2022

Efficient Assessment of Students' Proportional Reasoning

Michele Carney
Boise State University

Katie Paulding
California Polytechnic State University

Joe Champion
Boise State University

This is an Accepted Manuscript of an article published by Taylor & Francis in *Applied Measurement in Education* in 2022, available online: https://doi.org/10.1080/08957347.2022.2034825

Efficient Assessment of Students' Proportional Reasoning

Michele Carney
Boise State University

Katie Paulding
California Polytechnic State University, San Luis Obispo

Joe Champion
Boise State University

Abstract

Teachers need ways to efficiently assess students' cognitive understanding. One promising approach involves easily adapted and administered item types that yield quantitative scores that can be interpreted in terms of whether or not students likely possess key understandings. This study illustrates an approach to analyzing response process validity evidence from item types for assessing two important aspects of proportional reasoning. Data include results from an interview protocol used with 33 middle school students to compare their responses to prototypical item types to their conceptions of composed unit and multiplicative comparison. The findings provide validity evidence in support of the score interpretations for the item types but also detail important item specifications and caveats. Discussion includes recommendations for extending the research for examining response process validity evidence in support of claims related to cognitive interpretations of scores for other key mathematical conceptions.

Knowing What Students Know (Pellegrino, Chudowsky, & Glaser, 2001) recommends that for test results to meaningfully inform learning they need to include assessment of student cognition within a mathematical domain (i.e., how a student is thinking about the mathematics). However, assessment of student cognition is difficult to achieve in educational settings, and when attempted, is often time consuming (Cizek, 2010; Pellegrino, Chudowsky, & Glaser, 2001). This sets up a fundamental tension between efficiency and effectiveness of assessment in the mathematics classroom.

How can we provide teachers with high quality resources to efficiently assess students' cognition? One promising approach is to develop prototypical item types targeting students' mathematical cognition (Tjoe & de la Torre, 2014). Used successfully in projects such as Cognitively Guided Instruction (Carpenter, Fennema, Franke, Levi & Empson, 2015), these item types are templated to allow for multiple versions by varying aspects of the item (e.g., numbers or context). Administration is relatively quick and easy, and students' scores on a block of items are combined (e.g., 14 out of 20). However, it is crucial that the resulting quantitative value can be meaningfully interpreted in relation to student cognition. For example, based on a student's composite score for a block of items, can a teacher expect the student to likely possess certain mathematical understandings, while perhaps not possessing other understandings? However, we must ensure that these efficient items appropriately capture student cognition, which can be difficult.

In the technical language of assessment, this is about *score interpretation claims* (whether qualitative interpretations from the quantitative scores are valid), which require assessment developers to engage in gathering *response process validity evidence* to examine that claim (AERA, APA, & NCME, 2014). Unfortunately, response process validity evidence is seldom reported by developers of educational assessments, particularly in reference to score interpretation claims (Cizek, Rosenberg, & Koons, 2008; Padilla & Benítez, 2014).

The purpose of this report is to illustrate a process for addressing score interpretation claims through response process validity evidence through the specific context of an assessment of students' understanding of proportional reasoning. We propose fill-in-the-blank item types for efficiently assessing two aspects of students' proportional reasoning and examine response process interview data to determine if there is evidence to support score interpretation claims for the item types. In cases where the evidence may be insufficient, we describe how to adjust the item types and score interpretations to more closely target the desired aspects of cognition. We intend for this report to serve as an example to others' seeking to develop efficient mathematics assessments by illustrating ways to use response process evidence in relation to score interpretation claims.

Literature Review

The efficient assessment of students' proportional reasoning is complex. We start by summarizing related literature on the assessment of cognition in mathematics, followed by a framework for two components of students' proportional reasoning and associated item types. Lastly, we outline how validity evidence can support score interpretations, with a focus on response process evidence.

Assessing Cognition in Mathematics

Assessment items such as: 6 + 7 = ___ + 8 are designed to assess cognition related to an understanding of equality. Responses to these types of tasks can be used to infer students understand the meaning of the equal sign as representing equality between the two expressions. Studies suggest many students will respond '13' in the blank, which could be inferred as a lack of understanding of the meaning of the equal sign (Carpenter, Franke, & Levi, 2003; Matthews, Rittle-Johnson, McEldoon, & Taylor, 2012). If students correctly responded '5', this could be inferred as having some understanding of equality. Understanding student cognition in this way provides teachers with important information to determine next steps in instruction.

Designing tasks that assess a targeted aspect of student cognition is difficult, particularly when paired with a requirement that they be efficient to administer and score (Haja & Clarke, 2011). For example, teachers could individually interview students to assess cognition about a particular aspect of their thinking,, but the time required to administer and interpret results from this type of assessment is likely to be overly burdensome for many school personnel. Therefore, assessment items that can be relatively efficiently administered, scored, and interpreted (e.g., correct vs. incorrect, true/false, etc.) are ideal. Yet items that make use of those formats often do not allow for cognitively valid inferences related to the thinking a student used to arrive at the answer (Ward & Bennett, 2012). Therefore, claims that specific inferences about student cognition can be made based on easily administered and scored formatted items should be subject to scrutiny and supported by validity evidence. Prior to discussing the evidence needed, we examine the literature on students' proportional reasoning.

Proportional Reasoning

The specific construct for measurement in this study is students' proportional reasoning, which is a critical topic in mathematics with sufficient research around the general construct of middle grades' student cognition (Cetin & Ertekin, 2011; Ellis, 2013; Lobato, Ellis, & Zbiek, 2010; Ramful & Narod, 2013; Tucker et al., 2013). This research provides a clear target for operationalization of important aspects of student cognition into assessment items. Next we describe two key aspects of students' proportional reasoning.

In proportional situations, there are two mathematical relationships we want students to understand and flexibly make use of - scalar and functional. The *scalar relationship* describes the scale factor each quantity in the ratio can be multiplied or divided by to generate an equivalent ratio (Lobato et al., 2010). The scale factor in the scalar relationship changes as the equivalent ratio to be generated changes. In the Scalar Relationship Strategy presented in Figure 1, the student first scales down to a unit rate of 4 miles in 1 hour by dividing by 3, and then scales up to 8 miles in 2 hours by multiplying by 2. The student's focus is on coordinating the two quantities in the ratio in conjunction with one another by dividing and multiplying each quantity in the ratio by a common scale factor.

[Figure 1]

The *functional relationship* describes the constant multiplicative factor that exists between the two quantities in a rate situation (Lobato et al., 2010). The multiplicative factor in the functional relationship remains constant in all equivalent ratios. In the Functional Relationship Strategy presented in Figure 1, the student makes use of the constant multiplicative relationship of *x4* from hours to miles between the two quantities in the ratio and applies that relationship to the 2 hours to generate the 8 miles.

