tended to have a higher score than written code. This could have been due to specifications in the outcome space that clarified expectations around response to these assessment items. Three out of seven of these occurrences belong to the same student subject.

There were 19 missed opportunities for scoring this construct, 13 of which occurred on item 2d. In this item, students sketched a speculative topographic profile representing what they believed the landscape would look like in the absence of the river. I made the initial decision to score this item only with respect to the surface processes construct because I was uncertain whether students would use the features of the map or profile grid in their problem solving. However, for 13 of the 22 students there was enough evidence that they did this to code their written responses to this item with respect to both the surface processes and topographic maps constructs. Future revisions of the outcome space should include scoring criteria for the topographic maps construct.

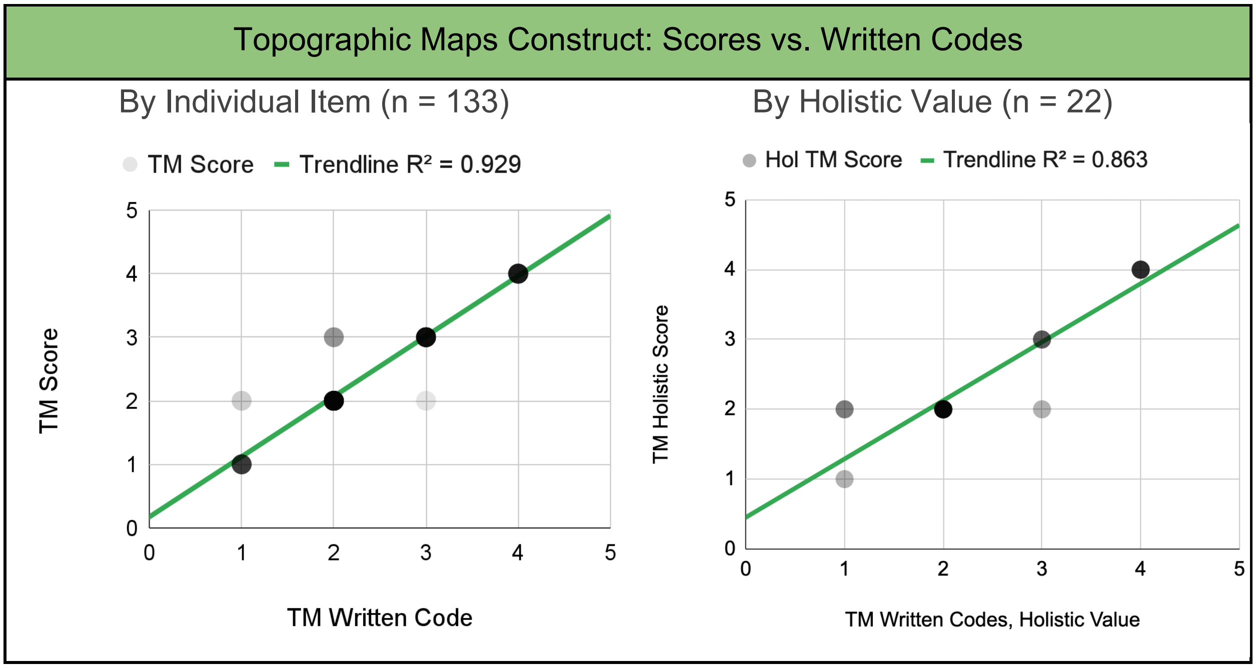
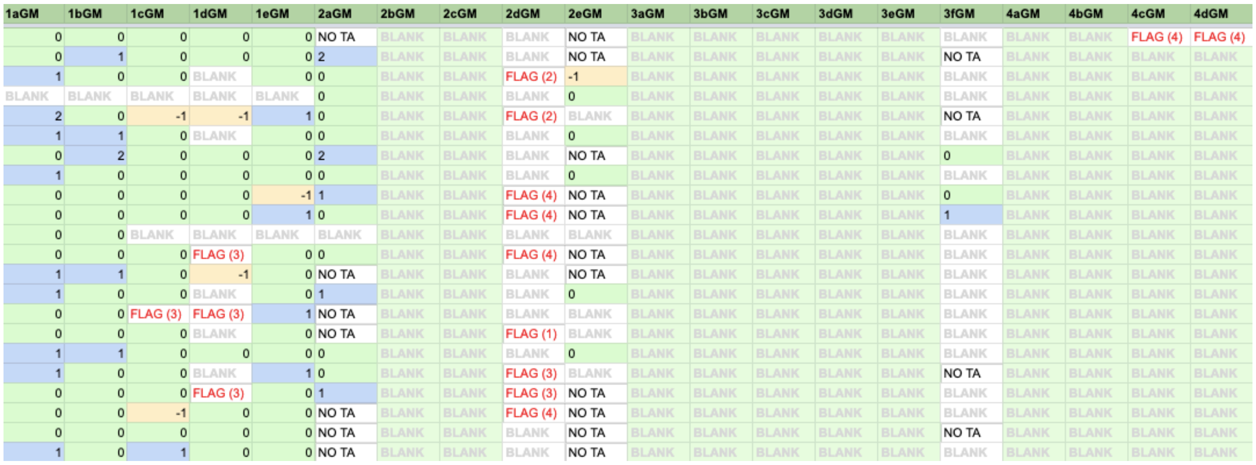


Figure 44: Linear Regressions on Individual Responses and Holistic Values Comparing TM Written Codes to Scores.

The scatter plot for individual items, seen on the left side of Figure 44, along with the mathematical comparisons above, indicate that there was a high degree of coherence between student responses and the resulting scores for the topographic maps construct on an individual item level. This tells me that the outcome space was well defined and accurately represented the general model of cognition for topographic maps from my construct map.

There is a subtle change in the pattern of score distributions that occurred in the transition from individual items to holistic scores for this construct. On individual items, there are a handful of instances where students were assigned a written code one level lower than the score they ultimately received on the same item (corresponding to the yellow cells in the difference table in Figure 43 above). This occurred at the lower end of the construct levels. However, when holistic scores were generated, level 2 became dominant and perhaps over-represented. The frequency pattern at written code level 1 is reversed. This warrants a revisitation of both the written prompts and outcome space to see if these components of the assessment system are artificially increasing the amount of level 2 scores received by students.

5.3.d.3 TM RQ4 Comparison. For the topographic maps construct, R^2 was higher for holistic values than for individual responses, both of which were statistically significant comparisons of think-aloud codes to scores. On individual responses, approximately 65% of variation in student scores was determined by variations in their thinking. On holistic values, the coefficient of determination increased to 75%. In both modalities, the rate of perfect matching between think-aloud code and score values was 73%. There were 116 data pairs for individual responses, 85 of which had identical values. Sixteen out of 22 students had the same holistic score as their holistic think-aloud code value.



1aGM	1bGM	1cGM	1dGM	1eGM	2aGM	2bGM	2cGM	2dGM	2eGM	3aGM	3bGM	3cGM	3dGM	3eGM	3fGM	4aGM	4bGM	4cGM	4dGM
0	0	0	0	0	NO TA	BLANK	BLANK	BLANK	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	FLAG (4)	FLAG (4)
0	1	0	0	0	2	BLANK	BLANK	BLANK	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	NO TA	BLANK	BLANK	BLANK	BLANK
1	0	0	BLANK	0	0	BLANK	BLANK	FLAG (2)	-1	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
BLANK	BLANK	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
2	0	-1	-1	1	0	BLANK	BLANK	FLAG (2)	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO TA	BLANK	BLANK	BLANK	BLANK
1	1	0	BLANK	0	0	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	2	0	0	0	2	BLANK	BLANK	BLANK	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK
1	0	0	0	0	0	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	0	0	-1	1	BLANK	BLANK	FLAG (4)	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK
0	0	0	0	1	0	BLANK	BLANK	FLAG (4)	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	1	BLANK	BLANK	BLANK	BLANK
0	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	0	FLAG (3)	0	0	BLANK	BLANK	FLAG (4)	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
1	1	0	-1	0	NO TA	BLANK	BLANK	BLANK	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
1	0	0	BLANK	0	1	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	FLAG (3)	FLAG (3)	1	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	0	BLANK	0	NO TA	BLANK	BLANK	FLAG (1)	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
1	1	0	0	0	0	BLANK	BLANK	BLANK	0	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
1	0	0	BLANK	1	0	BLANK	BLANK	FLAG (3)	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	NO TA	BLANK	BLANK	BLANK	BLANK
0	0	0	FLAG (3)	0	1	BLANK	BLANK	FLAG (3)	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	-1	0	0	NO TA	BLANK	BLANK	FLAG (4)	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK
0	0	0	0	0	NO TA	BLANK	BLANK	BLANK	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	NO TA	BLANK	BLANK	BLANK	BLANK
1	0	1	0	0	NO TA	BLANK	BLANK	BLANK	NO TA	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK

Figure 45: Difference Between Think-aloud Codes and Scores for TM Construct.

The difference table in Figure 45 above shows that the items in part 1 were most successful at generating scores that exactly matched the level of thinking students engaged in while completing the performance task components. Item 1c in particular had a high degree of agreement. When there was a mismatch, the think-aloud code was usually higher, suggesting that some information about student thinking is lost in the assessment process. This is consistent with the results from the first comparison of think-aloud to written codes.

There were 15 flags indicating missed opportunities for converting an observation of student thinking into a quantitative score for the topographic maps construct. The majority of these occurred at think-aloud levels 3 and 4, with six instances of each.

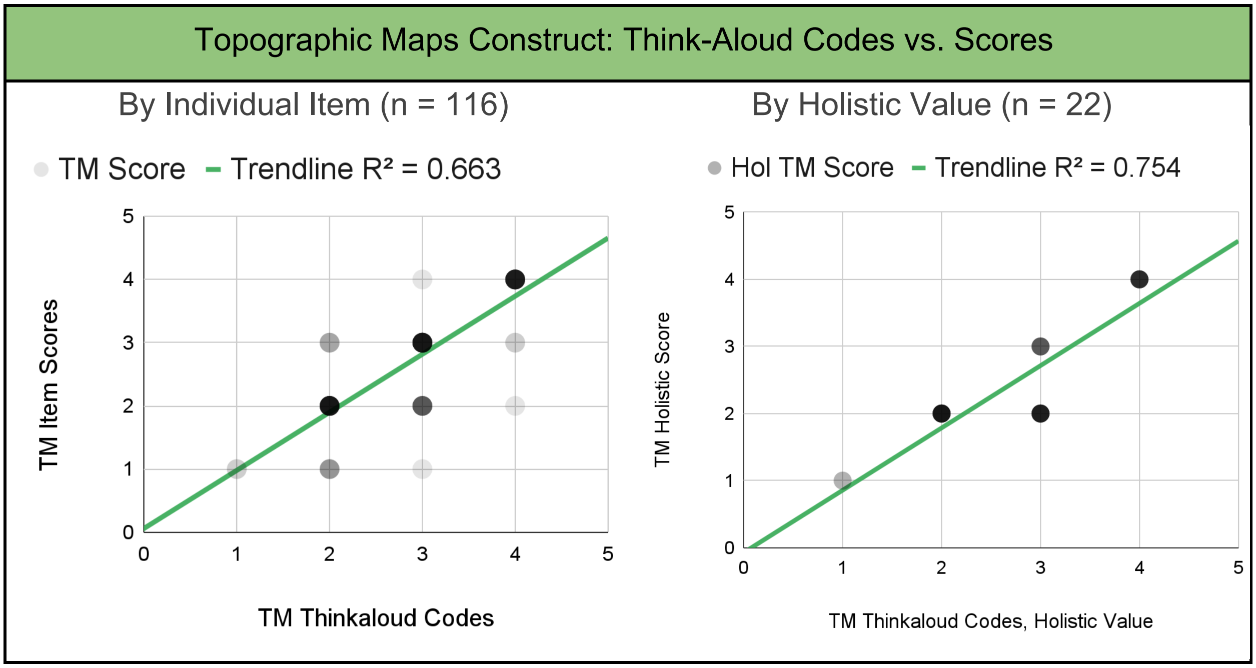


Figure 46: Linear Regressions on Individual Responses and Holistic Values Comparing TM Written Codes to Scores.

Figure 46 shows that there was the greatest degree of agreement between think-aloud codes and scores at level 1. In the holistic modality, there was also perfect agreement between think-aloud codes and scores at levels 2 and 4. On individual items, there was also strong agreement at level 4: out of 24 instances in which a student's think-aloud was coded at level 4, 21 of those items (87.5%) were also scored at level 4. For this construct, thinking at either the high or low extreme–as verbalized by students while they worked on the assessment task–was reliably captured by the scoring procedure. The “muddled middle” for the topographic maps construct indicates an assessment design flaw that could be improved on future iterations. I believe the source of this flaw is mainly in item design, less so in the outcome space or scoring procedure. As shown in the findings for research questions 2 and 3, above, the items successfully elicited student *thinking* aligned with all four levels of the construct, but did not successfully prompt students to record written responses that provided evidence of that thinking. In particular, the individual items did not successfully differentiate between students at levels 2 and 3.

5.3.d.4 Summary: Topographic Maps. For the topographic maps construct, there appears to be an effect of the coding and scoring that leads to a concentration of results at level 2. I believe this is an artificial effect that does not fully represent the range of student thinking, since the written codes and scores were highly correlated. To improve the coherence of this construct, I would revise the outcome space and potentially the item prompts using think-aloud data for guidance. When students reported thinking that clearly aligned with higher levels of the construct map, was there evidence of those thinking patterns captured in the written response? If yes, what can be added to the outcome space at the appropriate level? If no, what sub-prompts or structures can be added to the assessment item to ensure students have the opportunity to fully record the thinking that is relevant to this construct? This process would apply to all the items targeting the topographic maps construct because there was a consistent pattern of think-aloud codes trending higher than written codes and scores. Additionally, item 2d generated many flags for lost opportunities, raising the question of whether it would have been preferable to score this item with respect to the topographic maps construct. There were more flags in the written code to score comparison (research question 3) than in the think-aloud to score comparison (research question 4), which makes me suspect that my initial assumption that students might not think consciously about the map was partially correct. This warrants further examination.

5.4 Summary of Research Questions 2, 3, and 4

5.4.a Research Question 2: To what extent do the students' recorded responses to the assessment task correlate with their thinking during the task?

This question was answered via a simple linear regression analysis comparing **c**odes on think-alouds to codes on written responses, allowing me to ascertain the extent to which student thinking determined the response they wrote down in answer to the assessment task.

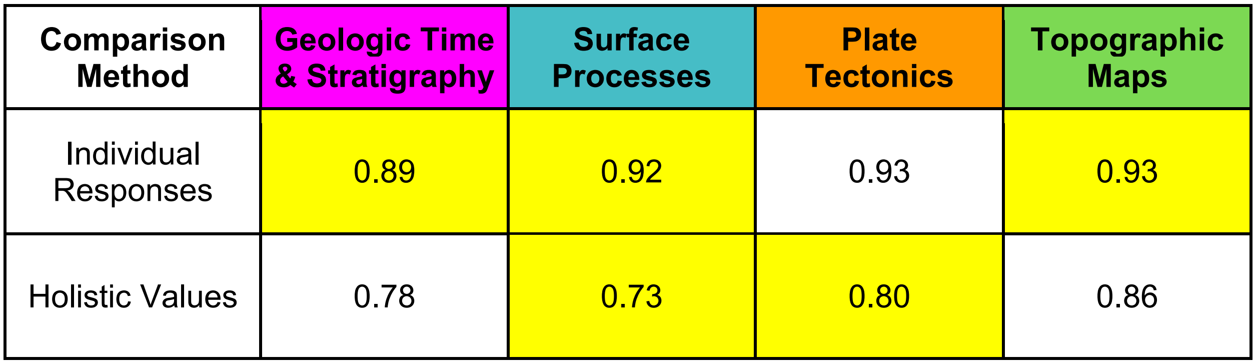
Comparison Method	Geologic Time & Stratigraphy	Surface Processes	Plate Tectonics	Topographic Maps
Individual Responses	0.60	0.57	0.46	0.69
Holistic Values	0.74	0.59	0.65	0.78

Figure 47: Coefficients of Determination for RQ2 Comparisons Across Constructs. Cells highlighted in yellow indicate statistically significant comparisons ($p < 0.05$).

Three out of the four content-based constructs showed significant correlation between student thinking, as captured by the think-aloud codes, and student written responses to the performance assessment tasks, as captured by the written codes. The exception, the plate tectonics construct, is the subject of further discussion in Chapter 6. For the remaining three constructs, R^2 ranged from 0.57 (geologic time & stratigraphy) to 0.69 (topographic maps) on individual responses, meaning that up to 69% of variation in the level of students' written responses can be accounted for by differences in the nature of student thinking. For the holistic values, which were reported as one number per construct per student, R^2 from a linear fit ranged from 0.48 (surface processes) to 0.78 (topographic maps). Across all four constructs, the holistic values had improved coherence between thinking and written responses in comparison to individual items. Out of the 84 total data pairs of holistic values, the holistic written code was identical to the holistic think-aloud code 70% of the time.

5.4.b Research Question 3: To what extent do the results of the scoring procedure correlate with the students' recorded responses?

This question was answered via a simple linear regression analysis comparing codes on written responses to scores generated via scoring procedure, allowing me to ascertain the extent to which written responses on the assessment task determined the score students were assigned by graders following a normed scoring procedure. I calculated p values for each result using a two tailed t-test. Values for R^2 across constructs and comparison methods are summarized in Figure 48.



Comparison Method	Geologic Time & Stratigraphy	Surface Processes	Plate Tectonics	Topographic Maps
Individual Responses	0.89	0.92	0.93	0.93
Holistic Values	0.78	0.73	0.80	0.86

Figure 48: Coefficients of Determination for RQ3 Comparisons Across Constructs. Cells highlighted in yellow indicate statistically significant comparisons ($p < 0.05$).

For these data sources, all four of the content-based constructs showed at least one significant correlation between student written responses, as captured by the codes on their written work, and scores assigned by graders. The plate tectonics construct once again exhibits some unusual behavior. For the remaining three constructs, R^2 ranged from 0.81 (surface processes) to 0.93 (topographic maps) on individual responses, meaning that the variation in scores assigned to single items can largely--up to 93%-- be accounted for by the variations in the student responses. It is reassuring, if unsurprising, that there is a high level of correlation since both were based on the written work produced by students. For the holistic values, which were reported as one number per construct per student, R^2 from a linear fit ranged from 0.73 (surface processes) to 0.80 (plate tectonics). Across all four constructs, there was a slightly lower coefficient of determination for holistic values than for individual items.

5.4.c Research Question 4: To what extent do scores represent the range of student thinking?

This question was answered via a simple linear regression analysis comparing codes on think-alouds to scores generated via scoring procedure, allowing me to ascertain the extent to which variations in student thinking elicited by the assessment task determined the scores they were assigned. I used the same three different modalities to compare the data as in research questions 2 and 3: individual responses to single items and holistic scores generated per student. Each of these three comparison methods for the linear regression was used to analyze each construct separately. I calculated p values for each result using a two tailed t-test. Values for R^2 across constructs and comparison methods are summarized in Figure 49.

Comparison Method	Geologic Time & Stratigraphy	Surface Processes	Plate Tectonics	Topographic Maps
Individual Responses	0.61	0.65	0.43	0.66
Holistic Values	0.81	0.74	0.58	0.75

Figure 49: Coefficients of Determination for RQ4 Comparisons Across Constructs. Cells highlighted in yellow indicate statistically significant comparisons ($p < 0.05$).

Of the 84 data pairs for holistic think-alouds and scores, 62 of them or approximately 74% were perfectly aligned. Along with the coefficients of determination for holistic values, this demonstrates that the assessment system succeeded in promoting partial, if not total, coherence between student thinking and resultant scores. For example, an R^2 value of 0.81, as in the case of the geologic time and stratigraphy construct, implies that the complete assessment process--from prompting student thinking in response to individual items to generating a single score for each construct--was successful in capturing about 80% of the variation in student thinking across this group of students. It is both interesting and important that the R^2 for correlations between holistic think-aloud codes and scores was comparable to, or higher than, the comparisons on individual responses. The holistic values captured a similar or better level of correlation than individual items did, even though they are the result of an additional, and substantial, data reduction step. This implies that the scoring procedure, including the final step of using item responses to generate a holistic score for each construct, played an important role in preserving the overall "snapshot" of student thinking across constructs. I will explore this idea further in chapter 6.

5.5 Additional Findings

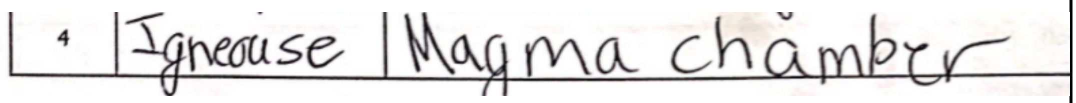
5.5.a Language Skills

Beyond the research questions, I observed a few additional patterns that were of particular interest to me. I noted how obstructive language skills can be, not just for the expected demographics (English language learners, bilingual students, students with special needs), but for many general education students. This builds on, e.g., the findings of O'Reilly and McNamara (2007), who demonstrated that performance on science assessment is mediated by literacy skills. We might expect science performance assessment to alleviate this problem, because it allows students to express their understanding in ways other than written language, such as diagrams, graphs, and map annotations. However, I observed numerous instances where a student's uncertainty about a word–e.g., how to pronounce it (when reading assessment prompts), how to spell it, or a specific vocabulary term–interfered with their work on the performance assessment. In the examples below I am including a portion of the student's think-aloud along with an excerpt of their written response on the relevant item, so that we may see what effect, if any, the student's linguistic uncertainty had on the response they committed to on paper.

In the following examples, students were distracted by their confusion over words while they thought about the performance assessment tasks. In cases like this, the students' written responses generally did capture their expressed thinking that was relevant to the assessment item in question, while leaving out their thinking related to words and pronunciations.

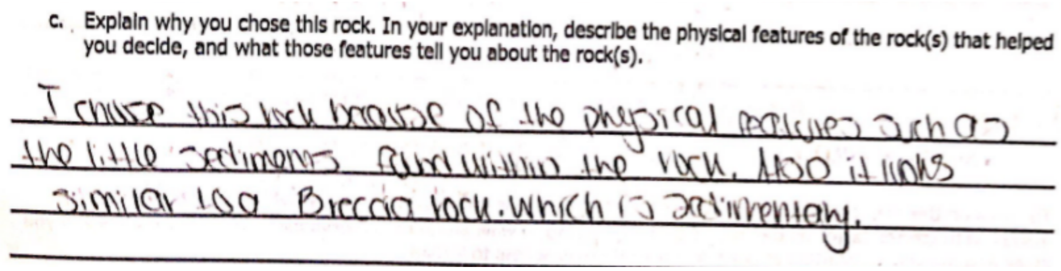
Debora was looking at rock samples from the local outcrops to make inferences about the past environment. She was using the igneous rocks chart on the Earth Science Reference Tables to determine the formation environment of an igneous rock sample.

Table 10: Debora's Language Struggles

Think-aloud	"okay I know the colors. It's black so it's over here [pointing to right-hand side of the color / composition / density arrows]. It's either andesite or … Deodorant? [points to 'DIORITE' on the chart] I was trying to say dorito! Dorito-rite! They need to be more relaxed with this stuff. It's too sciencey. Where would we find this? In a magma chamber? That's a thing. It sounds like a hot place. That could be a good name for a sauna."
Written Response	 4 Igneouse Magma chămber

Winnie was choosing a rock that she believed would have been transported by a river. She was using the Earth Science Reference Tables to find information about different rocks, and comparing it to the physical samples she had in her hands. The printed rock name "Breccia" in the *Characteristics of Sedimentary Rocks* table was a visual distractor to her.