Student proportional reasoning conceptions refer to how the student is thinking about the mathematical relationships involved in a proportional reasoning situation. While highly related to the mathematical relationships, student conceptions differ from mathematical relationships in that the focus is on how the students are thinking about the situation as opposed to the mathematical relationship they are making use of [1].

A *composed unit conception* of a ratio involves forming a unit that contains the two quantities in the ratio and recognizing these quantities must be scaled in conjunction with one another to create equivalent ratios (Lobato et al., 2010). When students are working with linear functions, a composed unit conception helps them to conceptualize how a change in one variable results in a change in the other variable in a constant manner (e.g., rise over run on a graph or input/output tables). Composed unit reasoning is evident with the exemplar assessment item provided in Figure 1 if a student makes a statement such as "If Sophie goes 12 miles in 3 hours, and I divide each of those quantities by 3, then she goes 4 miles in 1 hour. If I double each of those, in 2 hours she would go 8 miles."

A *multiplicative comparison conception* involves using one of the quantities in the ratio as the base and describing the other quantity in the ratio as a multiplicative comparison of the base quantity (Lobato et al., 2010). When students are working with slope and linear functions, a multiplicative comparison conception helps them to recognize the constant multiplicative relationship between the two variables present in the equation for the function. Multiplicative comparison reasoning is evident with the exemplar assessment item provided in Figure 1 if a student makes a statement such as "If Sophie goes 12 miles in 3 hours, then miles are always 4 times the hours. So if Sophie only goes for 2 hours, then I multiply that by 4 to get 8 miles."

Research made us wonder about the increased difficulty students experienced with reasoning around the functional relationship versus the scalar relationship (e.g., Steinthorsdottir & Sriraman, 2009). Carney, Smith, Hughes, Brendefur, & Crawford (2016) researched systematic manipulation of the location of an integer multiplier—to press the scalar or functional relationship—on item difficulty and student solution strategies. They found there was not a difference in difficulty when it came to the mathematical relationships themselves (i.e., scalar or functional) with small integer multipliers. Based on this and other research (e.g., Simon & Placa, 2012; Steinthorsdottir & Sriraman, 2009), we wondered if the difference was related to how students conceived of these relationships - i.e., the composed unit and multiplicative comparison conceptions. We developed prototypical assessment item types to tease these two conceptions apart, with a focus on items that were efficient to administer and score. In addition we wanted to design item types that could be modified by teachers to easily generate new items.

Proportional Reasoning Item Type Design

The item design used in this study starts with a common stem describing a rate[2] situation involving a relatively common context (e.g., mixing paint, hiking, buying cookies). The example used in the previous section is one such example: *Sophia can bike 12 miles in 3 hours.* From this common stem, five item types are presented, each designed to assess a different aspect of proportional reasoning. In Table 1 we name and describe each item type, the reasoning each item type assesses, and provide an example. The common stem and five item types make up a *testlet*. These item types were developed with the focus on use by teachers. By adjusting the context and quantities in the ratio, teachers can generate new testlets to assess composed unit and multiplicative comparison reasoning. Below is a brief rationale for each item type.

[Table 1]

Small Single-Digit Multiplier (SSDM). These item types are designed to be relatively easy for students and include an iconic representation of the initial ratio. Students who do not possess strong proportional reasoning understanding may still be able to solve these problems using more informal reasoning strategies. The intent is for the items to screen for students who primarily hold an incorrect additive perspective (Misailidou & Williams, 2003).

[1] This distinction is important when students divide one quantity in the ratio by another - making use of the functional relationship - to find the unit rate but do not actually conceive of the multiplicative comparison between the two quantities when performing the mathematical procedure.
[2] We use Thompson's (1994) perspective of rate as "... a set of infinitely many equivalent ratios (p. 42)".

Double-Digit Scalar Multiplier (DDSM). These item types are designed such that the multiplier is large enough that it is unlikely a student would solve them through informal strategies alone, but are still accessible without a calculator. We expect them to be solved via scalar multiplicative strategies and for students to likely express a composed unit conception.

Unit Rate. Previous research indicated students who demonstrated understanding of unit rate did not, for the most part, demonstrate an understanding of the multiplicative comparison conception (Carney & Crawford, 2016). These items are designed to determine if respondents can generate a unit rate, which typically involves expression of a composed unit conception.

Equations. Modeling a rate situation with an equation is critical for understanding algebraic representations of proportional and linear relationships. These items are designed to determine if respondents can fill in the blank with the correct multiplicative comparison multiplier, potentially indicating understanding of the multiplicative comparison conception.

Generalizing. These items are designed to determine if students can generalize the multiplicative comparison relationship, potentially indicating understanding of this conception (Staples & Truxaw, 2012) refer to this as the language of proportional reasoning).

Response Process Validity Evidence

In the case of mathematics assessment items such as we have described, assessment users need evidence that student responses are indicative of the process of thinking the assessment developer claims they are engaging in. In the *Standards for Education and Psychological Testing* (AERA, APA, & NCME, 2014), the evidence to support this type of claim or assumption is termed *response process* validity evidence. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) include response process validity evidence as one of five categories of evidence important to consider in developing validation arguments. They highlight the importance of gathering response process evidence, particularly for tests that claim results can be interpreted in relation to cognitive processes.

> Some construct interpretations involve more or less explicit assumptions about the cognitive processes engaged in by test takers. Theoretical and empirical analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers. For instance, if a test is intended to assess mathematical reasoning, it becomes important to determine whether test takers are, in fact, reasoning about the material given instead of following a standard algorithm applicable only to the specific items on the test (p. 15).

While the necessity of response process evidence is logically clear, particularly for cognitive interpretations, the reality is it is seldom provided. Cizek, Rosenberg, & Koons (2008) in a review of educational and psychological tests in the *Mental Measurements Yearbook*, found only 1.8% of studies provided response process validity evidence. Padilla & Benítez (2014) described validation studies that included response process validity evidence as scant and suggested that the lack of clear advice on how to conduct this work could be a contributing factor. Recently, more studies have been published that highlight the importance of reporting response process evidence (see Bostic, 2021).

Methodological information on gathering and examining response process validity evidence is increasing in availability (e.g., Castillo-Díaz & Padilla, 2013; Leighton, 2017; Padilla & Benítez, 2014; Padilla & Leighton, 2017). These tend to recommend cognitive interviews as a data gathering method for response process validity evidence. Approaches to administering and analyzing cognitive interviews to improve measurement mechanisms is also relatively well delineated (Miller, Chepp, Wilson, Padilla, 2014; Willis, 2015). While called for (Leighton, 2021), what is less articulated in the literature is how to connect response process evidence from cognitive interviews to validation arguments related to score interpretation claims.

Strength of Evidence Criteria for Response Process. Validation involves making an argument in support of score interpretations for proposed uses (AERA, APA, & NCME, 2014). This includes (a) clearly stating the interpretation(s) and proposed use(s), (b) identifying assumptions or claims that underlie the stated interpretation and use, (c) gathering

evidence to investigate the validity of the assumptions or claims, and (d) presenting the evidence in the form of a coherent argument for assessment users to use in evaluating the validity of the interpretation and proposed use(s) (Kane, 2016).

With regards to presenting evidence in support of particular claims or assumptions, there are generally accepted guidelines in the research literature (e.g., model fit or internal reliability guidelines) for quantitative evidence. However, guidelines related to strength of evidence necessary to support a particular assumption or claim are less clear with response processes, particularly when it comes to student cognitive processes. Work focused on student cognition has primarily focused on coding individual responses as congruent or not congruent with the construct being assessed (e.g., Hopfenbeck & Maul, 2011). If a claim is related to student use of a particular understanding when they correctly respond to a particular item, then a researcher might analyze the item level data for alignment and provide a count or percent of alignment as evidence (DiBello, Pellegrino, Gane, Goldman, 2017). However when the claim is related to the cognitive interpretation that can be made when a student solves a majority of a particular item type correctly or incorrectly (i.e., the score interpretation claim), the unit of analysis shifts from a student response to a single item to a set of the individual student's responses across a block of items.

To evaluate the strength of evidence of our score interpretation claims, we used Willis' (2015) framework for analysis of cognitive interviews by *a priori* developing explicit criteria to assess the relationship between student interview responses to assessment items and the related interpretations. While the decision-tree of criteria is specific to our work, we see the general approach of assessment developers explicitly developing and stating their strength of evidence criteria, the quality of which can then be evaluated by others, as an important step in the analysis and reporting of response process data related to score interpretation claims.

The score interpretation claims that correspond to the item types are students who solve the majority of a block of items:

- Correctly, likely possess the understanding of the conception - either composed unit or multiplicative comparison - that the block of items are designed to assess.
- Incorrectly, likely do not possess the understanding of the conception - either composed unit or multiplicative comparison - that the block of items are designed to assess.

Our research questions address the strength of evidence in support of score interpretation claims:

1. Composed Unit Understanding: What is the strength of evidence in support of the score interpretation claims that students *likely possess* or *likely do not possess* composed unit understanding when they solve the majority of the items correctly or incorrectly, respectively?
2. Multiplicative Comparison Understanding: What is the strength of evidence in support of the score interpretation claims that students *likely possess* or *likely do not possess* multiplicative comparison understanding when they solve the majority of the items correctly or incorrectly, respectively?
3. Item Features: For situations where the strength of evidence is less than strong, are there patterns in the data that suggest item features which may affect student reasoning?

Methods

Participants

We conducted and recorded full cognitive interviews with 32 students (Padilla & Leighton, 2017) recommend between 20-50 interviews to achieve theoretical saturation and relevance) in grades 6-8 (ages 11-14) at three different schools. The schools represented students from a range of (a) settings (suburban (n=1) and rural (n=2)), (b) proficiency levels on the end of year state test (~40-65% of students proficient or advanced), and (c) eligibility for the free-and-reduced lunch (~10-45%). Specific demographic information about the students (e.g., race/ethnicity, ELL status) was not collected. Students (who had returned IRB paperwork) were randomly selected from stratified groups based on their recent performance on a test of proportional reasoning (i.e., high, moderate, and low performance groups). The goal of stratification was to have a range of knowledge and skills represented in the responses (Padilla & Leighton, 2017).

A relatively equal group of students were selected from each of these bands, excluding those with extremely low performance on the assessment and slightly reducing the size of the low group because responses from these students would not provide as much insight to our understanding of the aspects of cognition assessed by the items.

- Low = 2-7 (of 20) items correct, 518 (34%) of all students, 8 (25%) of interviewed students.
- Medium = 8-12 (of 20) correct, 471 (31%) of all students, 12 (37.5%) of interviewed students
- High = 13-20 (of 20) items correct, 543 (35%) of all students, 12 (37.5% of interview students

Setting

The interviews were conducted on the school premises in offices or classrooms that were not being used for other purposes on the day(s) of the interviews. The rooms were in locations familiar to the students, relatively quiet, and isolated from outside disruptions. In addition to the student being interviewed and the interviewer, a third person was present who managed the video recording and assisted with paperwork.

Interview Protocol

The interview protocol consisted of the interview form - two different testlets presented to the student to solve - as well as the interview procedure used to explain the interview to students and elicit their thinking.

Interview Forms. There were two different interview forms (A & B), with each form consisting of two different testlets (see figure 2 for an example of Interview Form B). A total of 16 students received Form A and 17 students received Form B. Each interview form consisted of two testlets using the following contexts and initial ratio in the stem: Interview Form A - Testlet One: 6 ounces of red paint to 3 ounces of white paint, Testlet Two: 5 sugar cookies for $4; Interview Form B - Testlet One: 12 miles in 3 hours, Testlet Two: 5 miles in 2 hours.

The interview forms were designed to be similar in difficulty based on the contexts, number relationships, and past student performance on these items. Each set consisted of an easier number relationship for the functional relationship presented first (i.e., 6:3 or 12:3) and a harder number relationship for the functional relationship presented second (i.e., 5:4 or 5:2). We intentionally presented the easier number relationship first in order to reduce or eliminate excessive cognitive load for struggling students and to decrease the likelihood of triggering initial incorrect thinking due to a harder number relationship.

Two item blocks can be constructed across the two testlets within an interview form. Figure 2 highlights this organization. One item block was focused on assessing composed unit conception and consisted of 4 total items, two unit rate item types and two double-digit scalar multiplier item types, with one of each item type within each testlet. The other item block was focused on assessing the multiplicative comparison conception and consisted of 4 total items; two equation and two generalizing item types, with one of each item type within each testlet. The multiplicative comparison relationship was reversed between the two items within a testlet (i.e., the base quantity for the comparison was changed) so the answers were never the same.

[Figure 2]

***Interview Procedure*.** The interviews were conducted by both the first and second author. The students were informed the intent of the interview was to better understand how they were thinking when solving particular types of mathematics problems. Students were then presented with Testlet One (of Form A or B). They were asked to solve the five items within the block and prompted to explain their thinking as they solved the items. Following their responses to all items within Testlet One, we presented them Testlet Two (of the respective form) and prompted them to explain their thinking as they solved the items.