Table 11: Winnie's Language Struggles

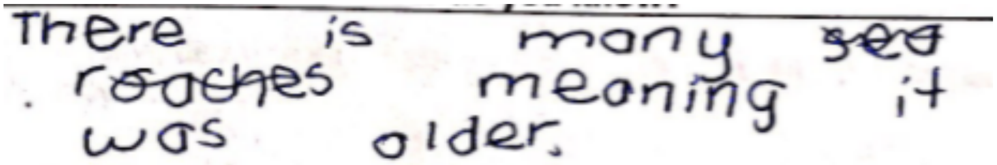
Think-aloud	"I'm looking at rock 3 because I already thought about that one. Like just little parts of sentiments. That one is the. . . How do you pronounce it? It reminds me of Russia? Berisha?"
Written Response	 c. Explain why you chose this rock. In your explanation, describe the physical features of the rock(s) that helped you decide, and what those features tell you about the rock(s). I chose this rock because of the physical features such as the little sediments found within the rock. Also it looks similar to a Breccia rock. Which is sedimentary. *I chose this rock because of the physical features such as the little sediments found within the rock. Also it looks similar to a Breccia rock. Which is sedimentary.*

Winnie's written response includes a correct usage of the word "sediments," even though she said "sentiments" when thinking aloud. She seems to know what sediments are ("little parts"). It is unclear whether she knows the word, since she could have copied it from reference tables. Her focus on the composition of the rock, rather its shape and texture, does help us fit her response within the cognitive model for the surface processes construct. It is notable that her thinking about the rock sample, and seemingly the assessment prompt, is abruptly cut off when she begins wondering how to pronounce the word Breccia.

In the examples that follow, the students' written responses *do not* fully capture the thinking they reported verbally during the think-aloud, after an instance of language-induced uncertainty.

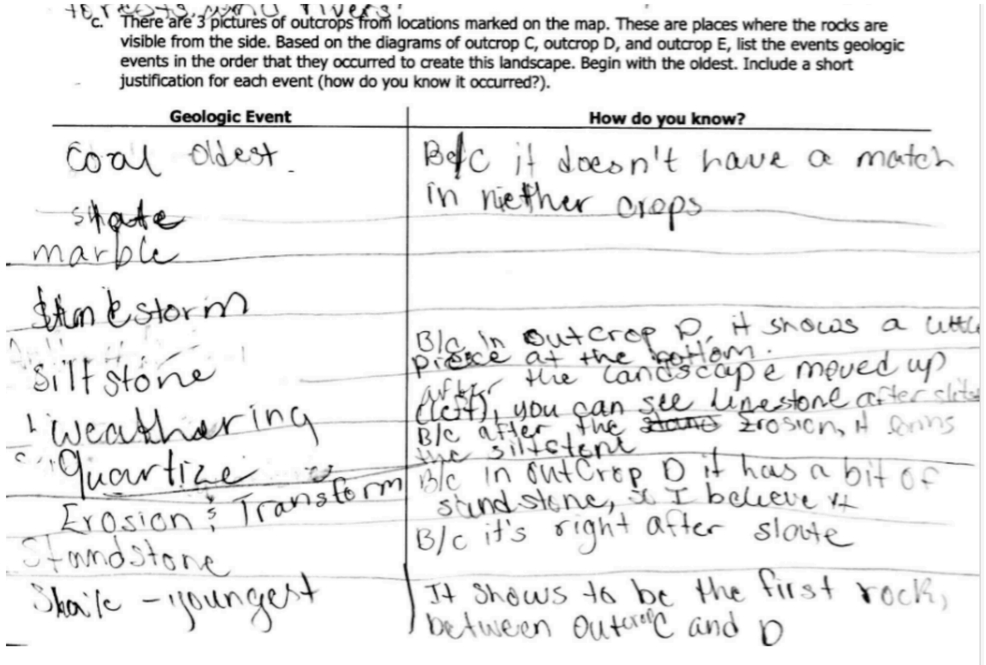
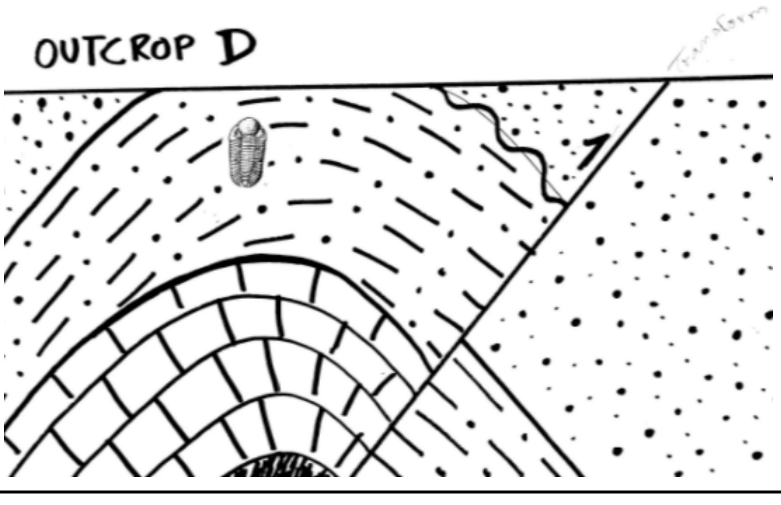
Dorothy was looking at the outcrop cross sections and trying to determine the earliest geologic event in the region. In her written response, seen in Table 12, Dorothy crossed out "sea roaches" but did not replace it with anything. It is unclear from what she wrote that she knows there are trilobite fossils in these rock layers.

Table 12: Dorothy's Language Struggles

Think-aloud	"I'm looking at the fossils and the shape of the rocks. I don't know exactly what they are. But the dot dot dot [pointing to sandstone symbol] I'm looking at that, and I'm looking at the fossils in there [pointing to Phacops illustration]. I think they're called trilobites. But I wrote sea roaches. Let me leave it, I don't know how to spell that."
Written Response	 There is many sea roaches meaning it was older.

Lucy was using evidence from the cross section diagrams to reconstruct a sequence of geologic events that shaped the region. The handwritten additions to outcrop D, shown here, were added by the student. This is, in fact, a reverse fault, in which the rocks were displaced vertically, not transform movement or a transverse fault, in which rock movement is predominantly horizontal. Based on her physical gestures, which reveal her spatial thinking, it seems that Lucy is conceptualizing the geologic movement correctly, but has named it inaccurately. In her written response, she writes that "the landscape moved up (left)" in her justification for a different geologic event. The construction of this assessment item did not allow Lucy to express her understanding of the large-scale rock movement. A better approach may have been to prompt students to annotate the outcrop diagrams with labels and symbols first, rather than relying on students to transfer their spatial thought directly into a written description.

Table 13: Lucy's Language Struggles

Think-aloud	Lucy: So, and then this is the, um, the line. Yeah. Where it's going, splitting. [points to diagonal fault line on outcrop D.] It shows, I don't know what this is. I forgot what this is called. Rabi: What are you thinking? Lucy: I'm thinking like, cause it showed in the arrow [points to half-arrow] how much you [points to left-side rock layers] like, moved up. So I guess that's the land splitting in a way that's trans... Transform? Transforming? [Lucy is holding her hands in the air next to each other, moving one upward toward the ceiling.] Like this is how the rocks would be moving.
Written Response	 terrests, and rivers: c. There are 3 pictures of outcrops from locations marked on the map. These are places where the rocks are visible from the side. Based on the diagrams of outcrop C, outcrop D, and outcrop E, list the events geologic events in the order that they occurred to create this landscape. Begin with the oldest. Include a short justification for each event (how do you know it occurred?). **Geologic Event** — **How do you know?** Coal Oldest — B/c it doesn't have a match in niether crops slate marble Limestorm silt stone — B/c in outcrop D, it shows a little piece at the bottom. weathering — after the landscape moved up (left), you can see Linestone after slate Quartize — B/c after the ~~stone~~ Erosion, it forms the siltstone Erosion ; Transform — B/c in Outcrop D it has a bit of standstone, so I believe it Standstone — B/c it's right after slate Shale - youngest — It shows to be the first rock, between Outcrop C and D
Assessment Material Annotation	

These examples are representative of an overall pattern I saw in the student think-alouds. Some other common fluency interruptions were caused by the words “topographic,” “tectonic,” and “extinct,” in addition to the more obscure rock names. This shows that when we design cognitively-based assessments, we need to make the role of language explicit in our cognitive models. If students can successfully think about and understand science independent of their knowledge and use of specific vocabulary, our cognitive models (construct maps) need to clearly describe the nature of that thinking and understanding. On the other hand, if correct use of vocabulary or science-specific language is a necessary feature of successful understanding within that construct, our cognitive model must name it as a criterion. When our cognitive model allows for students to express different *levels* of understanding about things like movement, spatial or temporal relationships, or physical transformations, we need to decide if there is a level at which students should be able to command the vocabulary that describes these things the way we might expect a geologist to do. These key distinctions will allow us to design cognitively-based assessment items that prompt for an appropriate mode of response.

There is a similar and related need to be specific in the outcome space. Especially in a cognitively-based performance assessment, we should be looking for evidence of student thinking in these different modalities. It is important for the outcome space to clearly name what that evidence might look like.

This study reinforces the idea that language and literacy skills interact strongly with assessment outcomes. When prompting thinking and responses, giving the opportunity for students to respond in modes other than writing may allow for a more complete capture of their knowledge and understanding. However, the use of domain-specific language is an important aspect of science practice. These dual aspects of the role of language in assessment are important for both developers and classroom teachers to consider.

5.5.b Student Interactions with Assessment Materials

I observed students' physical interactions with the assessment materials during the think-aloud protocol. I noted several trends in the ways students used the assessment components during their reasoning, summarized in Table 14. These patterns of behavior illustrate the ways in which science performance assessment is able to elicit different types of thinking than traditional assessments.

Table 14: Ways in Which Students Interacted with Assessment Materials

Description of Interaction	# (%) of Subjects
Picking up rock samples	22 (100%)
Proximal comparison of rock sample to another object	20 (91%)
Qualitative observation of rock texture via touch	17 (77%)
Quantitative estimation of rock grain size	2 (9%)
Rotating map	18 (82%)
Finger walking / tracing on map	16 (73%)
Physical correlation of outcrop diagrams	14 (64%)
Positioning outcrop diagrams to match map orientation	5 (23%)

5.5.b.1 Rock Samples. The students' physical interactions with the rock hand samples shows how important access to real-world samples is for geology thinking. Without fail, every student picked up the rocks to observe them. This is not surprising, but it shows that there is an important difference between using photographs or illustrations and using real geological objects. The students' purpose in examining the rocks fell into two categories: describing the composition and identity of the rock, or making connections that allowed them to mentally situate the rocks in a larger context.

Students engaged in the first kind of thinking across all levels of achievement. Those who were assigned low scores on the relevant constructs of geologic time & stratigraphy or surface processes were equally likely to talk about their tactile observations of the rocks' texture, grain size, or composition as students who were assigned higher scores. In the following example, Pluto was able to make a mental comparison to a known texture even though he didn't know how to describe it in words: "this feels very smooth. And like, not like cement like because cement feels like, I don't know how to describe it. But [the rock] feels soft. It must be made of dirt." T-Noona, whose written work was scored at a higher level, also compared the texture of a rock to a familiar material:

> The texture of this rock is softer compared to the other one. That one, you can feel small rough grains on it, it really feels just like sandpaper, that's how I know it's sandstone. But this one doesn't feel like sand. It's softer because the particles must be smaller. I'm looking at the reference tables and the smaller sediments are silt and clay. This does feel soft like clay so I'm thinking it could be a shale rock.

This shows that physical, real-life observations are an important cognitive entry point for students who are thinking about geology.

A second purpose behind students' manipulations of the rock samples was to facilitate their understanding of how the rock samples were connected to the larger context of the assessment scenario. This commonly took the form of placing the rocks on top of outcrop diagrams, maps, or both. The majority of students (91%) used physical proximity to compare the rocks to each other, map symbols on outcrop diagrams, or outcrop locations on the map.

5.5.b.2 Topographic Map and Outcrop Diagrams. Unlike in a traditional written assessment, in which a map would be integrated into the rest of the printed assessment materials, the map in this performance assessment was provided on a separate page. I suspect that treating the map as an object influenced the way students used it. A large majority (82%) of the students rotated the map to help visualize a viewpoint from a specific perspective. Interpreting overhead map data in the context of a specific ground-based location is a type of thinking specified in level 3 of the topographic maps construct, so providing the topographic map in a format that makes this more natural is an important feature that enhances the performance-based aspect of this assessment.

Another way in which students reasoned with the map was to use their fingers as a representation of themselves when planning or imagining movement across the landscape. Most of the students (73%) either traced a pathway with their finger or used their fingers to “walk” across the contour lines. As they did so, they frequently thought about how the elevation would be changing over that horizontal distance. For example, Tim Tam contrasted a region on the map with closely spaced contour lines to a longer route through a region where the lines were farther apart. As she did so, she used her fingers to show how she believed the difference in topography would affect her journey:

> If we’re walking down all the way, you’re not speeding like this. If you’re going down this side, you’re running like this [pantomimes running with fingers on the map]. So over here you’re just making a peaceful walk down the mountain. But from this 620 to outcrop C, you’re going to be speeding. The steeper it is, the faster it has to go.

The outcrop drawings in this assessment were also provided to students as separate materials, rather than as diagrams in line with the assessment prompts. As a result, students were able to move the cross sections to align similar symbols, or to orient them in the same spatial arrangement that was indicated by the map. For example, when determining the order of geologic events in this region, Trynessa correlated two outcrop diagrams using the limestone map symbol present in each cross section:

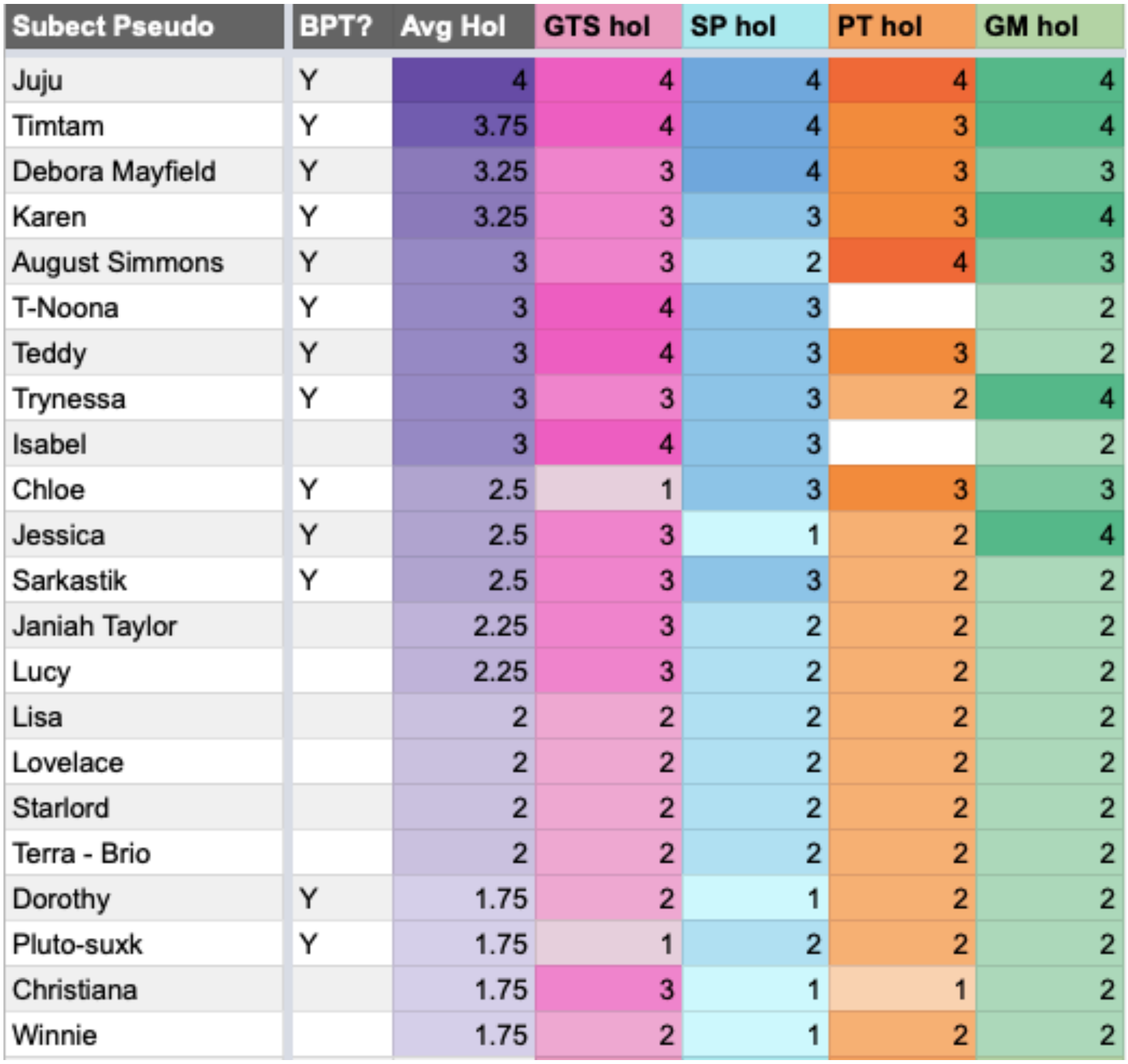
> I'm going to need to put the outcrops in order. I can match them up. [slides outcrops C and D next to each other to make the same rock symbols align] So I know this symbol, it's limestone. And because these outcrops match up, I know that this limestone (outcrop C) is the same age as this limestone (outcrop D).

Students who moved the outcrop diagrams in this way were much more likely to receive holistic scores of 3 or 4 for the geologic time & stratigraphy construct.

Students' ability to move, manipulate, and reorganize the assessment materials supported their thinking in a variety of ways. They were able to reason via spatial thinking, draw connections between different components of the assessment, and make authentic observations of real-world objects. These are ways that performance assessment can help maintain the desired alignment between domain-specific cognition patterns and classroom learning.

5.5.c "Big Picture" Thinking

One of the design features of this performance assessment was that it was centered around a single unifying scenario. In this case, students were making sense of the past, present, and future of a unique (though hypothetical) geologic setting, bringing to bear their knowledge of separate but related Earth Science constructs. Students made sense of the scenario as a whole, or thought about the "big picture," to varying degrees. In general, those who used big picture thinking were more successful on the assessment, meaning likely to receive higher scores overall. These students verbalized connections in their thinking between different parts of the assessment. Others considered each section in isolation, and did not express an understanding that all parts of the assessment were situated in the same physical location, or connected via cause-and-effect relationships. Students in the former group tended to perform better on the assessment than those in the latter group, but there were a few exceptions to this pattern on either side.



Subect Pseudo	BPT?	Avg Hol	GTS hol	SP hol	PT hol	GM hol
Juju	Y	4	4	4	4	4
Timtam	Y	3.75	4	4	3	4
Debora Mayfield	Y	3.25	3	4	3	3
Karen	Y	3.25	3	3	3	4
August Simmons	Y	3	3	2	4	3
T-Noona	Y	3	4	3		2
Teddy	Y	3	4	3	3	2
Trynessa	Y	3	3	3	2	4
Isabel		3	4	3		2
Chloe	Y	2.5	1	3	3	3
Jessica	Y	2.5	3	1	2	4
Sarkastik	Y	2.5	3	3	2	2
Janiah Taylor		2.25	3	2	2	2
Lucy		2.25	3	2	2	2
Lisa		2	2	2	2	2
Lovelace		2	2	2	2	2
Starlord		2	2	2	2	2
Terra - Brio		2	2	2	2	2
Dorothy	Y	1.75	2	1	2	2
Pluto-suxk	Y	1.75	1	2	2	2
Christiana		1.75	3	1	1	2
Winnie		1.75	2	1	2	2

Figure 50: Big Picture Thinking Compared to Average Holistic Scores

Note: The figure above shows averaged holistic scores for all 22 subjects. The holistic scores for each construct shown in the pink, blue, orange and green columns have been averaged to generate an overall score (purple column). This is not part of the scoring procedure, but is done here for the purpose of examining the relationship between scores and "big picture thinking." Colors in these columns are formatted with a gradient to make higher scores more visually apparent. The column labeled "BPT?" is labeled "Y" (yes) for students who expressed big picture thinking in their think-aloud.

Of the 22 subjects in my study, 13 (59%) of them described "big picture" thinking. I assigned this label to students based on the entirety of their think-aloud data. Key indicators included instances where students talked about multiple parts of the assessment in terms of a single unified context, made reference to evidence from different sections of the assessment when responding to a single prompt, or described an "aha!" moment when they saw the big picture. I speculate that this is in part due to the nature of the assessment, and is confirmation that this design principle was effectively enacted.

Students who used big picture thinking tended to have higher scores overall. In this group of subjects, an average holistic score of 2.5 or higher separated those who were likely to have made sense of the big picture from those who did not. An average holistic score of 3 or higher indicates that the student's overall thinking and understanding about these geology constructs has reached the level of sophistication and scientific accuracy prescribed by the New York State board of Regents for these topics; a score of 2.5 would mean the student is approaching, but has not yet reached, this target. Eleven out of the 12 students (92%) whose average holistic scores were 2.5 or greater used big picture thinking. This is consistent with my framing of the construct map levels, where level 3 was intended to describe application and analysis modes of thinking

(Anderson & Krathwohl, 2001). When students use these thinking processes, they are more apt to make connections and transfer knowledge between contexts (or parts of the assessment).

In contrast, students with an average holistic score below 2.5 were much less likely to use big picture thinking. These scores of 2 or lower indicate more novice level thinking about the geology constructs, and could be interpreted to mean that the student had not reached the goals of the NYS ESCC. In this group of ten students, only two of them (20%) used big picture thinking, including the student who had the lowest average score in the data set.

5.5.c.1 Characteristics of Big Picture Thinking at Different Levels. I observed an interesting pattern with these two students who thought about the big picture but received low scores. The nature of their sensemaking was different from the more high-achieving students. They were both persistent in creating a "big picture" story about the geologic history of the location represented in the performance assessment, even when doing so led them to abandon their scientifically accurate ideas in favor of an erroneous, but more complete, narrative. They expressed uncertainty or self-doubt in their think-alouds on multiple occasions. This uncertainty in their initial ideas seems to have been a major motivator for them to pursue an integrated understanding of the performance assessment context.

Dorothy's think-aloud illustrates this problem-solving pattern. We saw in an earlier example that Dorothy described trilobites as "sea roaches." These trilobite fossils, present in both the physical rock sample and the outcrop cross section diagrams, became the keystone of the narrative she constructed about this region's geologic past. In the selected excerpts from her think-aloud in Table 15 below, we can see how her thinking progresses over the course of the assessment. Highlighted portions of the text emphasize ideas from different constructs (pink =

geologic time and stratigraphy; blue = surface processes; orange = plate tectonics). She first considers the fossil specimen when she encounters it in a rock sample, making an only partly-accurate characterization of the animals and their environment; later she connects these ideas with evidence of trilobites in rock layers to make assumptions about the changes that have taken place in this location.

Ultimately, Dorothy reframed this big picture to tell the story of the trilobites, rather than the geologic story of this landscape. She used ideas from different constructs (plate tectonics, geologic time, and surface processes) to think about how major changes or events have affected the trilobites. She conceived of the current river as leftover water from the sea where the trilobites used to live, and the importance of the features present in the outcrops is that they suggest tectonic motion or events that affected living trilobites in the past. These conclusions are not justified by the scientific evidence presented in the assessment materials, nor are they supported by principles of geology. Nonetheless, Dorothy was able to create an internally consistent narrative that brought together multiple aspects of the performance assessment task.