Depending on the explanation provided, the interviewer occasionally asked follow-up questions to better understand the student process and/or underlying cognition, such as “Can you further explain how you were thinking?” One lesson learned from this approach is that students tended to do a sufficient job of explaining their approach to solving an item either when first explaining their approach or when prompted with a follow-up question. However, it was more difficult to get students to consistently articulate the underlying cognition involved in their approach. In asking questions to probe if students were using particular conceptions, it sometimes felt as if the questions were inducing

learning (i.e., prompting students to see the particular relationships), and it was difficult to discern whether students actually understood and used these relationships on their own or if their conceptions were a result of interviewer probing. We limited the amount of follow-up probing, which in turn influenced the level of articulation we received from students, and therefore, influenced the classification of the strength of the evidence used in support of a particular interpretation (further elaborated in the Results section).

Analysis

The data analysis can be broken into two segments with multiple steps in each segment. See Figure 3 for an overview of the analysis steps. The first segment (left side of figure 3), focused on articulating the score interpretation, involved analysis of student numerical responses to the testlet items from the perspective of correct or incorrect (step 1), and was followed by assigning a score interpretation for each item block (step 2).

The second segment (right side of figure 3) focused on response process, and involved cognitive coding (Willis, 2015) of the video and transcript data using the framework of composed unit and multiplicative comparison conceptions (step 3), applying the strength of evidence categories based on the score interpretation category in conjunction with the cognition code (step 4), and then for item blocks assigned a strength of evidence category below strong, using pattern coding (Willis, 2015) to analyze what factors might be impacting item functioning (step 5).

[Figure 3]

Results

The results are presented in the order of the steps presented in Figure 3. In step 1, all the testlet items were marked as correct or incorrect and scored as 1 or 0, respectively.

Step 2 Block Level Score Interpretation

In step 2, a total score was calculated for each student for each of the two item blocks - composed unit and multiplicative comparison - from the scoring conducted in step 1 with a range of 0-4 correct possible. Each student received two total scores, one from each item block. A score interpretation was then assigned for each item block depending upon the number of items correct (see Table 2 for the score interpretation). For example, a student who correctly answered three of the composed unit item types correctly would receive a score interpretation label of *likely possess composed unit conception*. This same student may have correctly answered one of the multiplicative comparison item types correctly so would receive a score interpretation of *likely does not possess multiplicative comparison conception*.

[Table 2]

Within the composed unit item block score interpretations, the frequency of *likely possess* versus *likely does not possess* was relatively similar with 44% and 35% of students in each respective category. Within the multiplicative comparison item block score interpretations, the frequency of *likely possess* versus *likely does not possess* leaned towards *likely does not possess* with 22% and 47% of students in each respective category.

Step 3 Item Level Cognitive Coding

Cognitive coding involves assigning predetermined codes based on the respondent cognitive processes (Willis, 2015). In step 3, all items were independently coded by two researchers based on the reasoning expressed during the interview. Codes were compared and discrepancies were discussed until agreement was reached.

The DDSM and unit rate item codes were (a) composed unit, (b) other, or (c) indeterminate. An item response was coded as composed unit when the verbal response indicated the student had formed a unit from the quantities in the original ratio and recognized these quantities must be scaled in conjunction with one another. This often, but not always, involved determining a unit rate from the original quantities and then scaling the quantities in the unit rate by a common scale factor. Students' understanding was often contextually based, and they were not required to generalize their understanding of the scalar relationship beyond the problem context.

The generalizing and unit rate items codes were (a) multiplicative comparison, (b) other, or (c) indeterminate. To be coded as multiplicative comparison, a student had to state the multiplicative relationship between the two quantities in the ratio with one of the quantities being expressed as a multiplicative comparison of the other quantity (Lobato et al, 2010; Staples & Truxaw, 2012). The requirement for indicating a multiplicative comparison conception was present was more stringent than for the composed unit item types because the generalizing item was essentially a direct statement of the multiplicative comparison relationship. Therefore, a student had to provide a response beyond simply re-reading the given statement and filling in their numerical answer as they read in order to indicate an understanding of the relationship; however, they often did not do so. This left the interview team in an awkward position of either probing further, which appeared to induce learning (and thus discontinued for the most part), or moving on with the interview, even when a student only read the generalizing item verbatim (which would not be coded as expressing multiplicative comparison reasoning). This is further addressed in the discussion.

Step 4 Block Level Strength of Evidence Categories

In step 4 the strength of evidence criteria (see figure 4) were applied based first on the assigned score interpretation categories from step 2 of *likely possess* or *likely does not possess* the particular aspect of cognition (step 2) in conjunction with the strength of evidence criteria based on the response coding from step 3. Table 3 presents the frequency of alignment between score interpretations and strength of evidence criteria across both the composed unit and the multiplicative comparison conceptions.

[Figure 4]

For *likely possess composed unit* conception, all responses fell within the strong (54%) or moderate (46%) evidence categories. For the *likely does not possess composed unit* conception, the majority of responses fell within the strong evidence category (64%) and some in the minimal evidence category (36%). For *likely possess multiplicative comparison* conception, the responses were more distributed across the evidence categories with two strong (29%), three moderate (43%), one minimal (14%), and one none (14%). Lastly, for *likely does not possess multiplicative comparison* conception, the majority of responses fell within the strong evidence category (80%) with one in moderate and two in minimal. All item block responses that did not meet the strong evidence category requirements were further investigated in step 5 using pattern coding to try to identify factors that might be impacting item functioning.

[Table 3]

Step 5 Block Level Pattern Coding of Item Functioning

In step 5, item blocks that were coded with moderate, minimal, or none strength of evidence categories were further investigated via pattern coding. Pattern coding is an inductive or bottom-up method of coding where the goal is to detect relationships between and within the various item features involved in an assessment. The "...emphasis is not so much on naming, labeling, or assigning codes, but rather on describing how item functioning varies as a function of these other factors" (Willis, 2015, p. 107). This section is divided into the four score interpretation categories: *likely possess composed unit*, *likely does not possess composed unit*, *likely possess multiplicative comparison*, and *likely does not possess multiplicative comparison*.

Likely Possess Composed Unit. We investigated the six moderate strength of evidence responses to determine if there were patterns in why the students only presented composed unit reasoning on one item block and not on both. Four of these students who received Interview Form A and the testlet with the stem - 6 ounces of red paint to 3 ounces of white paint - did not express composed unit reasoning due to the ease of the functional relationship. In other words, providing a multiplicative comparison relationship that involved doubling or halving made it more likely that students would make use of the multiplicative comparison conception. For example, on the DDSM item for 6 ounces of red paint to 3 ounces of white paint, one student stated '*And then for a hundred eight ounces of red paint you'd need fifty-four ounces of white because it's half of the red paint."* The other two students who had moderate strength of evidence responses received Interview Form B, and we could not detect a discernible pattern in their responses. The pattern in responses related to the testlet with a stem that had a doubling/halving functional relationship indicated the need for a testlet specification that a doubling/halving functional relationship should be avoided when items are created.

Likely Does Not Possess Composed Unit. We investigated the four minimal strength of evidence responses to determine if there were patterns indicating why students who only answered 0 or 1 of the composed unit item correctly were able to explicitly and correctly express composed unit reasoning two or more times within one item block. Three of these students correctly expressed composed unit reasoning on both the small single-digit multiplier item and the unit rate item, but were unable to do so on the double-digit scalar multiplier (DDSM) item. All four students determined a unit rate and expressed composed unit reasoning around the unit rate but had more difficulty identifying and/or expressing a scalar multiplier for the DDSM items. This difficulty appeared to be related to the inability to determine the multiplier when (a) it was a double-digit, and (b) the given quantity in the ratio being scaled to did not match with the quantity that was one in the unit rate.

Our analysis of the *likely does not possess composed unit conception* responses with minimal evidence indicates that while these students could express some composed unit reasoning around the relatively easy number relationships and generate a unit rate, they were unable to identify the double-digit scalar multiplier, indicating they could not consistently make use of composed unit reasoning. Thus the pattern of responses indicates the score interpretation is accurate, but that the strength of evidence criteria may need to be adjusted to indicate students need to consistently express composed unit reasoning for both unit rate and scalar multipliers.

***Likely Possess Multiplicative Comparison*.** We investigated the three moderate, one minimal, and one none strength of evidence responses to determine if there were patterns in why students who answered 3 or 4 of the multiplicative comparison items correctly were unable to consistently express multiplicative comparison reasoning. The three students in the moderate strength of evidence category all correctly expressed multiplicative comparison reasoning on the testlet with the stem - 6 ounces of red paint to 3 ounces of white paint - likely due to the ease of the relationship. Two of these same students correctly solved the multiplicative comparison items in the other testlet but did not express multiplicative comparison reasoning when describing their solution process. The other student, along with the two students whose responses fell into the *minimal* and *none* strength of evidence categories, all reasoned with a numeric equation. Essentially they used the quantities from the original ratio to form an equation and then performed mental calculations to see what quantities would make the equation true. While this is a valuable strategy, it is not clear whether or not they held a multiplicative comparison conception. For example, on the Equation item with the stem - 12 miles: 3 hours - one of these students stated, '*I did the number of hours is three and the number of miles is twelve, and you, to get twelve to three you'd have to divide by four, but it says multiplication so I multiplied by one fourth.'*

The pattern of responses provides insight into two factors. First, similar to our analysis related to the *likely possess composed unit conception response*s, the use of a common stem that has a doubling or halving relationship for the multiplicative comparison clouds the ability to determine if students truly conceive of the relationship between the two quantity in the ratio, therefore, use of a double/halving functional relationship should be avoided. Second, a caveat needs to be placed on the *likely possess multiplicative comparison* score interpretation related to the potential use of 'equation reasoning' to solve these items, and that it is difficult to determine if the students conceive of the multiplicative comparison between the quantities as intended.

Likely Does Not Possess Multiplicative Comparison. We investigated the one moderate and two minimal strength of evidence responses to determine if there were patterns in why students who answered 0 or 1 of the multiplicative comparison items correct were able to explicitly and correctly express the multiplicative comparison conception once (moderate) or two or more times (minimal) within one item block. Similar to earlier patterns, we found students expressed the multiplicative comparison once (moderate) or more than once (minimal) on the testlet with the stem - 6 ounces of red paint to 3 ounces of white paint - likely due to the ease of the doubling/halving functional relationship. Thus the pattern of responses indicates a doubling/halving functional relationship should be avoided when items are created.

Discussion

Generic, relatively easy-to-use item types designed to assess student thinking have the potential to assist teachers in better understanding individual student cognition and inform their instruction (Tjoe & de la Torre, 2014). To ensure these item types are assessing the intended aspects of cognition, it is important to gather response process validity evidence (AERA, APA, & NCME, 2014). This evidence can be used to (a) provide validity evidence for a validation argument, (b) clarify item specifications and (c) detail caveats related to the score interpretation.

In the context of efficient assessment of students' mathematical cognition, this study illustrates a novel way to use cognitive interviews alongside strength of evidence criteria in relation to score interpretation. Assessment developers need approaches for examining proposed qualitative interpretations of students' mathematical cognition from quantitative test scores related to response process validity evidence from cognitive interviews.The results of which can then be central to constructing a chain of reasoning in a validation argument (Kane, 2016).

In addition to supporting developing evidence for validity arguments, our research supports Shepard's (2018) call for meaningful qualitative interpretations of quantitative scores, particularly for classroom-based formative assessment purposes. Assessment developers commonly provide qualitative interpretations of quantitative scores, but these interpretations are often in relation to a skills continuum as opposed to student cognition (e.g., Northwest Evaluation Association's Measures of Academic Progress or Renaissance Learning's Star Math assessment). We see the response process validity evidence as a necessary aspect of both developing and supporting qualitative interpretations focused on student cognition. It is important to note, that to do this work well, developers need strong expertise in measurement and student cognition. When mathematics assessments scores are interpreted in relation to a skills continuum less mathematics expertise is required. However, when one proposes that scores can be interpreted in relation to student cognition, significantly more expertise is needed and that expertise is likely domain specific (e.g., proportional reasoning). This is an important consideration for test developers interested in conducting this type of work.

We have provided a simple example of examining the validity of interpretation statements focused on student cognition in relation to a strength of an evidence criteria model. These criteria will necessarily differ depending upon the particular situation or aspect of cognition being assessed. We invite others to explore and refine this approach for other cognitive domains. Work in this area has the potential to provide teachers with efficient ways to assess key cognitive understandings, and ultimately to better support students' individual and collective learning of mathematics.

References

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

Black, P., & Wiliam, D. (2005). *Inside the black box: Raising standards through classroom assessment*. Granada Learning.

Bostic, J. (2021). Think Alouds: Informing Scholarship and Broadening Partnerships through Assessment. *Applied Measurement in Education*, 34(1), 1-9.

Carney, M. B., & Crawford, A. (2016). Students' reasoning around the functional relationship. Proceedings in the *38th Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, Tucson, AZ.

Carney, M. B., Smith, E., Hughes, G. R., Brendefur, J. L., & Crawford, A. (2016). Influence of proportional number relationships on item accessibility and students' strategies. *Mathematics Education Research Journal, 28*(4), 503-522.