Table 15: Dorothy's Big Picture about Trilobites

While looking at rock samples in part 2	Like I said, rock 2 can use the fossil and it was in the Cambrian, the first period of time. And… oh, now I remember, they were like sea roaches. The sea roaches that bottom feed. [...] And weren't these really big? These were disgustingly big. These weren't the only species of roaches. There were plenty that were not that big. But some were bigger than me. Weren't they able to become like six feet tall because there wasn't no humans around to take the oxygen? They didn't have predators because nobody wanted them.
While interpreting geologic cross sections in part 3	I'm assuming this pattern [with the rocks in outcrop C forming first] because they were more… sea roaches in the ground. And then in D, they eventually died out and then they began being formed into rocks. And then in outcrop E when they were not existing they just got morphed from sedimentary rocks into these ones that are, I'm assuming metamorphic. In the outcrops, rocks are layered on top of layers. And over time there's more layers formed because there's like dead animals making another layer of rock, and then erosion, and then metamorphic from sedimentary.
While inferring past crust movement in part 3	Okay, oh yeah, I actually get this. Because the ones over here at C are like … there could have been an earthquake from all this movement, and that caused the trilobites to probably either be obliterated and they get rolled into rocks, and then those keep getting pushed up by other changes. And then maybe an earthquake on the other side.
While summarizing the "big picture" of change over time in this region in part 4	4b: I'm thinking about the fossils themselves, and not what happened before them. [Picking up rock 2, which contains the Phacops fossil] They look like little gerbils. 4c/d: The rocks shift, causing the water to change. Kind of like the sea level. The river. It like, became smaller. It was a sea before. An earthquake… or like…. People started to come. And that, that caused the trilobites to decrease. Not because, like, they have predators. But they're lacking oxygen now. And then they grew smaller. Then they're… speared. They're extinct. [returns to part a] 4a: I skipped that one because I couldn't think of the earliest geologic events, just from the fossil's existence because if it's a trilobite it was existing in the water, and because of the sand moving around, I'm not sure about the earliest, I just know it will be… I'm not quite sure about the earliest geologic event, because I know there were dinosaurs, and then … like their river was already there before. Their sea was already there before. It's very barren land. No life forms.

Students who expressed a more accurate big-picture understanding of the performance assessment scenario consistently gave written responses that were both scored and coded at levels 3 and 4. Juju, the only subject who achieved an average holistic score of 4, provides an example of this kind of thinking. (Note that this does not mean their assessment was “perfect,” only that their thinking showed enough evidence of level 4 criteria across constructs to generate *holistic* scores at this level.) This excerpt is from part four of the assessment, in which students are prompted to describe the landscape, using words and illustrations, at four geologically distinct times. In comparison with Dorothy’s response to part 4, which focused on the trilobites, Juju’s response is focused on the local geology and how it provides evidence for past events and environments. They use ideas and thinking grounded in three different constructs to build this narrative. Some examples are highlighted in the excerpt in Table 16 (pink = geologic time and stratigraphy; blue = surface processes; orange = plate tectonics).

Table 16: Juju's Big Picture about Geologic History

4a	Okay, at the time of the earliest geologic event, I said that's when limestone was forming. So I know it was a shallow sea, I know it was before the Devonian period, probably, or at least it was like the early Devonian. Because that fossil was Devonian so the rocks that formed before it were older. And, so, looking at that time, it was like … it could have been the Silurian, Ordovician, or even Cambrian time. But there's no erosion between them, so probably not so far before the Devonian. [...] So it's a shallow sea or ocean, and [...] there were probably organisms… oh, I know there were organisms because I know that limestone was made of shells. Um, I wonder what kind of shells lived then. I'm looking at my reference tables again. Looking at the index fossils. At that time, it looks like there's definitely corals, and also the shell Eospirifer was there, possibly, that's a brachiopod. So I'll put possibly brachiopods, corals, could have been eurypterids, trilobites, etc. Okay, so I'm going to just draw water. [...]
4b	At the time of the fossil's existence. So I know this was the middle Devonian period. I know it was still underwater. And the fossil was in siltstone. So I know that siltstone was really small sediments. And that's probably like from really calm water, so maybe a little deeper. Well, I'll make my water look calmer anyway. And I know that lots of trilobites were living there. So I know that, when I look at the stream velocity graph, I know that for silt the biggest silt particles are .006 and I know that would have a velocity -- a maximum velocity of like 0.4 centimeters per second [interpolating on graph with finger], and that's pretty small. So I'll just make my water look calm. [...] And there was a silt bed. I'll draw some silt sediments. It should be a pretty deep layer I think. They could've been from stuff on the land running into the ocean. I'm going to draw a phacops here. I know they like to crawl on the bottom of the ocean. [...]
4c	At the time of rock movement I know that there was convergence. So that means that there was convergent plate movement, and that was probably causing mountains. And then, so the land was coming above the sea level. So I'll draw land rising up. And the layers getting kind of mushed around, like we saw before, and maybe the little sea is down here. I don't know, there's probably like, volcanoes and stuff happening. I'll draw a volcano. I don't know when that happened. It looks like the water level went down because I see that the sand was on top. So there's probably like an ocean margin here. And there's probably sand and stuff in here. And… I think, let's see, Devonian… um, I'm looking in the reference tables again. [...] Oh, I see. Maybe this is all the way up here in the Permian time where it says the Alleghenian orogeny, caused by the collision of North America and Africa. And that would've happened about 300 million years ago when Pangea formed. Yeah, that seems right. I'm not totally sure, but it seems like that had to happen after. It couldn't be the Acadian orogeny because it says the catskill delta formed, or the phacops lived, after that. So it was underwater before that. So the Alleghenian orogeny seems like about the right time for it. And at that time there was reptiles and stuff. So probably reptiles lived on land, maybe around the early Permian time. No dinosaurs yet. I don't see any index fossils here, so I'll just draw a little lizard. I think they kind of looked like little salamanders then. Okay. It does say… yeah, abundant reptiles.

The difference in Dorothy and Juju's thinking is also apparent in their written responses to part 4, which was the section of the assessment intended to help prompt and capture big picture thinking. Juju's response, on the right of Figure 51, shows a changing landscape, incorporating their ideas from multiple parts of the assessment into a progressive story. The features in Dorothy's illustrations show little change to the landscape, because she focused her narrative on the living organisms, rather than on the evolving geology of the region.

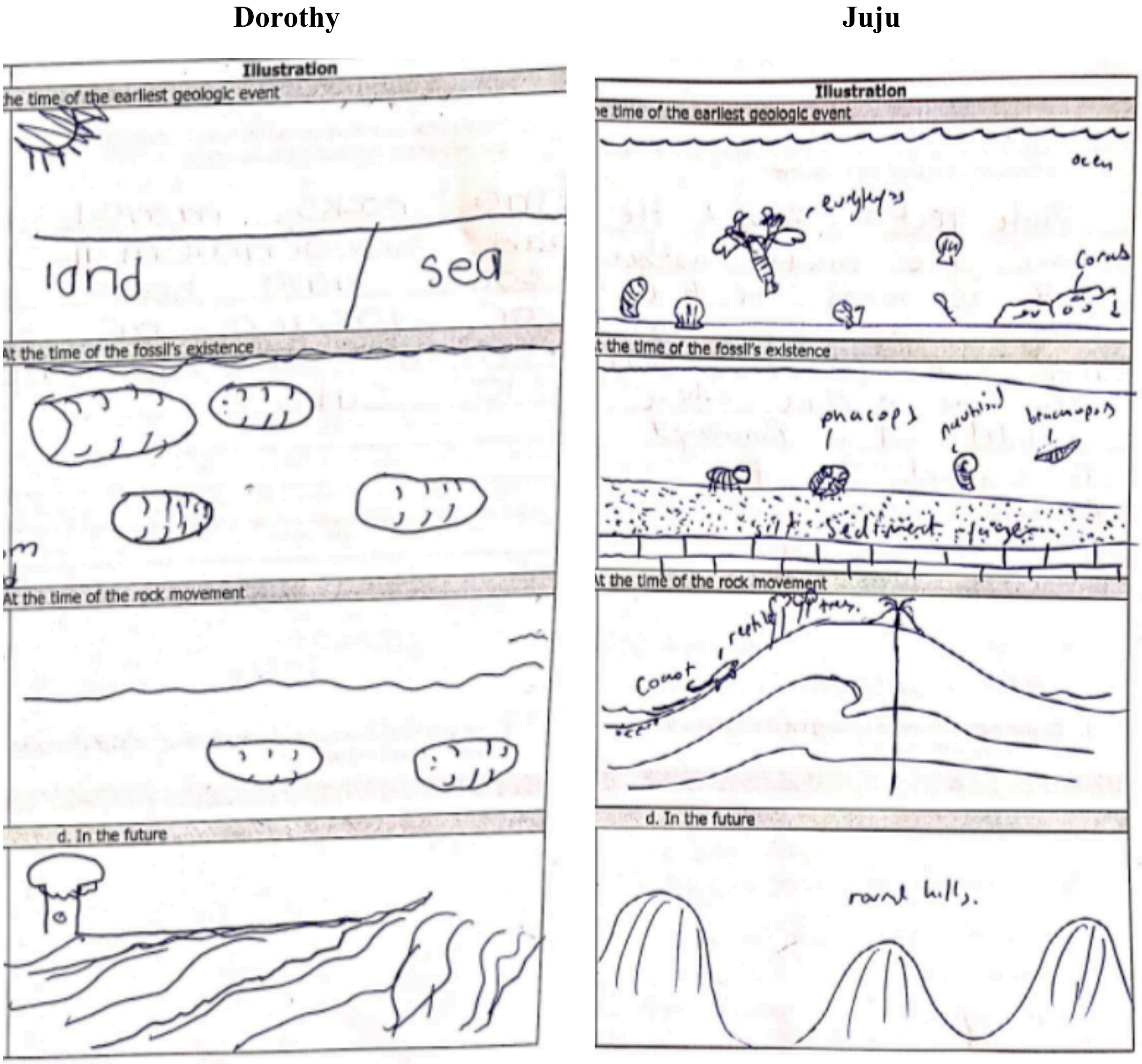

Dorothy
Juju
Illustration
he time of the earliest geologic event
land
sea
At the time of the fossil's existence
At the time of the rock movement
d. In the future
Illustration
he time of the earliest geologic event
ocen
eurypterys
Corals
t the time of the fossil's existence
phacops
nautiloid
brachiopods
silt sediment layer
At the time of the rock movement
Comot
reptile
trees.
d. In the future
round hills.

Figure 51: Big Picture Illustrations.

One implication of this comparison, and the overall trend I observed, is that big picture thinking is a necessary, but insufficient, part of developing and expressing a more expert understanding of geology constructs. This has further implications for the design principles for cognitively-based performance assessment in content domains like Earth Science that are highly systems-based. When designing cognitively-based assessment tasks that target multiple constructs, it is important to think about the relationship between constructs, and to intentionally design task components to leverage those relationships.

Additionally, we need to consider the interplay between individual constructs and big picture thinking when creating construct maps and other assessment system components. Is it possible for student thinking about these constructs to occur in cognitive isolation, especially at the higher levels? If it is the case that students need to use relationships between constructs to make sense of a task or phenomenon, claims about student thinking at the expert end of a learning progression are fundamentally different in nature from claims about thinking at the novice end. Pursuing a more cohesive assessment setting may interact with our ability to observe and quantify student thinking in distinct constructs.

At the same time, it's important to remember the goal of science assessment reform to move beyond assessments that measure only decontextualized, superficial knowledge (Kohn, 2004). If we wish to promote systems thinking and understanding that builds on connections between ideas, we need assessments that measure this kind of thinking as well. With that in mind, I do not believe these findings suggest that we should intentionally segregate constructs in a cognitively-based assessment, the way traditional assessment items (e.g. multiple-choice questions) have often done. However, we do need to be aware of the complexities introduce when students are thinking in a big picture context.

Chapter 6: Discussion

6.1 Cognitively-Based Science Performance Assessment System

In this study, I explored a process for designing and scoring science performance assessments that would illuminate student thinking more coherently than traditional assessments. This process combines the principles of cognitively-based assessment with those of science performance assessment. It draws from the integrated assessment model developed at the Berkeley Evaluation and Research Center (BEAR), which is founded on the idea that tools for observation, scoring, and interpretation should be designed to align with a developmental perspective on student learning (Wilson, 2005). I used existing literature on learning progressions and cognition in geosciences, in combination with my experience as a classroom teacher of Earth Science, to create a construct map for four distinct areas of Earth Science disciplinary understandings and practices. These construct maps served as the cognitive model for the development of both the performance assessment and the associated scoring procedure.

My goal was that this approach to assessment design would promote coherence between the three cornerstones of assessment: cognition, observation, and interpretation (NRC, 2001). *Cognition* is the intended student thinking and understanding developed by classroom activities and then elicited by the related assessment; *observation* is the representation of that understanding shown via the written responses prompted by the performance assessment tasks; and *interpretation* is the value assigned to written responses via a scoring procedure mediated by a predefined outcome space. In my analysis, I examined the coherence between pairs of each "cornerstone" in order to illuminate the successes, failures, and implications that arise from one specific instantiation of this approach. While the field's understanding of cognitively-based assessment has developed significantly in recent decades, its application to performance assessment is a new approach that needs to be understood more fully. In this discussion I will turn to a broader view of the cognitively-based science performance assessment process as a complete system.

Through this research, I have gained insight into both the potential of cognitively-based performance assessment to measure and quantify student thinking, and the significant challenges inherent in this approach. It is exceptionally difficult to fully capture student reasoning in written responses. This is a known problem for the assessment field; here, I will discuss what I've learned about the nature of this difficulty and some potential solutions.

The cognitive science research underlying the development and content of my construct map makes it clear that patterns of thinking and learning have the potential to be both complex and construct-specific (Mislevy & Baxter, 2005; Ormand et al., 2017). This complexity presents a significant challenge for assessment, and an important goal of cognitively-based performance assessment is to confront this challenge (NRC, 2014). My experience with the development and implementation of a cognitively-based performance assessment system has led me to understand that it is necessary to prompt for this complex, construct-specific thinking very intentionally and explicitly. A cognitively-based design model (i.e., a construct map) is a vital tool in this process. In order to be maximally useful for assessment design, the construct map should describe a clear progression from the novice end to the expert end of the spectrum. Wiser, Smith, and Doubler (2012) point to the importance of these "stepping stones" for instructional design purposes; I contend that they are equally important for assessment purposes. Cognitive models that are based on a dichotomy may help us design assessment items that students can get right or wrong. Cognitive models that articulate a more gradual change in thinking help us take a more nuanced approach. Descriptions of the "stepping stones" can point us to specific words or response formats for assessment prompts. They also guide the development of the outcome space.

6.1.a Outcome Space

I believe, at least for initial rounds of an iterative assessment design process, that "more is more" when it comes to the outcome space. The outcome space is the part of the assessment system that tells us how to put value on students' responses in a way that aligns with the underlying cognitive model. The greater the number of specific outcomes pre-defined for each level of the construct, the less interpretation will need to be done by individual human scorers. The outcome space can therefore make the "interpretation" corner of the assessment triangle increasingly objective and consistent – improving coherence-based validity – while preserving the capacity of performance assessment to observe students' understanding of science content and practice in a more authentic context than traditional assessments. The first two design principles that guided my approach to creating this cognitively-based science performance assessment demand both richness and authenticity. Without a thorough and well-developed outcome space, a "rich" and "authentic" performance assessment may quickly become unwieldy, generating student responses whose scoring is dauntingly time-consuming and overly subjective. The outcome space demands a significant amount of work prior to administering and scoring the assessment, but this work is an investment that will continue to deliver the benefits of cognitively-based performance assessment on subsequent administrations and versions of an assessment task.

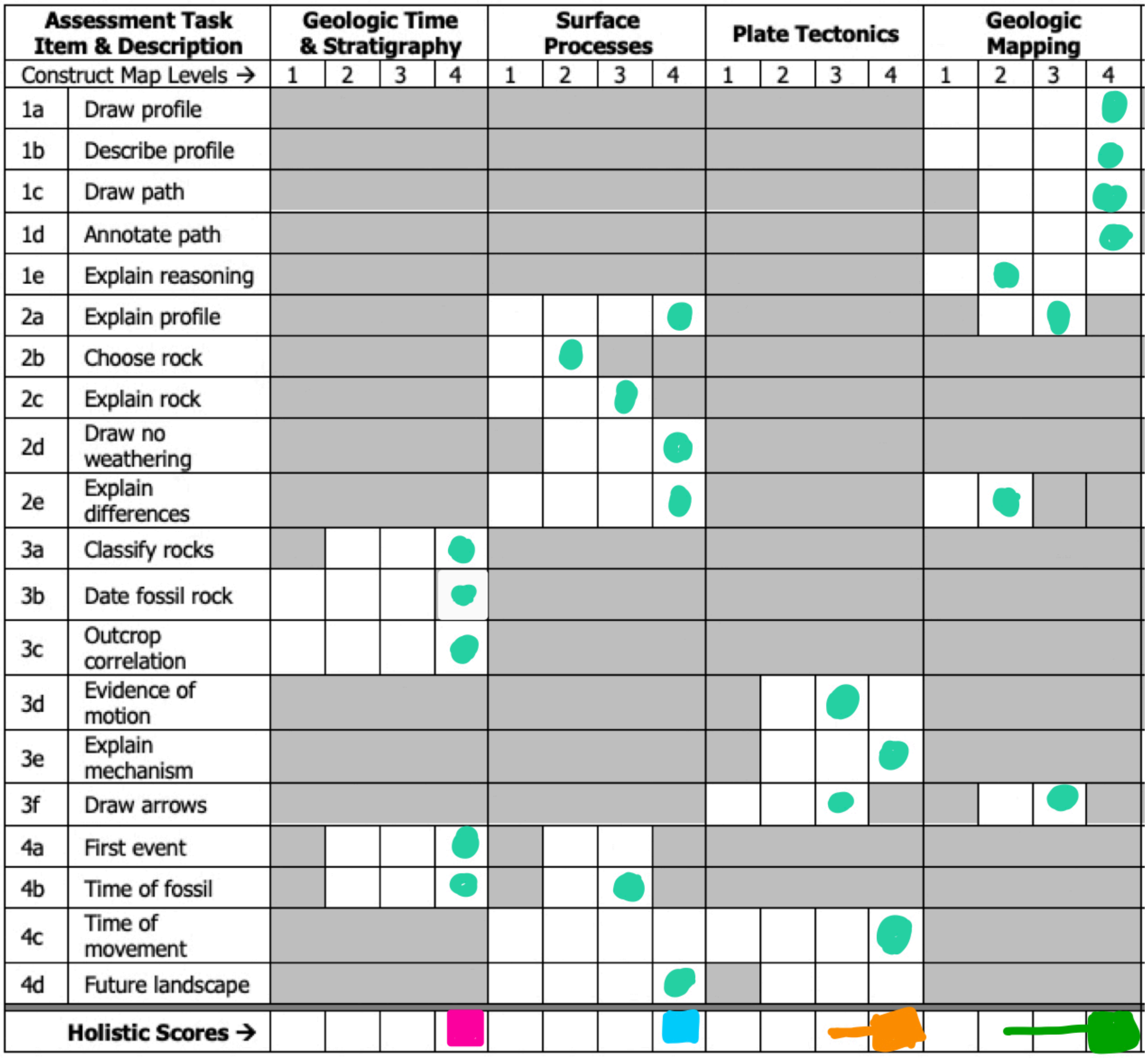
An open question remains about the role of the outcome space: should the scoring procedure allow for the scoring of constructs beyond that which is predefined in the outcome space? In other words, if the outcome space doesn't describe a response for construct X on item

Y, but a student response evokes construct X on item Y, what should we do with it? I found that instances such as this were more likely to reduce the coherence between student thinking, responses, and scores than those where the outcome space provided scorers with guidance. There are a few possible reasons for this threat to coherence. Students who bring to bear ideas about the construct(s) that an assessment task did *not* deliberately prompt for may do so in ways that are outside the typical progression identified by cognitive science research. Additionally, requiring scorers to interpret and apply the construct map introduces a potential source of error or subjectivity. On the other hand, including these data would give us more information that could inform a student's holistic score. Data from student responses in these "out of the box" contexts may be particularly valuable in capturing a holistic view of their thinking. The answer to this question may depend on the scale at which the assessment is being used. At larger scales, across classrooms, schools, or even districts, it would be important to minimize variations in the scoring protocol. For individual teachers using this approach for classroom assessment, the benefits of scoring outside the outcome space – or, perhaps better, having the flexibility to add outcome space criteria – could outweigh the drawbacks.

6.1.b Holistic Scoring

The role of holistic scoring in this assessment approach was more interesting than I anticipated. In some ways a concession to the uncomfortable pragmatic need to reduce children into data points, this final step of the scoring procedure also encapsulates the answers to many questions we need to consider when attempting to measure cognition. When a student demonstrates thinking across a range of sophistication levels, what is the most accurate way to characterize that thinking: with the maximum, the mean, or the mode? (I don't think anyone would suggest the minimum is appropriate, but there's another option!) What is the appropriate balance between providing detail and being concise in score reports? How can these data be summarized in a way that is readily accessible to teachers or other users?

I chose an approach to generating holistic scores that I hoped would strike a balance between giving students "credit" for their most sophisticated thinking and characterizing the range of their thinking within each construct. According to the scoring procedure, frequency is the first determinant of holistic score, but since some items were designed to target a limited range of levels for a given construct, I wondered if holistic scores would be artificially lowered. (This is one of many reasons I opted not to use the mean score as an overall indicator of student thinking.) This does not appear to have happened. The assessment provided enough opportunities for students to demonstrate their thinking at more sophisticated levels.



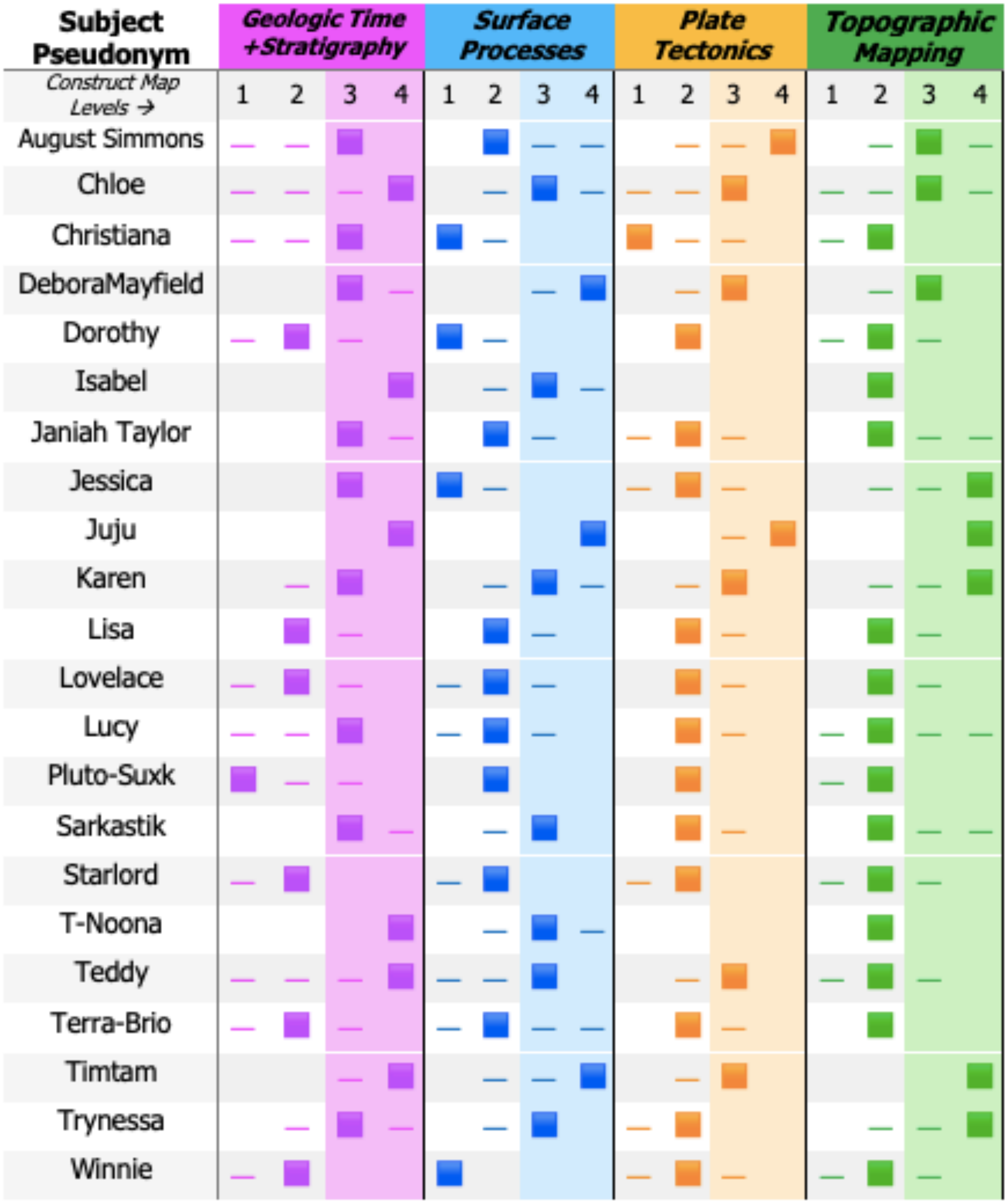
Assessment Task Item & Description		Geologic Time & Stratigraphy				Surface Processes				Plate Tectonics				Geologic Mapping			
Construct Map Levels →		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1a	Draw profile	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨				●
1b	Describe profile	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨				●
1c	Draw path	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨			●
1d	Annotate path	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨			●
1e	Explain reasoning	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨		●		
2a	Explain profile	▨	▨	▨	▨				●	▨	▨	▨	▨	▨		●	▨
2b	Choose rock	▨	▨	▨	▨		●	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
2c	Explain rock	▨	▨	▨	▨			●	▨	▨	▨	▨	▨	▨	▨	▨	▨
2d	Draw no weathering	▨	▨	▨	▨	▨			●	▨	▨	▨	▨	▨	▨	▨	▨
2e	Explain differences	▨	▨	▨	▨				●	▨	▨	▨	▨		●	▨	▨
3a	Classify rocks	▨			●	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
3b	Date fossil rock				●	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
3c	Outcrop correlation				●	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
3d	Evidence of motion	▨	▨	▨	▨	▨	▨	▨	▨	▨		●		▨	▨	▨	▨
3e	Explain mechanism	▨	▨	▨	▨	▨	▨	▨	▨	▨			●	▨	▨	▨	▨
3f	Draw arrows	▨	▨	▨	▨	▨	▨	▨	▨			●	▨	▨		●	▨
4a	First event	▨			●	▨			▨	▨	▨	▨	▨	▨	▨	▨	▨
4b	Time of fossil	▨			●	▨		●	▨	▨	▨	▨	▨	▨	▨	▨	▨
4c	Time of movement	▨	▨	▨	▨								●	▨	▨	▨	▨
4d	Future landscape	▨	▨	▨	▨				●	▨				▨	▨	▨	▨
Holistic Scores →					■				■				■				■

Figure 52: Student Score Report. The mode item score results in a single holistic value for each construct.