Carpenter, T. P., Franke, M. L., & Levi, L. (2003). *Thinking mathematically: Integrating arithmetic and algebra in elementary school.* Portsmouth, NH: Heinemann.

Carpenter, T. P., Fennema, E., Franke, M. L., Levi, L., & Empson, S. B. (2015). Children's mathematics : cognitively guided instruction (Second).

Castillo-Díaz, M., & Padilla, J. L. (2013). How cognitive interviewing can provide validity evidence of the response processes to scale items. *Social Indicators Research, 114*(3), 963-975.

Cengiz, N., & Rathouz, M. (2018). Making sense of equivalent ratios. *Mathematics Teaching in the Middle School, 24*(3), 148-155.

Cetin, H., & Ertekin, E. (2011). The relationship between eighth grade primary school students' proportional reasoning skills and success in solving equations. *International Journal of Instruction*, *4*(1), 47-62.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, *68*(3), 397-412.

Cizek, G. J. (2010). An introduction to formative assessment: History, characteristics, and challenges. In *Handbook of formative assessment* (pp. 15-29). Routledge.

DiBello, L. V., Pellegrino, J. W., Gane, B. D., & Goldman, S. R. (2017). The contribution of student response processes to validity analyses for instructionally supportive assessments. *Validation of score meaning in the next generation of assessments. The use of response processes*, 85-99.

Ellis, A. (2013). *Research brief: Teaching ratio and proportion in the middle grades*. Reston, VA. Retrieved from http://www.nctm.org/news/content.aspx?id=35822

Haja, S., & Clarke, D. (2011). Middle school students' responses to two-tier tasks. *Mathematics Education Research Journal*, 23, 67–76.

Hopfenbeck, T. N., & Maul, A. (2011). Examining evidence for the validity of PISA learning strategy scales based on student response processes. *International Journal of Testing*, *11*(2), 95-121.

Kane, M. T. (2016). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In S. Lane, M. Raymond, & T. M. Haladyna (Eds.), *Handbook of Test Development* (2nd ed.). New York, NY: Routledge.

Leighton, J. P. (2017). Collecting and analyzing verbal response process data in the service of interpretive and validity arguments. In *Validation of Score Meaning for the Next Generation of Assessments* (pp. 25-38). Routledge.

Leighton, J. P. (2021) Rethinking Think-Alouds: The Often-Problematic Collection of Response Process Data. *Applied Measurement in Education*, 34:1, 61-74, DOI: 10.1080/08957347.2020.1835911

Lobato, J., Ellis, A., & Zbiek, R. M. (2010). *Developing essential understanding of ratios, proportions, and proportional reasoning for teaching mathematics: Grades 6-8.* National Council of Teachers of Mathematics. Reston, VA.

Matthews, P., Rittle-Johnson, B., McEldoon, K., & Taylor, R. (2012). Measure for measure: What combining diverse measures reveals about children's understanding of the equal sign as an indicator of mathematical equality. *Journal for Research in Mathematics Education*, *43*(3), 316-350.

Miller, K., Chepp, V., Willson, S., & Padilla, J. L. (Eds.). (2014). *Cognitive interviewing methodology*. John Wiley & Sons.

Misailidou, C., & Williams, J. (2003). Diagnostic assessment of children's proportional reasoning. *The Journal of Mathematical Behavior*, *22*(3), 335-368.

National Council of Teachers of Mathematics (NCTM). (2014). *Principles to actions: Ensuring mathematical success for all.*

Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136-144.

Padilla J. L., & Leighton, J. P. (2017). Cognitive interviewing and think aloud methods. In Zumbo B. & Hubley A. (Eds.), *Understanding and Investigating Response Processes in Validation Research.* Social Indicators Research Series (Vol. 69). Springer.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Ramful, A., & Narod, F. B. (2014). Proportional reasoning in the learning of chemistry: levels of complexity. *Math Education Research Journal*, 26, 25–46. Retrieved from https://doi.org/10.1007/s13394-013-0110-7

Shepard, L. A. (2018). Learning progressions as tools for assessment and learning. *Applied Measurement in Education*, *31*(2), 165-174.

Simon, M. A., & Placa, N. (2012). Reasoning about intensive quantities in whole-number multiplication? A possible basis for ratio understanding. *For the Learning of Mathematics*, *32*(2), 35-41.

Staples, M. E., & Truxaw, M. P. (2012). An initial framework for the language of higher-order thinking mathematics practices. *Mathematics Education Research Journal*, 24, 257–281. https://doi.org/10.1007/s13394-012-0038-3

Steinthorsdottir, O. B., & Sriraman, B. (2009). Icelandic 5th-grade girls' developmental trajectories in proportional reasoning. *Mathematics Education Research Journal*, *21*(1), 6-30.

Tjoe, H., & de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal*, 26, 237–255. Retrieved from https://doi.org/10.1007/s13394-013-0090-7

Thompson, P. W. (1994). The development of the concept of speed and its relationship to concepts of rate. In G. Harel & J. Confrey (Eds.), *The Development of Multiplicative Reasoning in the Learning of Mathematics* (pp. 181-234). Albany, NY: SUNY Press.

Tucker, M., Daro, P., Snow, C., Pellegrino, J., Everson, H., Glasper, R., Mannes, K., & Fain, P. (2013). *What does it really mean to be college and work ready? The mathematics and english literacy required of first year community college students.* Washington, DC.

Ward, W. C., & Bennett, R. E. (Eds.). (2012). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment.* Routledge.

Willis, G. B. (2015). *Analysis of the cognitive interview in questionnaire design.* Oxford University Press.

Table 1. The five item types that make up the proportional reasoning testlet, including an example, item description, and reasoning assessed.

Reasoning	Item Type	Example	Description
Informal	Small Single-Digit Multiplier	Sophia can bike 12 miles in 3 hours. How far can Sophia bike in 2 hours?	A missing-value item where the scalar or functional multiplier was 2, 3, or 4.
Composed Unit	Double-Digit Scalar Multiplier	Sophia can bike ____ miles in 45 hours.	A missing-value item where the scalar multiplier is between 15 and 18.
	Unit Rate	In one mile, Sophia bikes ____ hours.	An item that asks for the value of one of the quantities in the ratio when the other quantity value is one.
Multiplicative Comparison	Equation	number of hours = _____● number of miles	An item that presents a contextual equation where the multiplicative comparison (constant of proportionality) is filled in the blank.
	Generalizing	The number of miles is always ____ times the hours.	An item that involves filling in the blank on a statement of the multiplicative comparison relationship.

Table 2. Score interpretation labels based on the number of items correct within an item block.

Conception (item types)	No. of items correct	Label	Score Interpretation	Freq
Composed Unit (double-digit multiplier and unit rate items)	3-4	Likely Possess	likely possesses an elementary understanding of the composed unit conception	14/32 (44%)
	2	Indeterminate	undetermined	7/32 (22%)
	0-1	Likely Does Not Possess	likely does not possess an elementary understanding of the composed unit conception	11/32 (35%)
Multiplicative Comparison (generalizing and equation items)	3-4	Likely Possess	the student likely possesses an elementary understanding of the multiplicative comparison conception	7/32 (22%)
	2	Indeterminate	undetermined	10/32 (31%)
	0-1	Likely Does Not Possess	the student likely does not possess an elementary understanding of the multiplicative comparison conception	15/32 (47%)

Table 3. Frequency of alignment between score interpretations and strength of evidence in student interviews.

Conception	Score Interpretation from Step 2[1]	Strength of Evidence from Student Interview to Support Interpretation[2]			
		Strong	Moderate	Minimal	None
Composed Unit	Likely Possess	7	6	0	0
	Likely Does Not Possess	7	0	4	0
Multiplicative Comparison	Likely Possess	2	3	1	1
	Likely Does Not Possess	12	1	2	0

Note. [1]Interpretations apply a majority as the threshold for item performance (e.g., at least 3 of 4 items correct). [2]See Figure 4 for strength of evidence criteria.

Figure 1. Two different student solution strategies making use of the scalar and functional relationships.

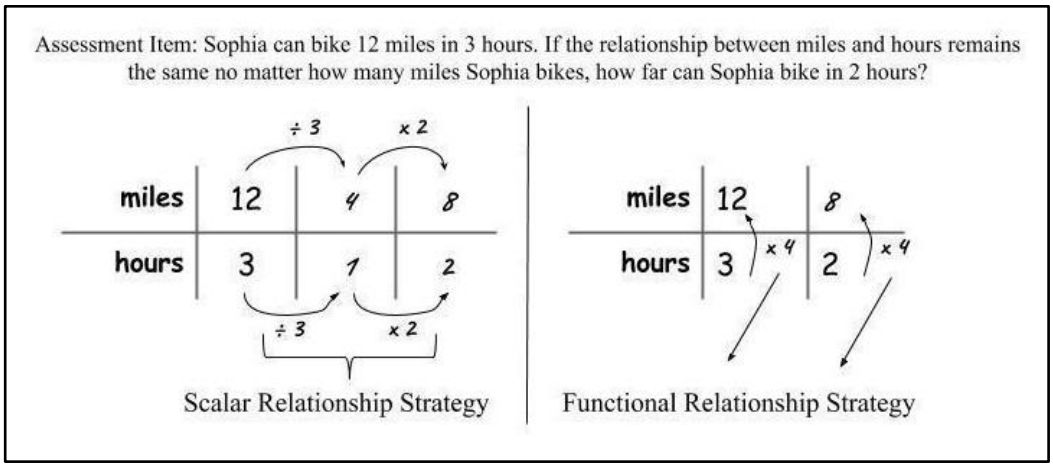


Figure 2. Layout of interview form detailing the item types, testlets, and item blocks across the two testlets.

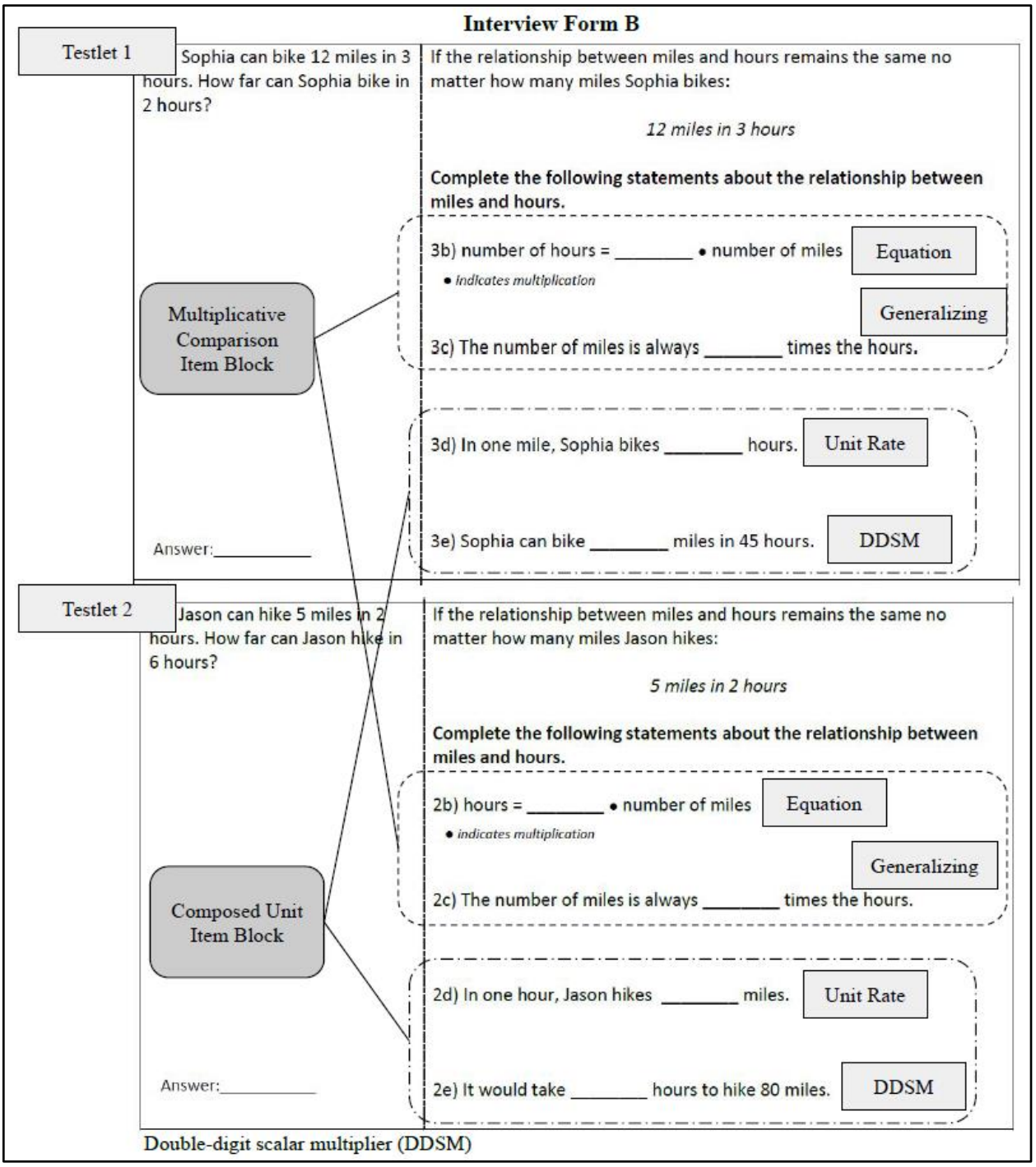


Figure 3. Steps of analysis for the score interpretation and response process.