For example, the student whose scoresheet is shown in Figure 52 above has individual item scores ranging from 2 to 4. By using frequency, the scoring procedure identifies their holistic score as level 4 for each construct, even though the average score for several constructs would have been around 3.5. For the surface processes and geologic mapping constructs, frequency alone is all that is needed to identify level 4. In the surface processes construct, this student "maxed out" each item, receiving the highest available score defined by the outcome space. On the topographic mapping construct, the student had one suboptimal response. On the plate tectonics construct, a tiebreaker was needed because both level 3 and level 4 were represented twice. The tiebreaker algorithm privileges items that measure across the tied levels; within this subset, level 4 is represented twice and level 3 only once. The plate tectonics construct demonstrates why there should be a minimum threshold for opportunities to score each construct. I feel a lot more confident in the holistic scores that were generated from six or more individual items. This confidence is reinforced by the significance in results for the surface processes and topographic mapping constructs compared to the plate tectonics construct. On future cognitively-based science performance assessments, I would require 6 - 10 items per construct, which could necessitate more intentional overlap on some items that would be designed to target more than one construct. (Such overlap items present their own challenge, since I found that students tend to focus their thinking on one construct at a time, especially at more novice levels.) This is important because it shows that holistic scoring allows us to accurately identify when students are *consistently* using more sophisticated modes of thinking in response to science assessment tasks, without sacrificing the specificity of items designed to distinguish between mid-levels of the construct map and/or provide access points for students whose thinking has not yet reached the more expert levels. Additionally, it is not necessary or appropriate to holistically score students at the highest level they achieved, as long as a sufficient proportion of the performance assessment tasks provide opportunities for students to demonstrate higher level thinking.

A vital function of holistic scores is to provide information that is useful and actionable for both formative and summative purposes. As described in chapter 3, the purpose of generating a holistic score on a cognitively-based assessment is to characterize the student’s learning along a known continuum. When teachers–or other people who make decisions about a student's school experience, including parents, counselors, and programmers–have access to this information, they can make decisions about appropriate next steps for that student's learning. However, such decisions are not typically made about individual students alone, but rather about groups of students (such as course sections) or subgroups (such as students with disabilities in a co-taught class). To facilitate these decisions, holistic scores need to be readily visible across such groups. Figure 53 shows an example of a classwide holistic score summary for this cognitively-based science performance assessment system. In this visual representation of the holistic score data, each student's holistic score is shown with the solid box, while dashes are used to represent the range of scores they had on that construct. Columns for levels 3 and 4 are shaded to provide a visual indicator of the proficiency target.

In this view, teachers or other users can look across rows to see a summary of information for a single student, and down the sectioned columns to see a snapshot of class performance by construct.



Subject Pseudonym	Geologic Time +Stratigraphy				Surface Processes				Plate Tectonics				Topographic Mapping			
Construct Map Levels →	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
August Simmons	–	–	■			■	–	–		–	–	■		–	■	–
Chloe	–	–	–	■		–	■	–	–	–	■		–	–	■	–
Christiana	–	–	■		■	–			■	–	–		–	■		
DeboraMayfield			■	–			–	■		–	■			–	■	
Dorothy	–	■	–		■	–				■			–	■	–	
Isabel				■		–	■	–						■		
Janiah Taylor			■	–		■	–		–	■	–			■	–	–
Jessica			■		■	–			–	■	–			–	–	■
Juju				■				■			–	■				■
Karen		–	■			–	■	–		–	■			–	–	■
Lisa		■	–			■	–			■	–			■	–	
Lovelace	–	■	–		–	■	–			■	–			■	–	
Lucy	–	–	■		–	■	–			■	–		–	■	–	–
Pluto-Suxk	■	–	–			■				■			–	■		
Sarkastik			■	–		–	■			■	–			■	–	–
Starlord	–	■			–	■			–	■			–	■	–	
T-Noona				■		–	■	–						■		
Teddy	–	–	–	■	–	–	■			–	■		–	■	–	
Terra-Brio	–	■	–		–	■	–	–		■	–			■		
Timtam			–	■		–	–	■		–	■					■
Trynessa		–	■	–		–	■		–	■				–	–	■
Winnie	–	■			■				–	■	–		–	■	–	

Figure 53: Classwide Score Report for Use by Teachers.

Including columns for all four levels of each construct emphasizes the continuum along which the holistic scores are situated. Shading in the columns allows the teacher to overall performance relative to a target threshold for proficiency (level 3) within that continuum. Including the "range" information helps teachers identify a zone of proximal development for each student (Vygotsky, 1978). For example, a teacher looking at the score report shown in Figure 53 would see that the majority of students are currently at level 2 thinking for the topographic maps construct, but are exhibiting some thinking at level 3 as well. This kind of information is indispensable for making decisions about future instruction, and allows this system of assessment to be used for formative assessment purposes. The teacher can look to the construct map for a clear description of the next step in the development of student thinking along the topographic maps learning progression, and subsequently engage students in classroom activities that help them bridge the cognitive gap between levels 2 and 3. The visualization of holistic data makes it simple to include information about the range in student thinking. I believe this inclusion is a significant enhancement over reporting single scores, for formative purposes.

The information used to generate holistic scores by construct could, in turn, be converted into an average holistic score – one that weights constructs, rather than individual items, equally – but I would not recommend this approach for decision making purposes, only when necessary for generating a single overall "grade" for the assessment.

These advantages of holistic scoring are meaningful only if the holistic scores are at least as good at characterizing student thinking as the individual items. In fact, in my study, holistic values consistently provided a *more* coherent alignment with think-aloud data than student responses to individual items. Furthermore, the coherence between think-alouds and scores was stronger than, or comparable to, coherence between think-alouds and written codes. Figure 54 below shows a summary of the R^2 values for comparisons across individual responses and holistic values, for all three research questions, and three selected constructs. (I removed the plate tectonics construct for the purposes of this discussion because the majority of results were not statistically significant. I will discuss this construct further in a later section of this chapter.)

Comparison Method	**RQ2** (TA → written)	**RQ3** (written → score)	**RQ4** (TA → score)
Geologic Time & Stratigraphy Construct			
Individual Responses	0.60	0.89	0.61
Holistic Values	0.74	0.84	**0.81**
Surfaces Processes Construct			
Individual Responses	0.57	0.77	0.65
Holistic Values	0.59	0.87	**0.74**
Topographic Maps Construct			
Individual Responses	0.69	0.94	0.66
Holistic Values	0.78	0.86	**0.75**

Figure 54: R^2 Values Across Research Questions

For these three constructs, a consistent pattern emerges. The holistic comparisons resulted in statistically significant correlations for research questions 2 and 4, and in those cases, the correlation was stronger than for individual responses. In the column for research question three, we see that correlation between written codes and scores was strong, but statistically significant only for individual responses, not for holistic values. I believe this is because the infrequent disagreements between the holistic scores and holistic written codes could go either way; i.e., it was not the case that disparities were always due to written codes being higher than scores. These disagreements arose from situations where the student's written response had a greater number of codes for a given construct than the prescribed number of scores available via the scoring procedure. It is worth noting that there were only 3 instances of disagreement for each of these constructs; the holistic values were identical for written codes and scores 82% of the time, and never diverged by more than one level.

We have already seen that a major challenge for assessment lies in the difficulty of fully capturing student thinking and distilling it into quantitative output. This is consistent with the pattern in which correlations between think-alouds and either of the other two data sources have lower R^2 values than correlations between written codes and scores. However, I believe it is important and encouraging that the R^2 for Research Question 4, comparing think-aloud codes to scores, is comparable to or higher than R^2 values for the comparison between think-aloud and written codes. Each of these research questions represents a data reduction step, first from student thinking to student response, then from student response to a score. The holistic values themselves represent an additional layer of data reduction. I initially expected that a loss of fidelity might occur at each of these steps, but this does not appear to have happened.

Another important aspect of the pattern is that, for these three constructs that "worked," R^2 for comparisons of holistic values was similar to or higher than the R^2 for comparisons on individual responses. The final output of this assessment system is a holistic score, intended to capture and characterize student thinking. The coherence of the entire system is, in a sense, encapsulated in the results for the holistic value correlations in Research Question 4. These results are shown in bold in Figure 54 above. They show that around 75% of the variation in holistic scores generated by this system can be attributed to variation in student thinking. While this isn't an ideally high percentage, it is, I believe, a successful starting point. The consistency in these values, and in the overall pattern across constructs, gives me confidence that this is a meaningful result. For all of these constructs, the possible scores on each item were bounded in different ways, while the think-aloud coding allowed for items to be coded at all four levels. In spite of this difference, the holistic scores of student work–which came out of the strictly regimented scoring procedure and were assigned by different human scorers–generally matched the holistic values from open coding of student thinking. The holistic scores have done a decent job of measuring student thinking in a broad sense, in a way that can be used for both summative and formative purposes, despite some messiness in the item-wise capture of student thinking via written responses. This speaks to the potential power and utility of holistic scoring on a cognitively-based assessment: it can smooth out some of the noise, and find the signal from a student's overall pattern of thinking. This power also confers the responsibility of getting the *right* signal, because it has taken such a large amount of information (and presumably an even larger amount of complex thought) and turned it into a single quantitative data point. The correlations in my study show that there is more we can do to fine-tune this process, but also that the goal of using cognitively-based performance assessment to accurately describe student thinking is within reach.

6.2 Construct-Specific Insights

In this section, I will revisit each of the four content-specific constructs. Selected findings about each construct include reflection on the interactions between assessment design and my results, as well as on how my observations of student thinking in this study connect to the theoretical framework underlying each construct. The full construct map is found in Figure 2.

6.2.a Geologic Time & Stratigraphy

This construct was based on Dodick and Orion (2006)'s research on how geology learners interpret stratigraphic information in rock outcrops, combined with my extrapolated ideas about how these different levels of understanding could apply to students' inferences about the environmental changes implied by a complex series of outcrops. Some of the criteria based on Dodick and Orion's work came out much more frequently than others in student thinking. At level 1, I did not observe any instances in which students expressed the idea that the number of layers in an outcrop was an indicator of the outcrop's age. However, I did observe some students whose thinking showed a dichotomy between "less ancient" and "extremely ancient." These students seemed to categorize anything that could create a fossil as "extremely ancient," and forces that caused surface changes like erosion as "less ancient." One student repeatedly referred to "dinosaur time" when thinking about anything that happened in the distant past. My observations did agree with the criterion described by Dodick and Orion in level 2: some students ignored crosscutting features (e.g. fault lines) in the outcrop diagrams, and others noticed changes to the rocks' original horizontality but were unable to explain what could have caused the change. At level 3, I did observe many students who expressed thinking about how the outcrops provided additional evidence, beyond the law of superposition, for past events. Students who correlated outcrops were significantly more successful at describing an overall history of the region, so this is in agreement with Dodick and Orion's assertion that correlation of rock units across outcrops is a feature of more sophisticated thinking in this construct. I did not observe any student attending to the thickness of rock layers, although this may be due to a design flaw: the thickness of rock layers in the outcrops did not vary significantly enough to demand attention. Finally, very few students attributed environmental changes to long-term geologic change, but those who did consistently scored or were coded at level 4 for this construct across the assessment items. In this way my results are also consistent with the model of thinking described by Dodick and Orion.

I found that there was evidence of particularly rich student thinking around this construct in the think-aloud data. The assessment was only partly successful in capturing this thinking. A strength of the assessment was that it prompted students to think about the geologic past of rocks in several different contexts and at different scales, even though it was all related to the same location-based scenario. Students were able to use individual hand samples of the rocks represented in the outcrop diagrams to make observations of their properties. Their physical manipulations and observations of these materials were a critical part of their thinking as they constructed their ideas about each rock, and then about the change over time to the region's environment. For example, one student made a direct comparison between sandstone and its metamorphic daughter rock quartzite, noting their similar color but distinct texture and (qualitative) density, which made him confident that some form of metamorphism had occurred in the region. Another student gently removed sediments from different rocks using the tabletop so that she could confirm the grain size, which she later used to infer a change from high energy to low energy in the depositional environment over time. This kind of thinking is unlikely to have occurred if students were interacting solely with illustrations or photos of the different rocks.

A weakness of the assessment design, for this construct specifically, is that it did not provide students with space or prompts that encouraged them to record the entirety, or even the majority of their thinking. This is a design flaw that led to some items having a consistent mismatch between think-aloud codes and the values for other data sources, and more generally fell short of capturing the complexity and depth of student thinking. For example, in item 3c, students are asked to list the sequence of geologic events that created the three outcrops visible in the location. Students at all levels of understanding spent a significant amount of time reasoning through this task. But the assessment tool did not give students room or reason to record their thinking.

It happened that this design flaw extended to the other items that targeted the geologic time and stratigraphy task as well. The majority of the items prompted open-ended thinking, without providing adequate space for students (especially those whose thinking aligned with higher levels of the construct map) to write it all down or otherwise record it, and without sufficiently cluing those students into the desired level of detail. If I were to redesign the items for this part of the assessment task, I would be more intentional about providing these prompts and adequate space.

6.2.b Surface Processes

This is the only construct for this assessment that had minimal grounding in existing learning science research. I cannot draw conclusions from my assessment data about whether the construct map "works" for agents of weathering and erosion other than water: the assessment did not prompt for this and no students brought them up unprompted. I found that students conceived of water as a powerful and important agent of weathering and erosion, as exemplified by August Simmons's language when describing it:

> "So running water tends to break down rocks because kinetic energy and stuff. And assuming the river was much larger in the past it perhaps has carved through the rock. The landscape I guess, which I'll assume is similar to, like a valley."

Students who recognized evidence of weathering in rock samples were highly likely to make a connection to personal experience, either "real life" experiences such as visiting a beach, or classroom activities in which they had previously participated. Many students were able to correctly identify which rock sample had experienced weathering in moving water, citing "softness," "smooth" textures, and "round corners" as salient features. Only a subset of these students were able to describe the abrasive mechanism through which those features are created. Although specific to this example, my observations here imply that the proposed construct map has put these understandings in the correct order of progression. This also shows the utility of designing items that can be scored separately to differentiate between levels of understanding. If the assessment went straight to asking students for an explanation, it would have been harder to identify the "recognition" of evidence that characterizes thinking at level 2.

I anticipated that an outcome of this study would be the revision of the criteria at each level of the surface processes construct, because it was not grounded in empirical research like the others. Based on the trends in student think-aloud data, I now have evidence that students think about the effects of surface processes on individual rocks or sediments before they progress to thinking about the effects on entire landscapes. Students whose overall performance was towards the more novice end talked about changes they would expect to see in rocks, such as cracks, missing pieces, or "crumbling." In contrast, students whose overall performance on this construct was towards the more expert end talked about changes they would expect to see over larger scales, such as the transformation of mountains to a more rounded shape. The original construct map hinted at this progression, but a revision could make it more specific and explicit.

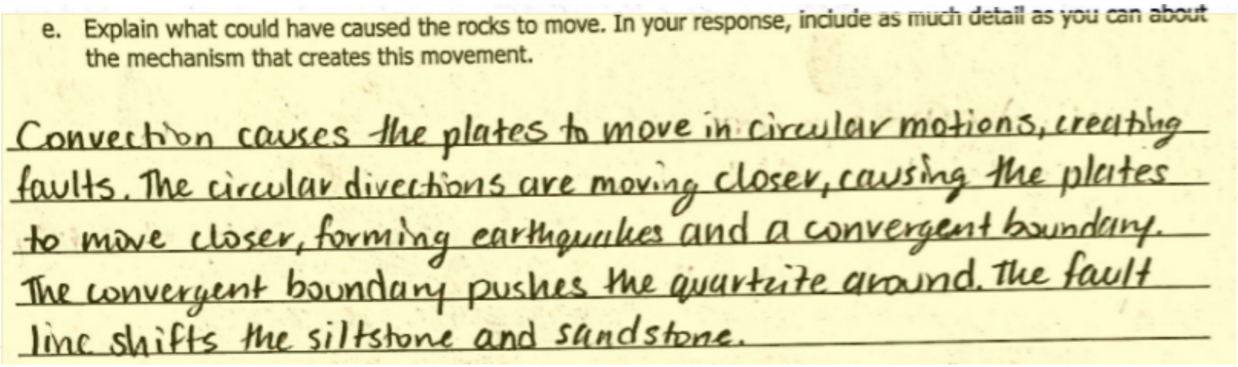
There was one item for this construct that I would particularly like to redesign. In 2e, students imagined what the landscape would look like in the absence of the river that currently runs through it; then, in 2f, they explained the differences in the two topographic profiles (with and without river) they had drawn. The format for item 2f asked students to give an explanation for each difference they identified. This was confusing to many students because they thought there was a common reason for each difference (water erosion). I would like to redesign these items to capture more of the students' thinking when they were reasoning about what the alternate landscape would look like.

6.2.c Plate Tectonics

As we saw previously in both this chapter and the preceding Findings chapter, this assessment was unsuccessful at generating scores with statistically significant correlations to student thinking. I believe this is a result of poor assessment design. There are several potential reasons for this failure. First, the performance assessment task did not include enough items that specifically prompted for this construct. In chapter 3 I suggested that this might be mitigated by addressing the plate tectonics construct on a separate assessment, but it is now clear that the reduced number of items undermined this assessment's capacity to accurately measure student learning, in addition to not addressing all the standards related to plate tectonics. There was only one item (3e) that gave students a reason to think about the *cause* of plate tectonics, which is necessary for demonstrating understanding level 3 or 4 on this construct. The prompt simply said to "include as much detail as you can" about the cause of rock movement, without giving students any indication of the nature of those desired details.

The multiple instances where there were differences between the think aloud code and the score across items for this construct provide insight into the poor results for this construct. In most–but not all–of the cases where the values did not match, the student's think-aloud was coded at a higher level than the score they received for that item. This was also one of the few constructs where the disagreement was sometimes more than 1. A two-point disparity is quite large on a four-point scale.

How could the assessment prompts have been constructed to better capture student thinking at higher levels? Patterns in student responses provide guidance. In think-aloud responses at levels 3 and 4, there is evidence that students are using spatial thought, visualizing movement (e.g. references to arrows, circles, directions). For example, Juju said, "I'm kind of like, picturing it. The picture of convection and how, when the mantle moves, it makes the little arrows flow around, and then the rocks, the tectonic plates on top get pushed." During the think-aloud protocol, Sarkastik used gesture to express her thinking: "The magma inside the earth is always moving. It doesn't just stay still. It pushes itself up, the way it's moving. It's always moving and it pushes the plates apart. [moving hands in big circles] Convection!" Tim Tam's written response, shown in Figure 55 below, also describes circles and motion.



e. Explain what could have caused the rocks to move. In your response, include as much detail as you can about the mechanism that creates this movement.

Convection causes the plates to move in circular motions, creating faults. The circular directions are moving closer, causing the plates to move closer, forming earthquakes and a convergent boundary. The convergent boundary pushes the quartzite around. The fault line shifts the siltstone and sandstone.

Convection causes the plates to move in circular motions, creating faults. The circular directions are moving closer, causing the plates to move closer, forming earthquakes and a convergent boundary. The convergent boundary pushes the quartzite around. The fault line shifts the siltstone and sandstone.

Figure 55: Tim Tam's Level 4 PT Written Response

This pattern suggests that it would have been beneficial to give students access to a more spatial mode of response, such as sketching or annotating a diagram. One item prompted students to indicate the direction of rock movement on the map, but this required them to use a different frame of reference (a bird's-eye view of the surface rather than the cross-sectional view of Earth's interior described by the students at level 4). Many students rightly expressed confusion about the relative scale of the map compared to tectonic plate movement, and scores on this item were generally low. It likely would have been more effective to

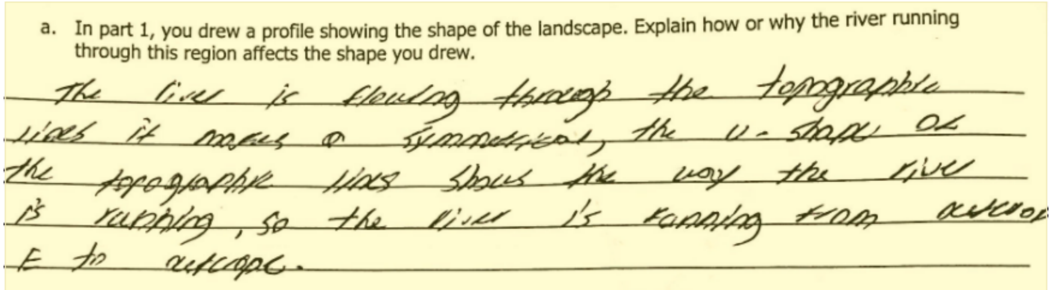
Additionally, in designing the assessment I made some assumptions that students' ideas about plate tectonics would come up in the context of other items. This did, in fact, happen–but not in a consistent manner across students. I also found that students were less likely to write down ideas about plate tectonics, even if they verbally reported those ideas during the think-aloud, on items that were more obviously targeting a different construct. In my mind, plate tectonics is deeply and closely connected to all three of the other content-specific constructs around which this assessment is designed; the students, however, were significantly more likely to talk about plate tectonics in relation to geologic time and stratigraphy than to either surface processes or topographic maps.

Common misconceptions or naive ideas that were expressed by multiple students, either in their think-alouds or in their written responses, included the following: the idea that the tectonic plates were somehow within the earth (located in the core or mantle); reversing the directionality of the cause-and-effect relationship between rock movement and observable rock properties (such as metamorphism); and that sudden events such as earthquakes were the only form of rock movement. My observations are in agreement with Gobert (2005)'s description of novice thinking about plate tectonics. The fact that I observed naive ideas more consistently than expert ideas described by Gobert is another reason to think the assessment task did not give students sufficient opportunity to demonstrate thinking that aligned with higher levels in the construct map.

6.2.d Topographic Maps

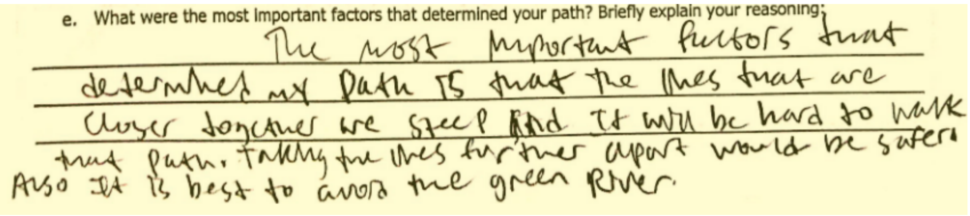
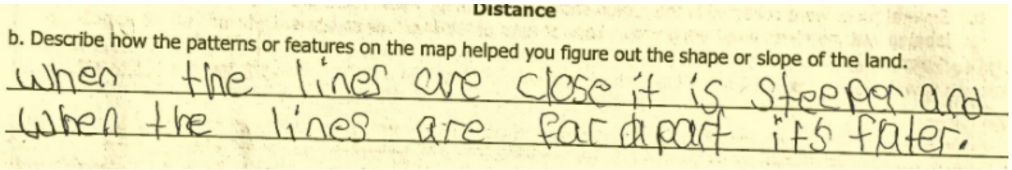
This construct was based on research that well defined the endpoints (levels 1 and 4), without a clear progression across levels in between. A main idea around which the endpoints were based was the recognition of patterns in contour lines representing real-world landscape features, as described by Kastens and Ishikawa (2006). I found that this happened in two primary ways: identifying shapes (e.g. circles) and interpreting line spacing. My observations of the progression in students' understanding and use of contour lines expands on the novice/expert dichotomy previously described in the literature. Table 17 shows examples of how students used shape schema in both their reported thinking and written responses when making sense of the topographic map.

Table 17: Using Shape Schema to Interpret Topographic Maps

Think-Aloud (Tim Tam)	"The spot where the lines make circles shows hilltops." "That's the top of the mountains [points to the area around B] I can say that the circles tell me that we're at the top of a mountain. Would I just say top of the land?"
Written Response (Jessica)	 a. In part 1, you drew a profile showing the shape of the landscape. Explain how or why the river running through this region affects the shape you drew. The river is flowing through the topographic lines it makes a symmetrical, the v-shape of the topographic lines shows the way the river is running, so the river is running from outcrop E to outcrop. *The river is flowing through the topographic lines it makes symmetrical. The v-shape of the topographic lines shows the way the river is running, so the river is running from outcrop E to outcrop C.*

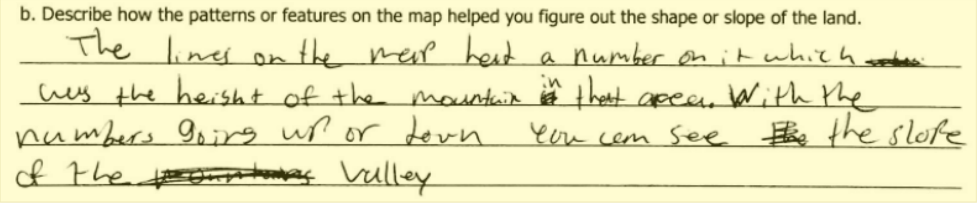
The examples in Table 18, below, show students using contour line spacing to make inferences about landscapes. These examples come from students who had holistic scores of 4, 3, and 2 for this construct, respectively.

Table 18: Using Contour Line Spacing to Make Inferences about Landscapes

Think-Aloud (August Simmons)	"I'm looking at the contour interval. It helps me find the pattern by counting by 20s. As it starts from 700, every time I cross the line I know the elevation of the land changed by 20 feet. I know that when contour lines are close together, the slope is pretty high because the elevation is changing quickly. So the left side of the map would be pretty strong. Steep. Risings-slash-falling pretty quickly on the map. And the opposite would be true for the other side. Where contour lines are spread out, the slope would be pretty not steep."
Written Response (Terra-Brio)	 e. What were the most important factors that determined your path? Briefly explain your reasoning: The most important factors that determhed my Path IS that the lines that are closer togather are steep And It will be hard to walk that path. Taking the lines further apart would be safer. Also It is best to avoid the green River. *The most important factors that determined my path is that the lines that are closer together are steep and it will be hard to walk that path. Taking the lines further apart would be safer.*
Written Response (Lisa)	 Distance b. Describe how the patterns or features on the map helped you figure out the shape or slope of the land. when the lines are close it is steeper and when the lines are far apart its flater. *When the lines are close it is steeper and when the lines are far apart it's flater.*

Students who thought about contour line spacing were more likely to have their responses coded at level 3 or 4. Of the 22 subjects, 11 of them mentioned line spacing in either a think-aloud or a written response. Seven of these students had holistic scores (and codes – 100% coherence in this case) at level 3 or 4 for the topographic maps construct, suggesting that attending to patterns in contour line spacing is a useful cognitive tool for interpreting topographic maps. This may seem obvious, since contour lines are the defining feature of topographic maps! However, it is instructive to compare this thinking to other approaches students used when interpreting the map. Students who did not explicitly think about line spacing tended to focus on the quantitative elevation labels on the contour lines, like in the examples in Table 19. Both of these students were holistically scored at level 2 for the topographic maps construct.

Table 19: Reasoning about Landscapes Without Contour Line Spacing

Think-aloud (Dorothy)	"And then the important factor that determined my path is the contour intervals, because you want to start going directly down because you're already at like 700 feet and then you only have to go… no, it's about the same. [comparing different paths] I'm looking at it, you're at 560 and you're already at 700, and it seems easier to me to just go down, as opposed to going down [pointing to C] and then you're going to have to travel more feet because you're already starting at a lower position. And it might seem easier but then you just have to struggle higher and then have an easier way down because you're going to be tired."
Written Response (Isabel)	 b. Describe how the patterns or features on the map helped you figure out the shape or slope of the land. The lines on the map hed a number on it which wes the height of the mountain in that area. With the numbers going up or down you can see the slope of the Valley *The lines on the map had a number on it which was the height of the mountain in that area. With the numbers going up or down you can see the slope of the valley.*

Based on this finding, I would revise the topographic maps construct map to emphasize a progression in student thinking from a reliance on quantitative elevation information, to an ability to interpret line spacing over two-dimensional lines, to an integrated thought process that takes into account line shapes, spacing, and contour intervals to produce a three-dimensional mental image of a landscape.

6.3 Limitations

This study comprised both a design phase and a data-gathering phase; limitations are present in both. The assessment design for a cognitively-based assessment hinges on the development of a construct map that characterizes student thinking along a predictable learning progression. At the time that I created my construct map, there were gaps in the learning progression literature for these constructs. In consultation with other teachers, I used classroom experience and best practice to fill in these gaps. Nonetheless, there are portions of the construct map that would be strengthened by more thorough research into student learning and cognition around these constructs.

A second limitation stemming from the design of this study is that it addresses only one high school science discipline, Earth Science, and only a subset of Earth Science topics. As we saw in the analysis, the strength and nature of the construct map had a significant impact on the coherence of the assessment and scoring for that construct. This makes it difficult to generalize the results of this study beyond these constructs.

In the implementation phase of the study, the main limitation was the relatively small number of subjects. There were only 22 students who participated. While this provided a large number of data points for each construct on individual assessment items, the holistic scoring reduces these data to one point per student per construct. Therefore, for comparisons of holistic codes and scores, this study had N = 22. A larger subject pool would have allowed for a stronger statistical analysis.

The use of a think-aloud protocol introduced some possible limitations. Requiring subjects to engage in think-aloud can influence their performance on assessment in a few different ways. The task of verbalizing and explicating thought processes constitutes an additional cognitive load that may reduce capacity for problem solving in ways that are more directly related to the assessment items. It is also feasible that the act of thinking aloud helps students reason more thoroughly than they would otherwise. The effects of the think-aloud protocol on individual student performance are likely to vary by student. For example, an NCEO study on the effectiveness of think-aloud studies for evaluating the design of large-scale assessments found that the methodology was useful when conducted with a wide variety of students, including English language learners and students with learning disabilities (Johnstone, Bottsford-Miller, & Thompson, 2006), noting that the benefits outweighed the limitations for all groups except students with cognitive disabilities.

Finally, the subject population was relatively homogenous. All the students in this study attend public school in a very large urban school district. Approximately 90% of them are Black or Latine, and approximately 75% of them identify as female. Although they provided valuable insight into thinking patterns of Earth Science students engaging with a cognitively-based science performance assessment, these patterns may not be representative of those we would observe in other demographic groups.

6.4 Implications

This study points towards the need for more complete research on learning progressions and cognition in geoscience, especially in K-12 contexts. The field has made progress in understanding both the cognition underlying geology thinking and the instructional tools that can support its development (e.g. Nazareth et al., 2019; Ormand et al., 2017), but much of this work has been done at the undergraduate level. A clearer picture of how thinking progresses at the more novice end of the spectrum would, in turn, clarify the design of assessment systems and related curriculum materials intended to support children's geoscience learning. The fourth design principle guiding this cognitively-based science performance assessment states that the task components must be based on research about learning. Based on this requirement alone, which is supported by guidelines for NGSS-aligned assessment (NRC, 2001), further research into geoscience learning is a prerequisite for continued development of cognitively-based assessment systems in this content area.

This work has practical applications across the k-12 educational landscape. Construct maps can be very powerful tools for designing curriculum, instructional materials, and assessment systems. They have the power to transform scholarly learning science research into useful, practical instruments that can be understood and applied by practitioners at all levels of education work. A uniform or universal format would be a good idea: four levels (a common number of levels teachers are used to thinking about for grades, rubrics, etc) in which each level has a known "profile" for the broad type or sophistication of thinking. This would facilitate their usefulness by reducing the amount of "overhead" teachers would need to invest in learning new constructs. It would be necessary to reduce or translate some learning progressions into this format – that is valuable work that could be done by teams of specialists.

Classroom teachers should feel empowered to use the principles of cognitively-based performance assessment in their work. Use cognitive models (construct maps) to design well-aligned curriculum materials *and* assessments. A strong alignment between these two makes both more effective. Teachers are also well suited to defining and refining the outcome space; they have experience with students and can predict or envision what the manifestation of different levels of the construct map might look like in the context of a specific task or prompt. In addition, teacher teams can build libraries of exemplars for student work at different levels for different constructs, thus contributing to our understanding of what that type/level of cognition looks like in real life. What teachers should not be held responsible for is the learning science research and creation of construct maps; that work should be the responsibility of other people in educational spheres (learning scientists, curriculum designers, etc).

However, there is the opportunity to grow teachers' understanding of how student thinking develops within each of these constructs. My recommendation is not intended to exclude teachers from the process of understanding construct-specific student cognition patterns or learning progressions, but to recognize that the creation of construct maps lies beyond the scope of what is reasonable to expect of teachers given their extensive professional responsibilities. It would therefore be appropriate to for schools and districts engaging in the use of cognitively-based curriculum and assessment to consider professional learning opportunities for teachers who are motivated to learn more. Such opportunities could facilitate teachers' thinking around student cognition in specific science domains, and support their ability to use construct maps as tools for day-to-day curriculum and assessment planning on a classroom level.

Curriculum designers and assessment developers should simply *not* create materials that aren't grounded in a cognitive model, yet this still happens frequently. We know that curriculum design and, especially, assessment design is ideally an iterative process: the insight gained from initial rounds of implementation inform revisions to instructional materials and assessment tools. The cognitive model, embodied in a construct map, should also be subject to iterative revisions. Assessment data from pilot programs or studies like this one tell us more than just how effective the item design was. They also provide information about the nature of student thinking in different contexts related to the construct. This information can then be used to clarify individual levels of the construct map underlying the assessment design, and even more importantly, to more precisely describe the differentiation or progression between different levels.

Anyone who uses assessment data to make decisions – from teachers to district leaders to politicians doling out funding – should understand how difficult it is for any assessment, even a cognitively-based assessment, to capture the full complexity of student thinking. Assessment items may be "valid" or "reliable" in a psychometric sense, but that does not mean they are providing a window into a student's mind. *Knowing what Students Know* emphasized the need for an assessment system to be built on a foundational model of cognition. This study demonstrates that, *even when that is the case,* the observational tools and human interpretation introduce a layer of uncertainty, and we should be cautious when making claims about truly "knowing what students know."

Given the difficulties in accurately measuring student thinking and the potential drawbacks of doing so imperfectly, why am I not advocating to get rid of assessment altogether? It is clear to anyone who works in any part of the education sector that formalized systems for assessment are here to stay. The use of assessments in education is pervasive and entrenched, and not without value: they have the potential to give us valuable information about how to better support student learning and guide curriculum development in service of 21st-century science education goals. Given both their promise and their widespread use, it is vital that we ensure our assessment tools are of the highest quality possible. When assessments are used to determine school or district funding, to gatekeep student promotion, or to make other equally high-stakes decisions, it is especially important that they are valid, in the sense that they measure the things they claim to measure. Since the nature of our current education landscape is clear—and it is one that is both shaped by assessments and engenders their continued proliferation—the necessity of efforts to reform and improve assessment is equally clear. We know assessments play a large role in current education policy. To prevent them from being used in unintentionally punitive ways, we need to critically examine the strengths and weaknesses of our current assessment tools, and continue our ongoing efforts to improve them. This is why work around cognitively-based assessment is key to informing the role of assessment in policy. Understanding the challenges around accurately characterizing student thinking and the limitations of assessment systems helps prevent their inappropriate use; efforts to refine assessment tools make it more likely that they will fulfill their intended purpose.

Next steps for this particular line of research can be taken in a variety of directions. To understand more about the nature of this approach to strengthening coherence between student thinking, responses, and scores, we would apply this same approach to more wide-ranging constructs – not just in geoscience or Earth Science, but in any subject area for which cognitively-based performance assessment is feasible. As assessments designed through this approach become more refined, it would be appropriate to investigate their validity and reliability through more traditional psychometric research. To understand more about the application of cognitively-based science performance assessment at different scales, we would build on validity and reliability work to test implementation across schools or districts. My belief is that an assessment system that provides coherent information about student thinking is valuable at *any* scale, including in individual classrooms under the practice of individual teachers. So we should not discount the value of this approach even if it is not immediately scalable to larger contexts. However, because so many high-stakes decisions about funding, programming, and promotion are made based on large-sale assessment, it is equally important to pursue research that can shift those assessments towards coherence with student cognition and classroom instructional activities.

Finally, and of most personal interest to me, we could take steps towards a better understanding of student thinking in geoscience, their expression of that thinking, and the specific ways in which teachers and curriculum designers can support its development. This line of research would bring together questions about learning progressions as well as assessment design in the specific context of Earth Science. As we grapple with climate change and sustainability, there has never been a more important time for Earth Science to take precedence in our efforts as educators to provide meaningful support for the development of student thinking.

References

AAAS. (2007). Getting Assessment Right. *2061 Today*, *17*(1), 3–4.

Abell, S. K., & Lederman, N. G. (2007). *Handbook of research on science education*. (S. K. Abell & N. G. Lederman, Eds.) *Science Education* (Vol. 20, p. 1344). Mahwah, NJ: Erlbaum. doi:10.1007/s11191-010-9294-3

Anderson, D. L. (2006). Plate tectonics; the general theory: Complex Earth is simpler than you think. In C. A. Manduca & D. W. Mogk (Eds.), *Earth and Mind How Geologists Think and Learn about the Earth* (pp. 29–38). Boulder, CO: Geological Society of America.

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. (L. W. Anderson & D. R. Krathwohl, Eds.)*Theory Into Practice* (Vol. Complete e, p. xxix, 352 p.). Longman. doi:10.1207/s15430421tip4104_2

Baxter, G. P., & Glaser, R. (1998). Investigating the Cognitive complexity of Science Assessments. *Educational Measurement: Issues and Practice*.

Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1990). Evaluation of Procedure-Based Scoring for Hands-On Science Assessment, *29*(1), 1–17.

Bell, B. (2007). Classroom Assessment of Science Learning. In S. K. Abell & N. G. Ledermen (Eds.), *Handbook of Research on Science Education* (pp. 965 – 1006). Mahwah, NJ: Erlbaum.

Biggs, J., & Collis, K. (1982). *Evaluating the Quality of Learning: the SOLO taxonomy*. New York: Academic Press.

Black, P. (1998). Assessment by teachers and the improvement of students' learning. In B. J. Fraser & K. Tobin (Eds.), *International handbook of science education* (pp. 811–822). Great Britain: Kluwer Academic Press.

Black, P., & Wiliam, D. (1998). Inside the Black Box : Raising Standards Through Classroom Assessment. *Phi Delta Kappan*, *80*(2), 139–148. doi:10.1002/hrm

Black, P., & Wiliam, D. (2005). Developing the theory of formative assessment. (J. Gardiner, Ed.)*Educational Assessment Evaluation and Accountability*, *21*(1), 5–31. Retrieved from http://eprints.ioe.ac.uk/1119/

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

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, *4*, 87–100.

Broadfoot, P., & Black, P. (2004). Redefining assessment? The first ten years of assessment in education. *Assessment in Education. Principles, Policy & Practice*, *11*(1), 7–26. doi:10.1080/0969594042000208976

Bybee, R., McCrae, B., & Laurie, R. (2009). PISA 2006: An assessment of scientific literacy. *Journal of Research in Science Teaching,* 46(8), 865-883.

Chi, M., Glaser, R., & Farr, M. (1988). *The Nature of Expertise*. Hillsdale, NJ: Erlbaum.

Clark, L. A., & Watson, D. (1995). Constructing Validity : Basic Issues in Objective Scale Development The Centrality of Psychological Measurement, *7*(3), 309–319.

Committee on Assessment in Support of Instruction and Learning, & Committee on Science Education K-12. (2003). *Assessment in Support of and learning: Bridging the gap between large-scale and classroom assessment*.

Cooper, B. C., Shepardson, D. P., & Harber, J. M. (2002). Assessments as Teaching and Research Tools in an Environmental Problem-Solving Program for In-Service. *Journal of Geoscience Education*, *50*, 64–71.

Cronbach, L. J., & Meehl, P. E. (1955). CONSTRUCT VALIDITY IN PSYCHOLOGICAL TESTS Lee J. Cronbach and Paul E. Meehl (1955) [1], 281–302.

Cronin, J., Dahlin, M., Adkins, D., Kingsbury, G. G., Finn, C. E., & Petrilli, M. J. (2007). October 2007, (October).

Daley, B. J., Watkins, K., Williams, S. W., Courtenay, B., Davis, M., & Dymock, D. (2001). Exploring Learning in a Technology-Enhanced Environment Studies of Learning Results in Technology-Enhanced Environments, *4*(3), 126–138.

Darling-Hammond, L. (1993). Setting Standards for Students: The Case for Authentic Assessment. *NASSP Bulletin*, *77*(556), 18–26. doi:10.1177/019263659307755604

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality Control in the Development and Use of Performance Assessments. *Applied Measurement in Education*, *4*(4), 289–303.

Duncan, R.G. & Rivet, A. (2013, January 25). Science learning progressions. *Science*, 339.

Duncan, R. G., Rogat, A. D., & Yarden, A. (2009). A learning progression for deepening students' understandings of modern genetics across the 5th-10th grades. *Journal of Research in Science Teaching*, 46(6), 655-674.

Embretson, S., & Gorin, J. (2001). Improving Construct Validity with Cognitive Psychology Principles Improving Construct Validity With Cognitive Psychology Principles. *Journal of Educational Measurement*, *38*(4), 343–368.

Ericsson, K. A., & Simon, H. A. (1993). Inferences from Verbal Data. *Protocol Analysis* (pp. 169–220). MIT Press.

Ferrara, S., & Institutes, A. (2006). Toward a Psychology of Large-Scale Educational Achievement Testing : Some. *Educational Measurement: Issues and Practice*.

Furtak, E. M., Morrison, D. M., Iverson, H., Ross, M., & Heredia, S. C. (2011). A conceptual analysis of the conceptual inventory of natural selection: Improving diagnostic utility through within-item analysis. Paper presented at the annual meeting of the National Association of Research in Science Teaching Annual Meeting, Orlando, FL.

Frisbie, D. a. (2005). Measurement 101: Some Fundamentals Revisited. *Educational Measurement: Issues and Practice*, *24*(3), 21–28. doi:10.1111/j.1745-3992.2005.00016.x

Gewertz, C. (2018, April 4). ESSA Offers Testing Flexibility. So Why Aren't States Using It? *Education Week*, *37*(25), 21.

Gobert, J. D. (2005). The Effects of Different learning Tasks on Model-building in Plate Tectonics: Diagramming Versus Explaining. *Journal of Geoscience Education*, *53*(4), 444–455.

Goertz, M., & Duffy, M. (2003). Mapping the landscape of high-stakes testing and accountability programs. *Theory into Practice*, *42*(1), 4–11.

Gorin, J. S. (2007). Test Design with Cognition in Mind. *Educational Measurement: Issues and Practice*, *25*(4), 21–35. doi:10.1111/j.1745-3992.2006.00076.x

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, 843–868. doi:10.1002/tea.v49.7

Guskey, T. R. (2005). Multiple Sources of Evidence : An Analysis of Stakeholders ' Perceptions of Various Indicators of Student Learning. *Educational Measurement Issues and Practice*, *26*(1), 19–27. doi:10.1111/j.1745-3992.2007.00085.x

Harlen, W. (2006). *Teaching, Learning, and Assessing Science 5-12* (4th ed., p. 264). Sage Publications Ltd.

Herbert, B. E. (2006). Student understanding of complex earth systems. In C. A. Manduca & D. W. Mogk (Eds.), *Earth and Mind How Geologists Think and Learn about the Earth* (pp. 95 – 104). Boulder, CO: Geological Society of America.

Herman, J., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Hussain. (2001). *Physics Performance Task*. State University of New York, Buffalo.

Johnstone, C.J., Bottsford-Miller, N.A., & Thompson, S.J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners* (Technical Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Jones, P., Carr, J. F., & Ataya, R. L. (Eds.). (2007). *A pig don't get fatter the more you weigh it: Classroom assessments that work*. New York: Teachers College Press.

Joy Cumming, J., & Maxwell, G. S. (1999). Contextualizing Authentic Assessment. *Assessment in Education. Principles, Policy & Practice*, *6*(2), 177–194. doi:10.1080/09695949992865

Kastens, K. A., & Ishikawa, T. (2006). Spatial thinking in the geosciences and cognitive sciences: A cross-disciplinary look at the intersection of the two fields. (C. Manduca & D. Mogk, Eds.)*Earth and Mind How Geologists Think and Learn about the Earth*, *2413*(2006), 53–76. doi:10.1130/2006.2413(05).

Klassen, S. (2006). Contextual assessment in science education: Background, issues, and policy. *Science Education*, *90*(5), 820–851. doi:10.1002/sce.20150

Kohn, A. (2004). *What does it mean to be well educated? And more essays on standards, grading, and other follies.* Boston: Beacon Press.

Koretz, D. M. (2009). *Measuring Up: what educational testing really tells us* (p. 368). Cambridge: Harvard University Press.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. (R. Pea & J. S. Brown, Eds.)*Learning in doing* (Vol. 95, p. 138). Cambridge University Press. doi:10.2307/2804509

Leighton, J. P., & Gierl, M. J. (2004). Defining and Evaluating Models of Cognition Used in Educational Measurement to Make Inferences About Examinees ' Thinking Processes, 3–16.