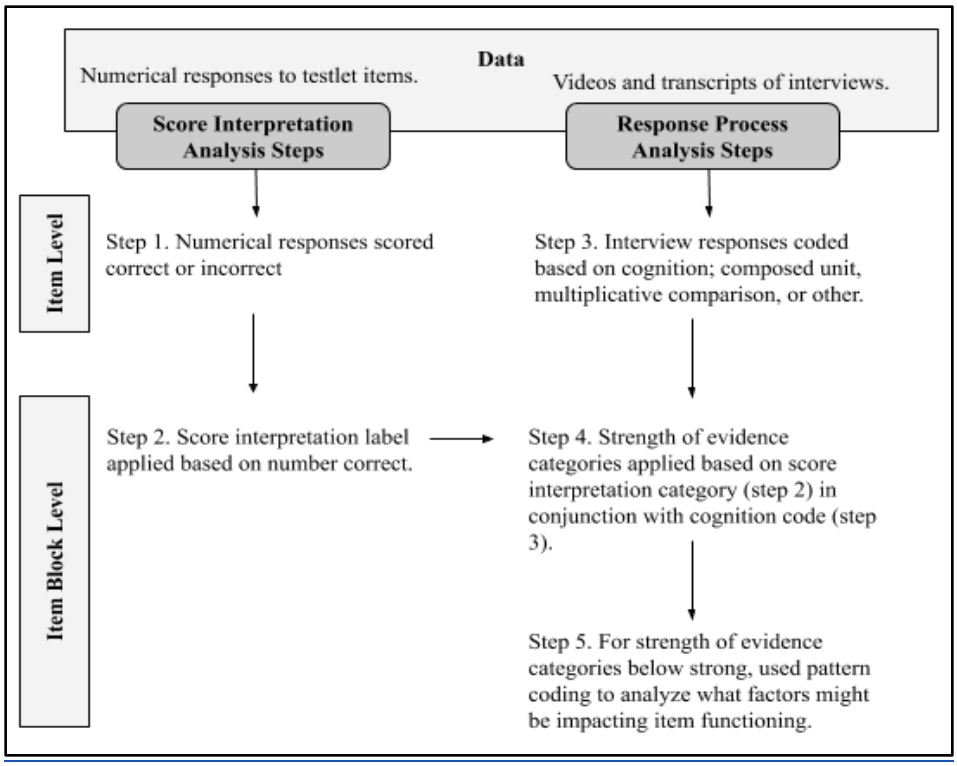
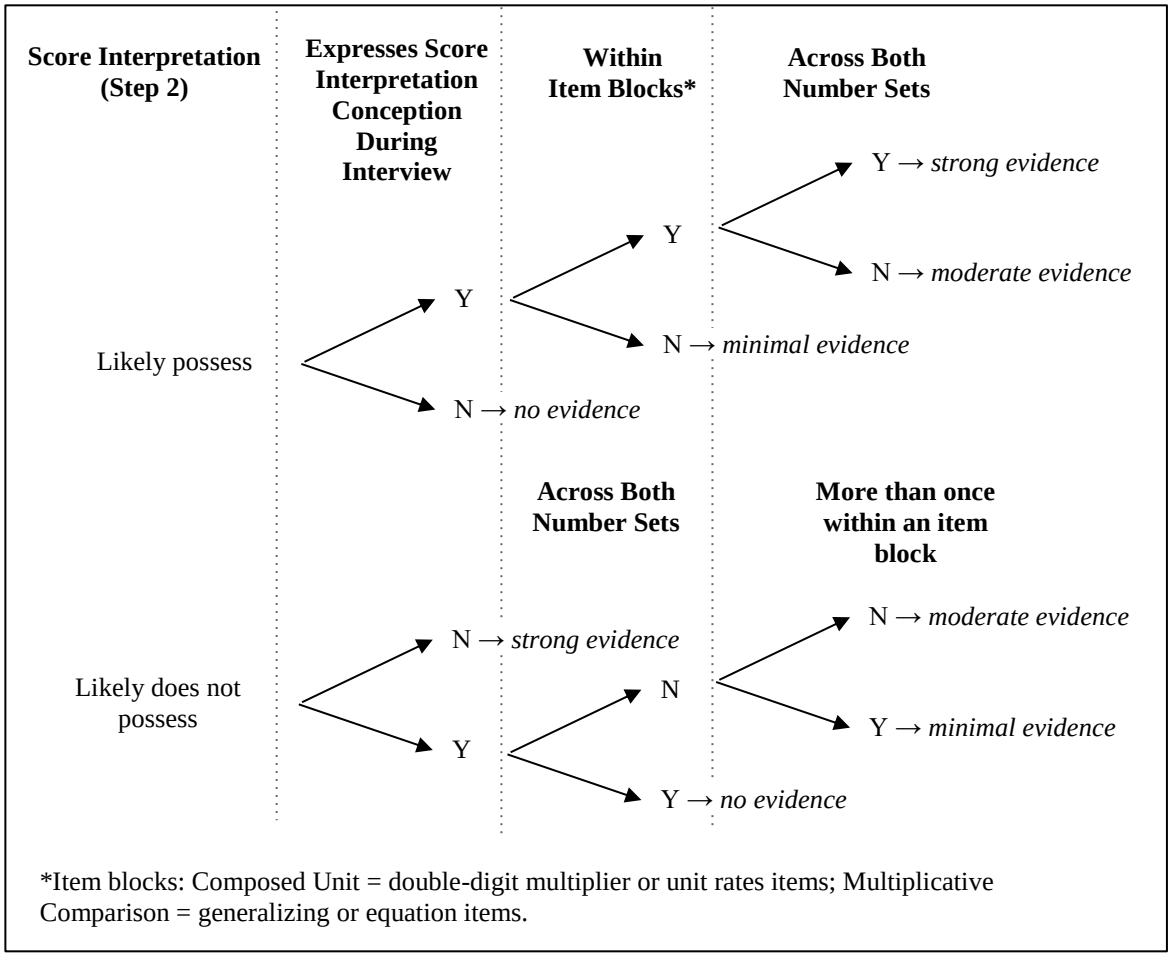


Figure 4. Strength of evidence criteria[3] used in Step 4 of the analysis.



[3] See strength of evidence table supplemental file for the detailed criteria.