Lai, Kobrin, J. L., DiCerbo, K. E., & Holland, L. R. (2017). Tracing the Assessment Triangle With Learning Progression-Aligned Assessments in Mathematics. Measurement (Mahwah, N.J.), 15(3-4), 143–162. https://doi.org/10.1080/15366367.2017.1388113

Li, J., Klahr, D., & Siler, S. (2006). What lies beneath the science achievement gap: The challenges of aligning science instruction with standards and tests. *Science Education*, *15*(1), 1–12.

Li, M., & Shavelson, R. (2001). Examining the linkage between science achievement and assessment. *Annual Meeting of the American Educational Research Association*. Seattle, WA.

Linn, Robert L. (1994). Criterion-Referenced Measurement: A Valuable Perspective Clouded by Surplus Meaning. *Educational Measurement: Issues and Practice*.

Linn, R. L. (2000). "Assessments and Accountability." Educational Researcher, 23(9), 4-14.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of the requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, *31*(6), 3–16. doi:10.3102/0013189X031006003

Linn, Robert L, Baker, E. L. V. A. L., Dunbar, S. B., & Dunbar, B. (2007). Assessment : Expectations and Validation Criteria. *Educational Researcher*, *20*(8), 15–21. doi:10.3102/0013189X020008015

Linn, Robert L., & Gronlund, N. E. (2000). *Measurement and Assessment in Teaching* (8th ed.). Des Moines, IA: Prentice Hall.

Lyons, C. (2014). *Relationships between Conceptual Knowledge and Reasoning about Systems: Implications for Fostering Systems Thinking in Secondary Science*.(Unpublished doctoral dissertation). Columbia University, New York.

Manduca, C. A., & Mogk, D. W. (Eds.). (2006). *Earth and Mind: How Geologists think and learn about the earth*.

Marion, S. F. (2006). A validity framework for Evaluating the Technical Quality of Alternate Assessments. *Educational Measurement: Issues and Practice*.

Mazzeo, John & Von Davier, Matthias. (2009). Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results. *Education Working Papers EDU/PISA/GB (2008*). 28. 23-24.

McClure, J. R., Sonak, B., & Suen, H. K. (1999). Concept map assessment of classroom learning: Reliability, validity, and logistical practicality. *Journal of Research in Science Teaching*, *36*(4), 475–492. doi:10.1002/(SICI)1098-2736(199904)36:4<475::AID-TEA5>3.0.CO;2-O

Mislevy, R., & Baxter, G. P. (2005). *The Case for an Integrated Design Framework for Assessing Science Inquiry*.

Mislevy, R. J. (2004). *The Case for an Integrated Design Framework for Assessing Science Inquiry* (Vol. 1522).

Mislevy, R. J., Haertel, G. D., & International, S. R. I. (2006). Implications of Evidence-Centered Design for Educational Testing. *Educational Measurement: Issues and Practice*.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 Assessment Frameworks*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2019/frameworks/

Nagy, G., Nagengast, B., Frey, A., Becker, M., & Rose, N. (2018). A multilevel study of position effects in PISA achievement tests: Student- and school-level predictors in the German tracked school system. *Assessment in Education: Principles, Policy & Practice*.

National Research Council. (1996). *National Science Education Standards*. Washington, DC: National Academies Press.

National Research Council. (2001). *Knowing What Students Know: the science and practice of educational assessment*. (J. W. Pellegrino, N. Chudowsky, & R. Glaser, Eds.) (p. 382).

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. (H. Quinn, H. Schweingruber, & T. Keller, Eds.)*Social Sciences* (Vol. Chapter 10, pp. 1–6). The National Academies Press. Retrieved from http://www.nap.edu/catalog.php?record_id=13165

National Research Council. (2013). *Next Generation Science Standards*. Washington, DC: National Academies Press.

National Research Council. (2014). Developing Assessments for the Next Generation Science Standards. Committee on Developing Assessments of Science Proficiency in K-12. Board on Testing and Assessment and Board on Science Education, James W. Pellegrino, Mark R. Wilson, Judith A. Koenig, and Alexandra S. Beatty, Editors. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Nazareth, A., Newcombe, N.S., Shipley, Thomas F ; Velazquez, Mia ; Weisberg, Steven M. (2019). Beyond small-scale spatial skills: Navigation skills and geoscience education. *Cognitive research: principles and implications*, 2019-12, 4(1), 1-17.

NGSS Lead States. (2013). *Next Geneartion Science Standards: For States, By States.* Washington, DC: The National Academies Press.

Nichols, P., & Wisconsin-Milwaukee, U. (1999). The Lack of Fidelity Between Cognitively Complex Constructs and Conventional Test Development Practice. *Educational Measurement: Issues and Practice*.

NYSED. (2001). Physical Setting / Earth Science: Core Curriculum.

OECD (2019), *PISA 2018 Assessment and Analytical Framework*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/b25efab8-en.

Opfer, Nehm, R. H., & Ha, M. (2012). Cognitive foundations for science assessment design: Knowing what students know about evolution. *Journal of Research in Science Teaching*, 49(6), 744–777. https://doi.org/10.1002/tea.21028

O'Neil, T., Sireci, S. G., & Huff, K. F. (2004). Evaluating the consistency of test content across two successive administrations of a state-mandated science assessment. *Educational Assessment*, *9*(3-4), 129–151.

O'Reilly, T., & McNamara, D. (2007). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional "high stakes" measures of high school students' science achievement. *American Educational Research Journal*, 44, 161–196.

Ormand, C. et al. (2017) The Spatial Thinking Workbook: A Research-Validated Spatial Skills Curriculum for Geology Majors, *Journal of Geoscience Education*, 65(4), 423-434.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "Two Disciplines " Problem : Linking Theories of Cognition and Learning with Assessment and Instructional Practice Chapter 9. *Review of Research in Education*, *24*, 307–353.

Popham, W. J. (2006). Assessment for Learning: An endangered species? *Educational Leadership*, *63*(5), 82–83.

Raia, F. (2005). Students' understanding of complex dynamic systems. *Journal of Geoscience Education*, *53*(3), 297-308.

Rapp, D. N., & Uttal, D. H. (2006). Understanding and Enhancing Visualizations: two models of collaboration between earth science and cognitive science. *Geological Society of America*.

Rice, D. C., Ryan, J. M., & Samson, S. M. (1998). Using concept maps to assess student learning in the science classroom: Must different methods compete? *Journal of Research in Science Teaching*, *35*(10), 1103–1127. doi:10.1002/(SICI)1098-2736(199812)35:10<1103::AID-TEA4>3.0.CO;2-P

Rothman, R. (1995). *Measuring Up: Standards, assessment, and school reform*. San Francisco: Jossey-Bass.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the Stability of Performance Assessments. *Journal of Educational Measurement*, *30*(1), 41–53. doi:10.1111/j.1745-3984.1993.tb00421.x

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching*, *33*(6), 569–600. doi:10.1002/(SICI)1098-2736(199608)33:6<569::AID-TEA1>3.0.CO;2-M

Salomon, G., & Perkins, D. N. (1989). Rocky Roads to Transfer : Rethinking Mechanisms of a Neglected Phenomenon. *Educational Psychologist*, *24*(2), 113–142.

Schneider, M. C., & Andrade, H. (2013). Teachers' and administrators' use of evidence of student learning to take action. [Special issue]. *Applied Measurement in Education*, 26(3). doi:10.1080/08957347.2013.793189

Shavelson, R. J. (1997). Development of Performance Assessments in Science : Conceptual , Practical, and Logistical Issues. *Educational Measurement: Issues and Practice*.

Shin, N., Stevens, S. Y., & Krajcik, J. (n.d.). Tracking student learning over time using construct-centered design.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of Research on Children 's Learning for Standards and Assessment : A Proposed Learning Progression for Matter and the Atomic-Molecular Theory. *Measurement*, *14*(1), 1–98.

Sparks, S. D. (2019, March 13). Testing; "A Consumer's Guide to Testing Under the Every Student Succeeds Act: What Can the Common Core and Other ESSA Assessments Tell Us?" *Education Week*, *38*(25), 5.

Stern, L., & Ahlgren, A. (2002). Analysis of students' assessments in middle school curriculum materials: Aiming precisely at benchmarks and standards. *Journal of Research in Science Teaching*, *39*(9), 889–910. doi:10.1002/tea.10050

Tamir, P. (1998). Assessment and evaluation in science education: Opportunities to learn and outcomes. In B. J. Fraser & K. G. Tobin (Eds.), *International handbook of science education* (pp. 761–789). Great Britain: Kluwer Academic Press.

Todd, A., Romine, W. L., & Cook Whitt, K. (2017). Development and Validation of the Learning Progression–Based Assessment of Modern Genetics in a High School Context. *Science Education*, 101(1), 32–65. https://doi.org/10.1002/sce.21252

Treagust, D. F., Jacobowitz, R., Gallagher, J. L., & Parker, J. (2001). Using assessment as a guide in teaching for understanding: A case study of a middle school science class learning about sound. *Science Education*, *85*(2), 137–157.

Uchinella. (2002). *Determination of Validity and Reliability of performance assessment tasks developed for selected topics in high school chemistry*. State University of New York Buffalo.

Ujifusa, A. (2019). In Drive to Revamp Tests, Some States in Pilot's Seat: Louisiana and New Hampshire have stepped up for ESSA's experiment in crafting new student assessments. *Education Week*, 38(27), 14–15.

Vygotskiĭ, L.S., & Cole, M. (1978). *Mind in society: the development of higher psychological processes.* Harvard University Press.

Wiggins, G. (1989). Teaching to the (Authentic) Test. *Educational Leadership*, *46*(7), 41–47.

Wiggins, Grant, & McTighe, J. (1998). *Understanding by design*. *Design* (Vol. 6066, pp. 1–16). Association for Supervision and Curriculum Development.

Wilmot, D. B., & Champney, D. (2008). Assessing Progress toward college readiness with cognitive and psychometric models of student learning in mathematics. *American Educational Research Association*. New York.

Wilmot, D. B., Champney, D., Wilson, M., Schoenfeld, A., & Zahner, W. (2009). Using cognitive and psychometric models of student learning in mathematics to validate a measure of college readiness. *American Educational Research Association*. San Diego, CA.

Wilmot, D. B., Schoenfeld, A., Wilson, M., Champney, D., Zahner, W., & Zabner, W. (2011). Validating a Learning Progression in Mathematical Functions for College Readiness. *Mathematical Thinking and Learning*, *13*(4), 259–291.

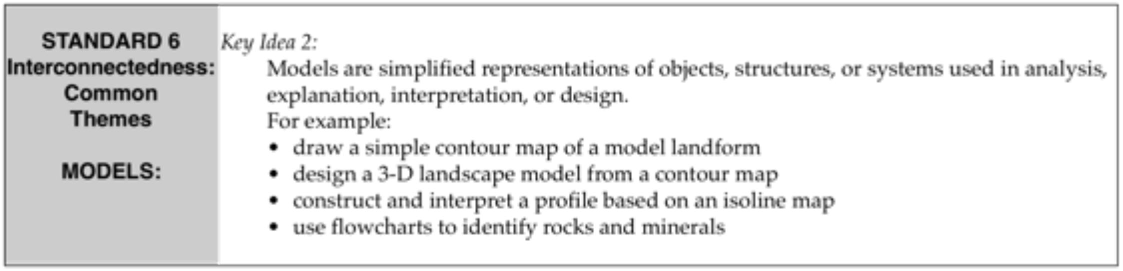
Wilson, M. (2005). *Constructing measures: An item response modeling approach*. *Constructing measures an item response modeling approach* (p. 228). Lawrence Erlbaum Associates.

Wiser, M., Smith, C. L., & Doubler, S. (2012). Learning progressions as tool for curriculum development. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progression in science* (pp. 359-403). New York, NY: Sense Publishers.

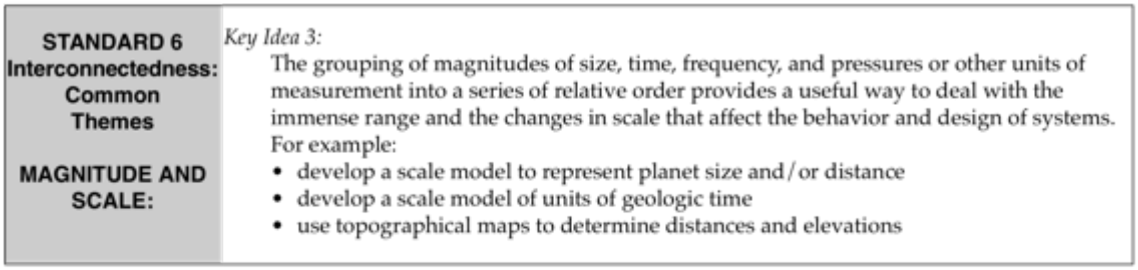
Wright. (2002). *Development of Performance Tasks, an alternative assessment for New York State Regents Biology Courses*. State University of New York Buffalo.

Yeh, S. S. (2006). Reforming Federal Testing Policy to Support Teaching and Learning. *Educational Policy*, *20*(3), 495–524.

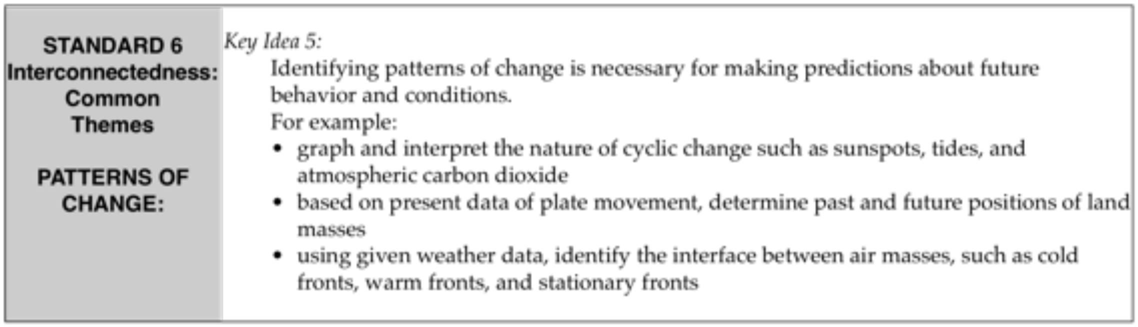
Appendix A: Key Ideas from the New York State Earth Science Core Curriculum



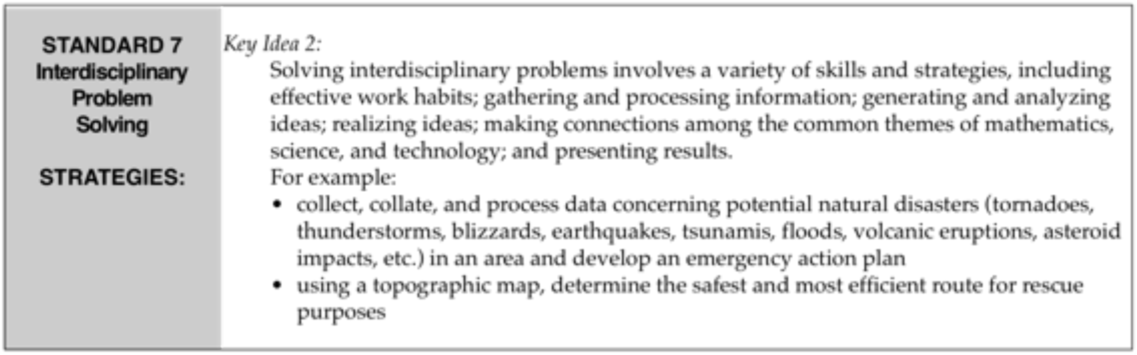
STANDARD 6 **Interconnectedness: Common Themes** **MODELS:**	*Key Idea 2:* Models are simplified representations of objects, structures, or systems used in analysis, explanation, interpretation, or design. For example: • draw a simple contour map of a model landform • design a 3-D landscape model from a contour map • construct and interpret a profile based on an isoline map • use flowcharts to identify rocks and minerals



STANDARD 6 **Interconnectedness: Common Themes** **MAGNITUDE AND SCALE:**	*Key Idea 3:* The grouping of magnitudes of size, time, frequency, and pressures or other units of measurement into a series of relative order provides a useful way to deal with the immense range and the changes in scale that affect the behavior and design of systems. For example: • develop a scale model to represent planet size and/or distance • develop a scale model of units of geologic time • use topographical maps to determine distances and elevations



STANDARD 6 **Interconnectedness: Common Themes** **PATTERNS OF CHANGE:**	*Key Idea 5:* Identifying patterns of change is necessary for making predictions about future behavior and conditions. For example: • graph and interpret the nature of cyclic change such as sunspots, tides, and atmospheric carbon dioxide • based on present data of plate movement, determine past and future positions of land masses • using given weather data, identify the interface between air masses, such as cold fronts, warm fronts, and stationary fronts



STANDARD 7 **Interdisciplinary Problem Solving** **STRATEGIES:**	*Key Idea 2:* Solving interdisciplinary problems involves a variety of skills and strategies, including effective work habits; gathering and processing information; generating and analyzing ideas; realizing ideas; making connections among the common themes of mathematics, science, and technology; and presenting results. For example: • collect, collate, and process data concerning potential natural disasters (tornadoes, thunderstorms, blizzards, earthquakes, tsunamis, floods, volcanic eruptions, asteroid impacts, etc.) in an area and develop an emergency action plan • using a topographic map, determine the safest and most efficient route for rescue purposes

PERFORMANCE INDICATOR 1.2 Describe current theories about the origin of the universe and solar system.

1.2i The pattern of evolution of life-forms on Earth is at least partially preserved in the rock record.

- Fossil evidence indicates that a wide variety of life-forms has existed in the past and that most of these forms have become extinct.
- Human existence has been very brief compared to the expanse of geologic time.

1.2j Geologic history can be reconstructed by observing sequences of rock types and fossils to correlate bedrock at various locations.

- The characteristics of rocks indicate the processes by which they formed and the environments in which these processes took place.
- Fossils preserved in rocks provide information about past environmental conditions.
- Geologists have divided Earth history into time units based upon the fossil record.
- Age relationships among bodies of rocks can be determined using principles of original horizontality, superposition, inclusions, cross-cutting relationships, contact metamorphism, and unconformities. The presence of volcanic ash layers, index fossils, and meteoritic debris can provide additional information.
- The regular rate of nuclear decay (half-life time period) of radioactive isotopes allows geologists to determine the absolute age of materials found in some rocks.

PERFORMANCE INDICATOR 2.1 Use the concepts of density and heat energy to explain observations of weather patterns, seasonal changes, and the movements of Earth's plates.

2.1k The outward transfer of Earth's internal heat drives convective circulation in the mantle that moves the lithospheric plates comprising Earth's surface.

2.1l The lithosphere consists of separate plates that ride on the more fluid asthenosphere and move slowly in relationship to one another, creating convergent, divergent, and transform plate boundaries. These motions indicate Earth is a dynamic geologic system.

- These plate boundaries are the sites of most earthquakes, volcanoes, and young mountain ranges.
- Compared to continental crust, ocean crust is thinner and denser. New ocean crust continues to form at mid-ocean ridges.
- Earthquakes and volcanoes present geologic hazards to humans. Loss of property, personal injury, and loss of life can be reduced by effective emergency preparedness.

2.1n Many of Earth's surface features such as mid-ocean ridges/rifts, trenches/subduction zones/island arcs, mountain ranges (folded, faulted, and volcanic), hot spots, and the magnetic and age patterns in surface bedrock are a consequence of forces associated with plate motion and interaction.

2.1p Landforms are the result of the interaction of tectonic forces and the processes of weathering, erosion, and deposition.

2.1q Topographic maps represent landforms through the use of contour lines that are isolines connecting points of equal elevation. Gradients and profiles can be determined from changes in elevation over a given distance.

2.1r Climate variations, structure, and characteristics of bedrock influence the development of landscape features including mountains, plateaus, plains, valleys, ridges, escarpments, and stream drainage patterns.

2.1s Weathering is the physical and chemical breakdown of rocks at or near Earth's surface. Soils are the result of weathering and biological activity over long periods of time.

2.1t Natural agents of erosion, generally driven by gravity, remove, transport, and deposit weathered rock particles. Each agent of erosion produces distinctive changes in the material that it transports and creates characteristic surface features and landscapes. In certain erosional situations, loss of property, personal injury, and loss of life can be reduced by effective emergency preparedness.

2.1u The natural agents of erosion include:

- *Streams (running water):* Gradient, discharge, and channel shape influence a stream's velocity and the erosion and deposition of sediments. Sediments transported by streams tend to become rounded as a result of abrasion. Stream features include V-shaped valleys, deltas, flood plains, and meanders. A watershed is the area drained by a stream and its tributaries.
- *Glaciers (moving ice):* Glacial erosional processes include the formation of U-shaped valleys, parallel scratches, and grooves in bedrock. Glacial features include moraines, drumlins, kettle lakes, finger lakes, and outwash plains.
- *Wave Action:* Erosion and deposition cause changes in shoreline features, including beaches, sandbars, and barrier islands. Wave action rounds sediments as a result of abrasion. Waves approaching a shoreline move sand parallel to the shore within the zone of breaking waves.
- *Wind:* Erosion of sediments by wind is most common in arid climates and along shorelines. Wind-generated features include dunes and sand-blasted bedrock.
- *Mass Movement:* Earth materials move downslope under the influence of gravity.

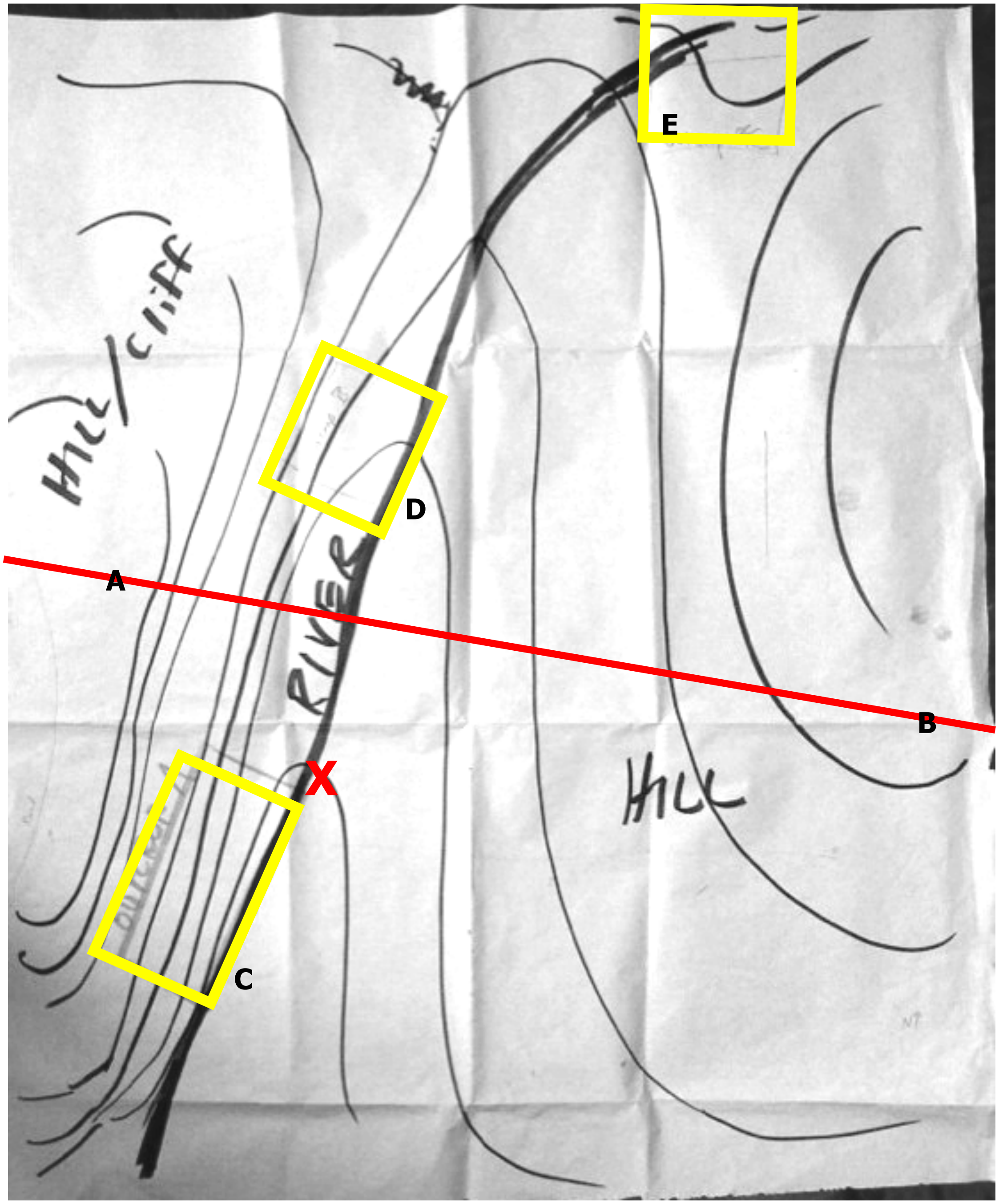
2.1w Sediments of inorganic and organic origin often accumulate in depositional environments. Sedimentary rocks form when sediments are compacted and/or cemented after burial or as the result of chemical precipitation from seawater.

PERFORMANCE INDICATOR 3.1 Explain the properties of materials in terms of the arrangement and properties of the atoms that compose them.

3.1c Rocks are usually composed of one or more minerals.

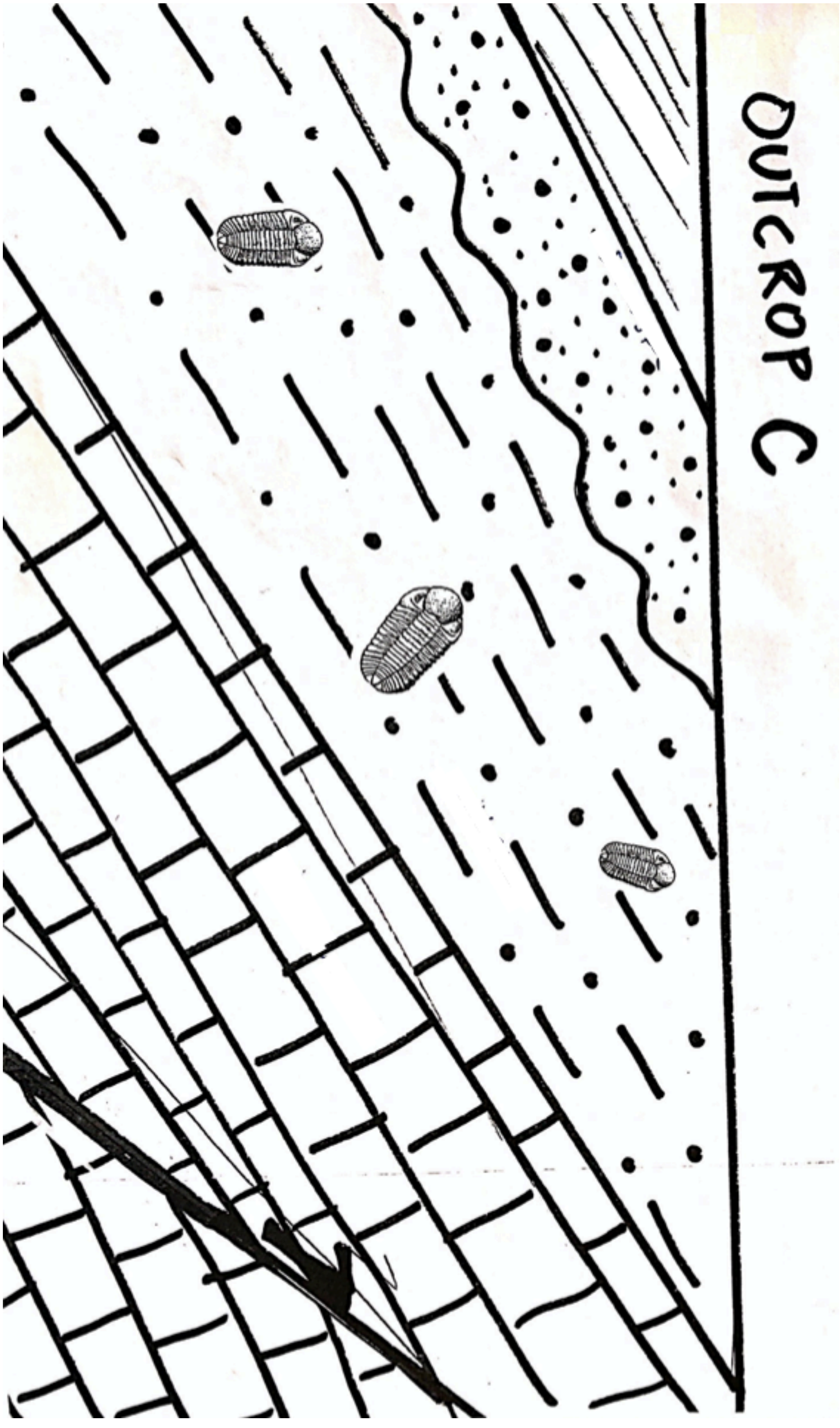
- Rocks are classified by their origin, mineral content, and texture.
- Conditions that existed when a rock formed can be inferred from the rock's mineral content and texture.
- The properties of rocks determine how they are used and also influence land usage by humans.

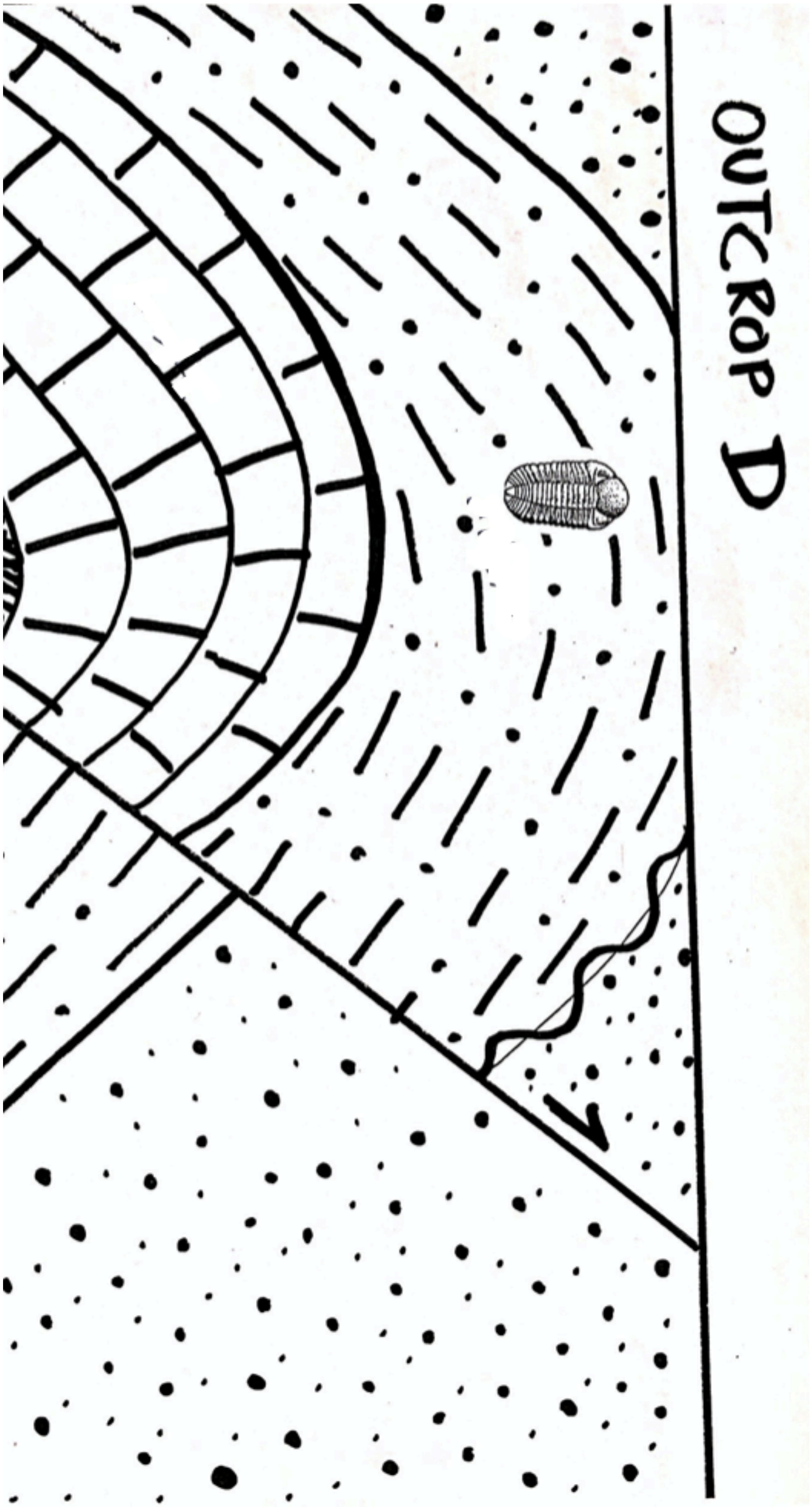
Appendix B: Cognitively-Based Performance Assessment Task Materials

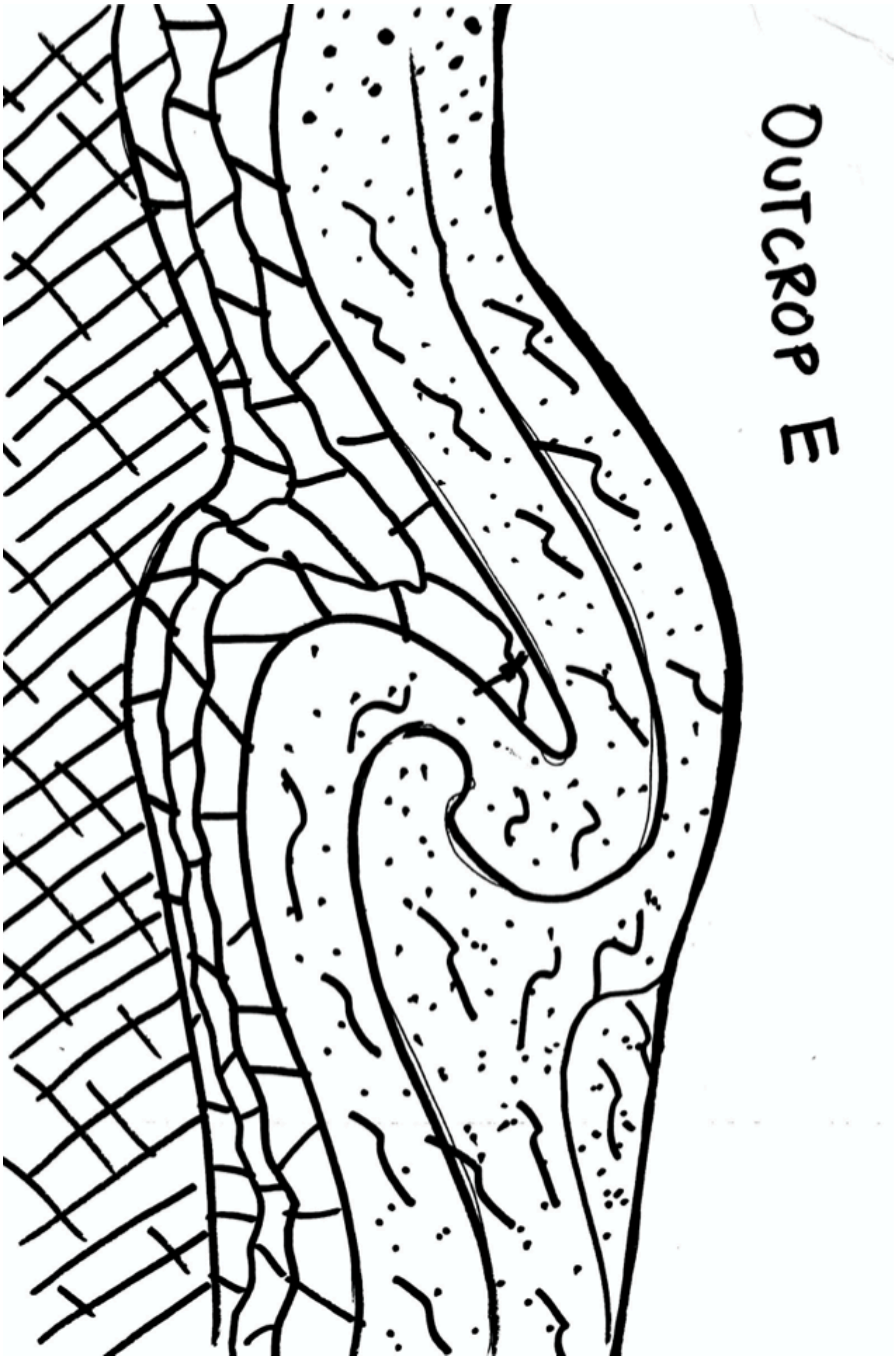


Yellow boxes show locations of outcrops (see sketches below)
Red line is reference line for topographic profile as seen from point X

Contour interval of map = 20 feet
Map will be sized so that the length of line AB is the same width as the line in Part 1 of the student response form (about 7 inches; the map will be 8.5 x 11)


OUTCROP C

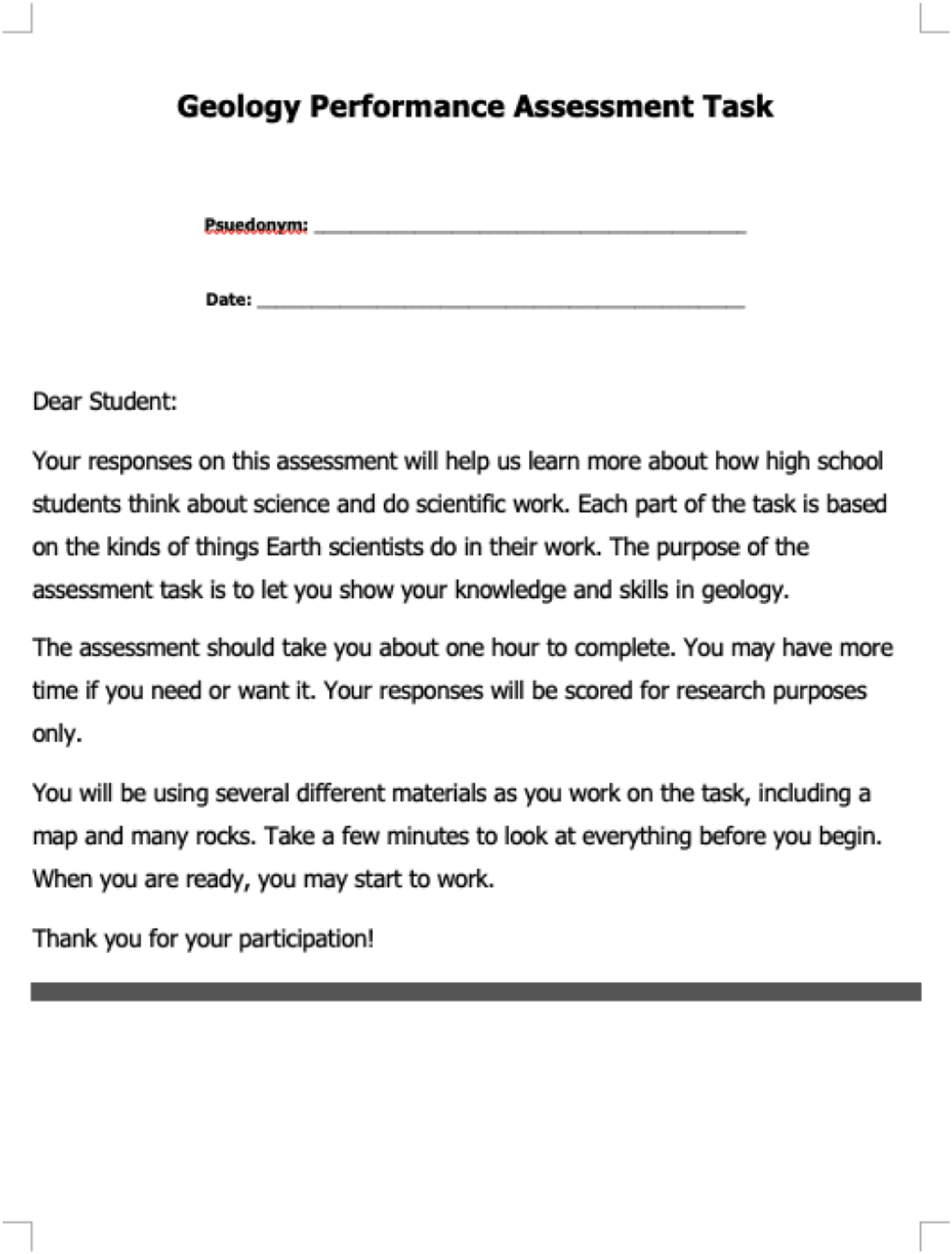

OUTCROP D


OUTCROP E

List of Additional Materials

- Several hand samples of rocks that correspond to visible outcrop layers
    - EITHER sandstone or limestone
    - Daughter regional metamorphic rock of selected sedimentary rock (either quartzite the same color as sandstone, or marble)
    - 1 fossiliferous siltstone containing fossils characteristic of the late Devonian period (ideally Phacops or other trilobite)
    - 1 rounded igneous (preferably granite) (weathering by transport in water), with no connection to local geology.
- Copy of the Earth Science Reference Tables
- Pencil and eraser
- Metric 30 cm ruler
- Four-function calculator

Appendix C: Student Response Form for Performance Task



Geology Performance Assessment Task

Psuedonym: ____

Date: ____

Dear Student:

Your responses on this assessment will help us learn more about how high school students think about science and do scientific work. Each part of the task is based on the kinds of things Earth scientists do in their work. The purpose of the assessment task is to let you show your knowledge and skills in geology.

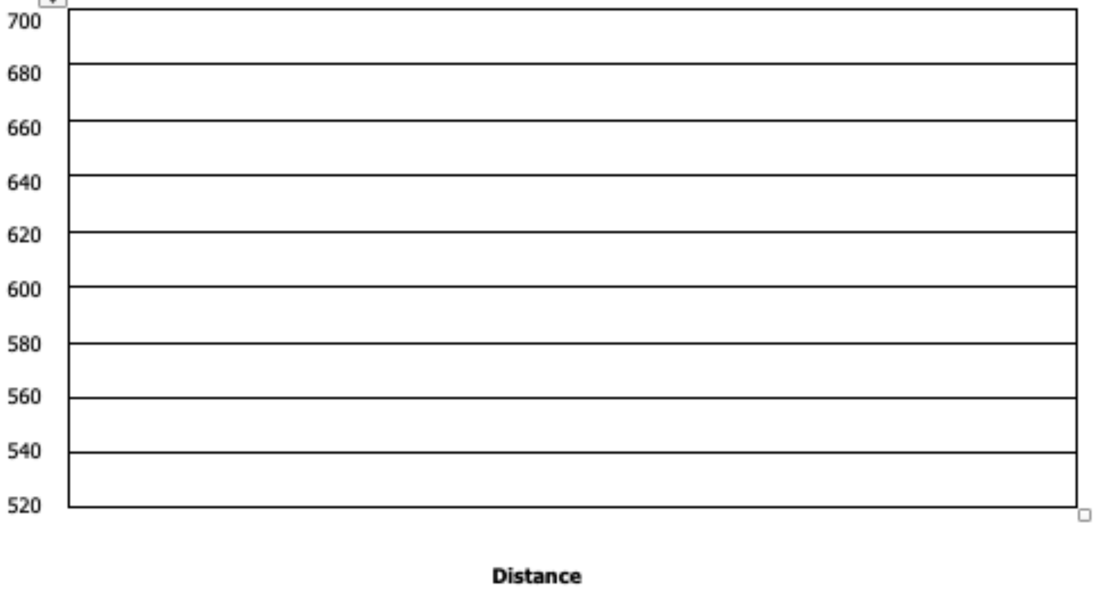
The assessment should take you about one hour to complete. You may have more time if you need or want it. Your responses will be scored for research purposes only.

You will be using several different materials as you work on the task, including a map and many rocks. Take a few minutes to look at everything before you begin. When you are ready, you may start to work.

Thank you for your participation!

Part 1: Mapping

a. If you were standing at the point marked with an **X** on the map, what would the landscape in front of you look like? In the space below, draw a topographic profile that shows the shape of the land between point **A** and point **B**. A grid has been provided for you to use however you see fit.



b. Describe how the patterns or features on the map helped you figure out the shape or slope of the land.

The area on this map contains a number of interesting outcrops and fossils. The areas with exposed outcrops are marked with colored boxes on the map. If you were going to walk around to observe each one, what path would you take? Assume you are starting at point **A**, where there is access to a road.

c. *On the map,* draw the path you would follow. Your path should go near each of the three outcrops.

d. Make notes along the line that you drew explaining the path that you chose. Imagine you are guiding someone else along the path.

e. What were the most important factors that determined your path? Briefly explain your reasoning:

Part 2: Surface Processes

The river running through this region has a big effect on the local landscape.

a. In part 1, you drew a profile showing the shape of the landscape. Explain how or why the river running through this region affects the shape you drew.

____

____

____

____

____

____

b. Several rocks were collected in the region shown on the map. These rocks are labeled with numbers one through four. Look at each of the four rock samples. Which rock number was probably found in the river?

Rock # __________

c. Explain why you chose this rock. In your explanation, describe the physical features of the rock(s) that helped you decide, and what those features tell you about the rock(s).

____

____

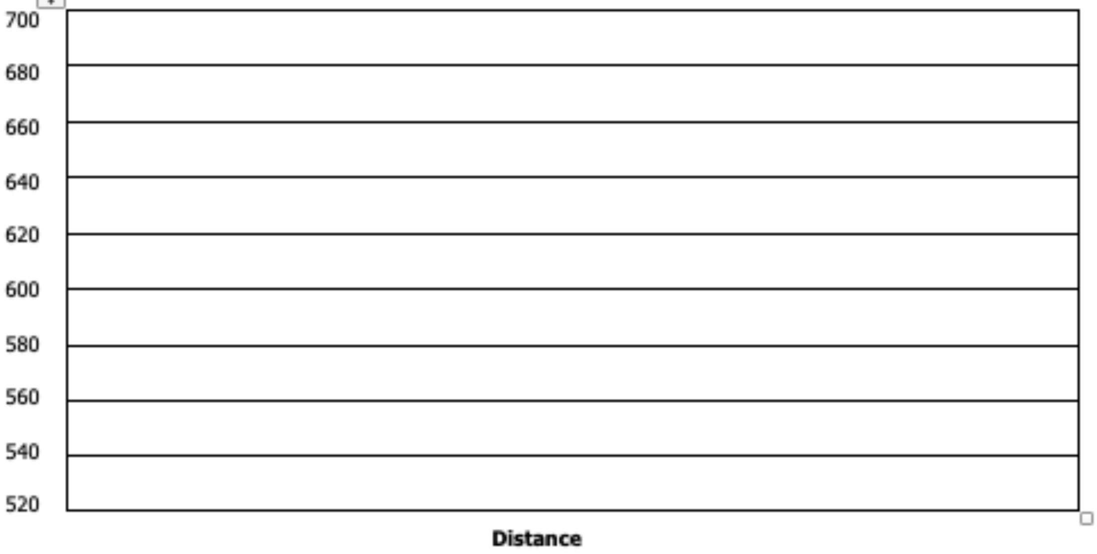
____

____

____

____

d. On the previous page you drew a profile of the landscape as seen from point **X**. What would this same view look like if the river had not formed here? Make a sketch below:



e. Complete the chart below comparing the difference(s) between your two drawings. In the left column, describe each major difference you can see between the shape of the profile with the river and the shape of the profile without the river. For each of these differences, use the right column to explain how it was caused by the running water. You do not need to use all the rows in the chart.

Difference between Profiles	**Explanation:** What effect did the river's presence have?

Part 3: Geologic History

a. Complete the table below showing the type of each rock sample. Include as much detail as you can, including the rock's classification, name, or composition. Then briefly describe what you know about the environment in which each rock formed.

Rock #	What kind of rock is it?	Where / how did this rock form?
1		
2		
3		
4		

b. Rock #2 contains fossil remains of an organism that lived on Earth in the past. Based on this fossil evidence, what do you know about the time period that the rock formed? Be as specific as possible in your answer.

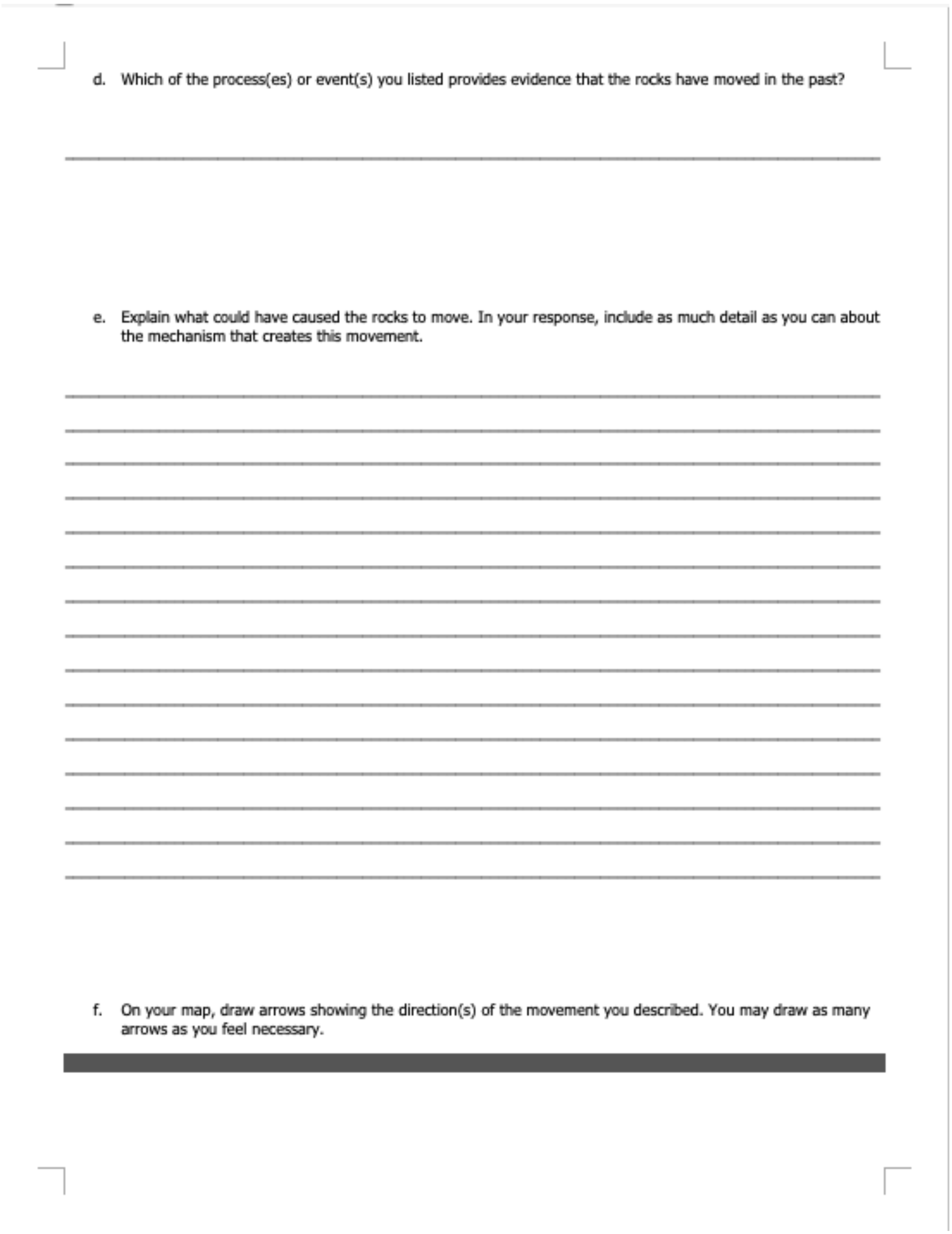
____

____

____

c. There are 3 pictures of outcrops from locations marked on the map. These are places where the rocks are visible from the side. Based on the diagrams of outcrop C, outcrop D, and outcrop E, list the geologic events in the order that they occurred to create this landscape. Begin with the oldest. Include a short justification for each event (how do you know it occurred?).

Geologic Event	How do you know?



d. Which of the process(es) or event(s) you listed provides evidence that the rocks have moved in the past?

e. Explain what could have caused the rocks to move. In your response, include as much detail as you can about the mechanism that creates this movement.

f. On your map, draw arrows showing the direction(s) of the movement you described. You may draw as many arrows as you feel necessary.

Part 4: The Big Picture

In Part 3, you described changes that took place to this region in the past. Now you are going to show how the local landscape looked at various points over time. For each of the following point in time, write a one- or two-sentence description of what you think the environment looked like at that time. Your description may include information about the type of environment, organisms that lived there at the time, the shape of the land's surface, and/or anything else you think is important. After you write your description, make a sketch that shows the important components of the local landscape.

Written Description	Illustration
a. At the time of the earliest geologic event	
b. At the time of the fossil's existence	
c. At the time of the rock movement	
d. In the future	

Appendix D: Scoring Instructions

- Each white box in the table below represents a construct and level that can be measured by the listed assessment task item
- Assign each item a score based on the outcome space.
- Place a mark in the appropriate box(es) for each item.
- Mark any box for which evidence of thinking / understanding at that level is shown, even if there is more than one per item.
- When all responses providing sufficient evidence for scoring have been marked, determine holistic scores for each construct as follows:
    - Fill in the box that corresponds to the most frequently demonstrated level of understanding for that construct (the mode)
        - When two or more levels within a construct are equally frequent, select the level that is not limited by available boxes
    - Extend a horizontal line out into adjacent boxes corresponding to any other level that was marked on individual items
    - This creates a visual indication of the scores (similar to a box and whisker plot)
    - The holistic score can also be represented numerically in the following format: (mode, min, max) for each construct.

Assessment Task Item & Description		**Geologic Time & Stratigraphy**				**Surface Processes**				**Plate Tectonics**				**Geologic Maps**			
Construct Map Levels →		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1a	Draw profile	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨				
1b	Describe profile	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨				
1c	Draw path	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨			
1d	Annotate path	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨			
1e	Explain reasoning	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨				
2a	Explain profile	▨	▨	▨	▨					▨	▨	▨	▨				▨
2b	Choose rock	▨	▨	▨	▨			▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
2c	Explain rock	▨	▨	▨	▨				▨	▨	▨	▨	▨	▨	▨	▨	▨
2d	Draw no weathering	▨	▨	▨	▨	▨				▨	▨	▨	▨	▨	▨	▨	▨
2e	Explain differences	▨	▨	▨	▨					▨	▨	▨	▨			▨	▨
3a	Classify rocks	▨				▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
3b	Date fossil rock					▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
3c	Outcrop correlation					▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
3d	Evidence of motion	▨	▨	▨	▨	▨	▨	▨	▨	▨				▨	▨	▨	▨
3e	Explain mechanism	▨	▨	▨	▨	▨	▨	▨	▨	▨				▨	▨	▨	▨
3f	Draw arrows	▨	▨	▨	▨	▨	▨	▨	▨				▨	▨			▨
4a	First event					▨			▨	▨	▨	▨	▨	▨	▨	▨	▨
4b	Time of fossil					▨			▨	▨	▨	▨	▨	▨	▨	▨	▨
4c	Time of movement	▨	▨	▨	▨									▨	▨	▨	▨
4d	Future landscape	▨	▨	▨	▨					▨				▨	▨	▨	▨
Holistic Scores →																	

Appendix E: Outcome Space

construct	Q	LVL	Response
Topographic Maps	1a	1	drawing is not a continuous line (e.g. several line segments, individual points)
Topographic Maps	1a	1	profile is flat
Topographic Maps	1a	1	profile shows a one-way slope
Topographic Maps	1a	2	profile shows a valley but doesn't show difference in east/west bank gradient
Topographic Maps	1a	2	profile shows a valley but doesn't reach from A to B
Topographic Maps	1a	2	profile shows a valley but its lowest point is not at X
Topographic Maps	1a	3	profile shows valley that is steeper along AX than along XB
Topographic Maps	1a	3	endpoints of profile are plotted at correct elevations but may be estimated otherwise
Topographic Maps	1a	3	profile is aligned to show viewpoint from X (near lowest point)
Topographic Maps	1a	3	profile is constructed with elevations at each contour line connected w/o smoothness
Topographic Maps	1a	4	profile shows accurate elevation throughout w/ smooth continuity
Topographic Maps	1b	1	response does not refer to the shapes or spacing of contour lines
Topographic Maps	1b	1	response does not refer to a valley
Topographic Maps	1b	1	describes the land as flat
Topographic Maps	1b	1	response does not refer to the contour lines on the map
Topographic Maps	1b	1	describes an incorrect landform such as a plateau
Topographic Maps	1b	2	identifies numbers on the map as representing changing elevation of the land
Topographic Maps	1b	2	states the land is going down based on elevation values on contour lines
Topographic Maps	1b	2	states a general relationship between slope and line spacing
Topographic Maps	1b	2	describes the slope going down, then up
Topographic Maps	1b	2	identifies a hill on the right / east side of the map
Topographic Maps	1b	2	refers to the river at the bottom of the hill or valley
Topographic Maps	1b	3	correctly describes the difference in slope on either side of the river; steeper on the left/west
Topographic Maps	1b	3	describes a process or procedure for constructing the profile
Topographic Maps	1b	4	describes v-shaped contour lines indicating a river or stream valley
Topographic Maps	1b	4	describes closed contours indicating a hill or mountain top
Topographic Maps	1cd	1	path does not go adjacent to all outcrops
Topographic Maps	1cd	2	path does not take elevation/slope into account (going up and down repeatedly, or down an excessively steep slope)
Topographic Maps	1cd	2	path goes from high elevation to low elevation, visiting all 3 outcrops, doesn't indicate taking relative slope into consideration

Topographic Maps	1cd	3	path follows a gentle downhill trend but crosses some contour lines more than once, visiting all 3 outcrops.
Topographic Maps	1cd	4	path follows a downhill trend, taking contour lines into account to avoid going up and and down
Topographic Maps	1e	1	explanation does not refer to topography or describes topographic features not on the map
Topographic Maps	1e	2	explanation refers to "steep," "flat," "uphill," "downhill" and/or mentions single locations
Topographic Maps	1e	3	explanation makes comparisons such as "steeper," "higher," or a change along the line
Topographic Maps	1e	4	explanation describes patterns in contour lines and connects to a change in slope or topography
Topographic Maps	2a	1	Describes a shape other than a river valley, e.g. a ridge, mountain, etc.
Topographic Maps	2a	2	Identifies a valley by name or description e.g. lower elevation
Topographic Maps	2a	3	Describes a v-shaped valley
Surface Processes	2a	1	describes processes unrelated to erosion by water, e.g "pressure" or "glaciers"
Surface Processes	2a	1	says the rocks were broken down or weathered
Surface Processes	2a	2	running water causes erosion
Surface Processes	2a	3	makes a connection between the process of erosion and the formation of the valley or change in elevation
Surface Processes	2a	3	identifies this location as the part of the rock that was susceptible to weathering/erosion
Surface Processes	2a	4	(unlikely) describes future change that the valley will get deeper and/or wider
Surface Processes	2b	1	incorrect rock (not the rounded one)
Surface Processes	2b	2	chooses correct rock #4 (rounded due to abrasion in riverbed)
Surface Processes	2c	1	explanation refers to irrelevant physical features of the rock, such as dark color
Surface Processes	2c	2	cites rounded edges or smooth texture as evidence for erosion, without explaining how they were created
Surface Processes	2c	3	cites rounded edges or smooth texture as evidence for erosion in water, citing abrasion or explaining how the movement of the river caused erosion of the rock via collisions with other rocks
Surface Processes	2d	2	profile shows a shape that is neither a valley nor a flat/gentle slope from A to B (e.g. hill, many hills or valleys, . cliff, etc)
Surface Processes	2d	3	profile shows a valley shape that is less steep / deep than the original
Surface Processes	2d	4	accurate "prediction": if no erosion, there would be no minimal point between A and B
Topographic Maps	2e	1	answer shows a misunderstanding of the topographic features
Topographic Maps	2e	2	answer describes qualitative or relative changes in the profile
Surface Processes	2e	1	refers to weathering and/or erosion incorrectly or without making a distinction between them
Surface Processes	2e	2	describes the results or evidence of erosion without clearly naming a process or mechanism for the change in landscape
Surface Processes	2e	3	describes erosion as a result of moving water that created the valley, using a correct definition
Surface Processes	2e	4	identifies and describes change in both elevation and slope due to erosion, stating that the land would remain un-eroded (higher elevation, flatter slope) in the absence of water
Surface Processes	2e	4	identifies and describes change in both elevation and slope due to erosion, stating that the water is responsible for creating the valley's depth (lower elevation) and steepersloped sides.
Geologic Time & Stratigraphy	3a	2	identifies a location or process that is based on on an incorrect rock ID, but matches the given description

Geologic Time & Stratigraphy	3a	2	identifies accurate **location** for rock formation based on the rock class
Geologic Time & Stratigraphy	3a	3	identifies accurate **process** for rock formation based on the rock class
Geologic Time & Stratigraphy	3a	4	describes a **place-based process** specific to the rock with accurate environmental details
Geologic Time & Stratigraphy	3b	1	says that rock formed a long time ago, without specifying a correct period or age
Geologic Time & Stratigraphy	3b	2	names a time period (anything in the paleozoic: carboniferous, permian, pennsylvanian, devonian, silurian, ordovician, or cambrian) that the rock formed
Geologic Time & Stratigraphy	3b	2	names the organism (trilobite or phacops) in the rock
Geologic Time & Stratigraphy	3b	3	accurately describes life or events on earth based on a horizontal correlation in the ESRT
Geologic Time & Stratigraphy	3b	4	states a possible absolute age of the rock / fossil (between 251 - 542 mya)
Geologic Time & Stratigraphy	3c	1	events are ordered or described based on outcrop or something else, not rules of stratigraphy
Geologic Time & Stratigraphy	3c	2	shows evidence of understanding rule of superposition
Geologic Time & Stratigraphy	3c	3	lists events in (all or mostly) correct order including crosscutting events: erosion, fault
Geologic Time & Stratigraphy	3c	4	lists or describes events in correct order with reference to an absolute time
Plate Tectonics	3d	2	identifies rapid change events only such as faulting or cracks
Plate Tectonics	3d	3	identifies longer term events such as folding, tilting etc. or processes such as metamorphism
Plate Tectonics	3d	4	identifies evidence / effects in the outcrops and describes the plate motion that caused them
Plate Tectonics	3e	2	describes multiple tectonic plates without describing movement
Plate Tectonics	3e	3	describes multiple plates moving at a boundary
Plate Tectonics	3e	3	identifies or describes mantle convection, without connecting to surface effects
Plate Tectonics	3e	4	describes mantle convection, with a connection to effects on the tectonic plates / surface crust
Plate Tectonics	3f	1	arrows are drawn that are not converging, or there is only one arrow
Plate Tectonics	3f	2	arrows are moving towards each other, but may move past each other (like transform movement)
Plate Tectonics	3f	3	pair of convergent arrows is drawn
Topographic Maps	3f	2	arrows are drawn on or between outcrops
Topographic Maps	3f	3	arrows are drawn on the map to indicate movement over a larger scale
Geologic Time & Stratigraphy	4a	1	No connection or reference to a specific rock formation environment, e.g. depicts a dry/static surface landscape
Geologic Time & Stratigraphy	4a	2	earliest geologic event (deposition of limestone) - limestone is mentioned or drawn, OR nonspecific shelly sea creatures (limestone composition)
Geologic Time & Stratigraphy	4a	3	describes and/or draws a credible formation environment for the first geologic event listed in student's respons to 3c even if not limestone
Geologic Time & Stratigraphy	4a	3	describes and/or draws generic water environment or deposition in water
Geologic Time & Stratigraphy	4a	4	describes and/or draws a shallow sea specifically characteristic of the early paleozoic (correct fauna such as trilobites, eurypterids, brachiopods, etc)
Surface Processes	4a	2	n/a I think

Surface Processes	4a	3	specifically shows or describes deposition processes
Geologic Mapping	4b	1	i don't know why i thought this was relevant?
Geologic Mapping	4b	2	i don't know why i thought this was relevant?
Geologic Mapping	4b	3	i don't know why i thought this was relevant?
Geologic Mapping	4b	4	i don't know why i thought this was relevant?
Geologic Time & Stratigraphy	4b	1	no connection to fossil, e.g. shows a non-marine environment or does not relate fossil & sedimenaty rock to formation env
Geologic Time & Stratigraphy	4b	2	fossil - describes and/or draws fossil or rocks in water // minimal or zero change from a
Geologic Time & Stratigraphy	4b	3	fossil - describes and/or draws water or deposition in water with phacops present
Geologic Time & Stratigraphy	4b	4	description includes an absolute time measure (e.g. MYA or named geologic time period / epoch)
Geologic Time & Stratigraphy	4b	4	fossil - describes/draws a change in marine environment compared to a, with justification
Surface Processes	4b	2	draws rock layer with unconformity, as if the surface were eroded
Surface Processes	4b	3	references change to environment based on sediment size
Surface Processes	4b	3	illustrates or describes wave action acting on rock layers
Surface Processes	4b	3	specifically shows or describes deposition processes
Plate Tectonics	4c	1	drawing/description does not show any change to the crust
Plate Tectonics	4c	2	drawing/description shows generic plate boundary features like cracks, lava
Plate Tectonics	4c	3	describes/shows orogeny due to convergent movement (mountains, volcanoes, increased elevation)
Plate Tectonics	4c	4	describes/shows mantle movement along with convergent crust movement
Plate Tectonics	4c	4	describes/shows both long and short term changes due to convergent movement (e.g. folding + volcanoes)
Plate Tectonics	4d	2	depicts/describes earthquakes due to plate movement
Plate Tectonics	4d	3	depicts/describes a change in landscape (e.g. mountains, coastal margin formation) due to plate movement
Plate Tectonics	4d	4	depicts/describes a change in climate or environment due to long-term plate movement
Surface Processes	4d	3	shows/describes the landscape as it is currently, with present-day surface processes such as water erosion to create a valley
Surface Processes	4d	4	prediction shows a change in landscape that takes local climate into account

Appendix F: Think-aloud Protocol

1. Read aloud to the student: *I am interested in what you think to yourself as you perform some tasks related to Earth Science. I will ask you to talk aloud as you work on the problems. What I mean by "talk aloud" is that I want you to say out loud everything that comes into your mind while doing the task. Put another way, I want you to say out loud what you say to yourself inside your head. Just act as if you are alone in the room speaking to yourself. If you are silent for any length of time I will remind you to keep talking aloud.*

2. Complete a short practice think-aloud. Read to the student: *Before we begin the assessment, we will start with a practice problem. I want you to talk aloud while you do this.* ***Multiply 10 times 15 in your head.*** *Be sure to talk aloud.*

3. Prepare to record. Ask the student: *I would like to begin videotaping you now. You may ask me to stop videotaping at any time. May I record?*

4. Begin recording. Read aloud to the student: *Today's date is [date]. Please state your name. …* ***You may begin working on the assessment.*** *Remember to talk aloud while you are working.*

5. While the student works, prompt think-aloud responses as necessary. Avoid asking any questions that relate to the specific components of the task. Instead, use prompts that relate to the students actions, such as:

    a. Please tell me what you are thinking now.
    b. What are you thinking about as you are writing?
    c. What are you thinking about as you are drawing?
    d. What are you thinking about as you are [other action]?

Appendix G: Inter-rater Reliability Notes

For **codes on student written responses:**

- We did this by construct
  - First, looked at the construct map and read through each level
  - Looked at one example of question and talked through how student written response corresponded to elements of construct map
  - Then each of us independently coded 5 examples
  - We compared codes and noted where we agreed and disagreed
  - When we disagreed on a code, we discussed and came to an agreement about what the code should be. (Note: we had zero instances of codes diverging by more than one level -- all disagreements were about, e.g., whether it should be GTS2 vs GTS3)
  - After this conversation, we independently coded 3 students' responses with the goal of reaching 20 codes for the construct.
  - We entered our codes into a spreadsheet
  - Used a formula to calculate the % of codes on which we agreed
  - I updated any codes that were changed or added in data analysis sheet

Construct	**Rater Agreement**	**Rater agreement after recoding**
Geologic Time and Stratigraphy	80% ** this is before recoding	**95%** Recoded items 3a, 4ab
Surface Processes	90% ** no items needed recoding	90% (no recoding)
Plate Tectonics	78.9% ** this is before recoding	**94.7%** Recoded item 3f
Topographic Maps	75% ** this is before recoding	**85%** Recoded item 1e

Recoded items based on conclusions from interrater reliability conversation:

- 3a (GTS)
- 4a & 4b (GTS)
- 3f (PT) → any arrow that identifies surface movement of land is a 2? Even if it is divergent or only one arrow. (bc level 1 is "static, not dynamic") ** arrow must be on map, not on blank paper space
- 1e (TM) → construct element in question is whether this task can ever evoke analysis of a "real world location," since the students are not physically at the location represented by the map. We counted the profile as a possible 4 for this. Also, what happens if students describe a landform (e.g. cliff) without including info about the contour lines specifically? Is that a "symbolic representation" or is it "patterns in contour lines"? We

can't assume the latter bc could also infer from elevations, so it's a 3 not a 4 if they name "cliff" w/o contour line pattern description

For **codes on student thinkalouds:**

- Constructs: GTS and SP (Friday), then PT and TM (Monday)
- Same process as before, except this time looking at thinkaloud transcripts
- I pulled the portions of the transcript corresponding to each item so that we could look at constructs, rather than subjects

Construct	**Rater Agreement**	**Rater agreement after recoding**
Geologic Time and Stratigraphy	75% ** this is before recoding	**90%**
Surface Processes	70% ** before recoding	**85%**
Plate Tectonics	**90%** ** no recoding needed	**90%** (no recoding)
Topographic Maps	**90%** ** no recoding needed, but I added one code that had been missing	**95%** (agreed to add code for one subject on item 2a)

- Notes for recoding:
  - GTS
    - 3a - possibility for many levels since subjects are observing multiple rock samples. Ignore answers based on wrong wrock identity if they are able to demonstrate level 2 or 3 with a sample they interpret correctly.
    - 3b - if they match fossil diagram to correct location on the ESRT and find something horizontally aligned with it (e.g. time period, information about life on earth, important geologic events), that is level 3 (kind of like correlation)
    - don't overscore for "variety of timescales" -- long term needs to be way long term
    - Difference between 2 and 3 -- at level 2, they identify materials or components that must have existed in the past. At level 3, they put it into an environmental context and/or describe the process that acted upon the materials.
  - SP
    - How to capture the difference between recognizing weathering & erosion on a small scale and a large scale? Or generic vs specific to individual agents of weathering and erosion?
      - If Ss describe large scale landscape effects accurately, that's a 3 -- "evidence of past processes in current landscape features"

        - If Ss **describe** specific mechanisms (not just name), that’s a 4
        - If they name the specific mechanism or agent without a description, that’s a 3
        - If they just name the resultant features, thats a ...2 ? “evidence of weathering and erosion processes”
    - PT
        - Agreement was high, but we agreed there was only one item where students were likely to express level 4 understanding -- assessment construction not a great match for construct map.
    - TM
        - Conversation about “analyze a real-world location” from last time served us well and led to high interrater reliability on the first pass
        - Keep an eye on level 4 codes in previously completed thinkalouds