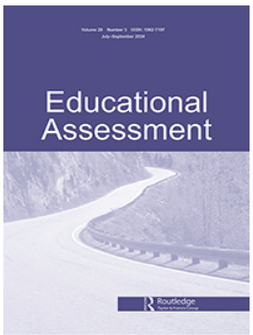


Routledge
Taylor & Francis Group

Educational Assessment

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/heda20

Using Evidence from Task-Based Interviews for Development and Validation of Classroom Assessments: Approaches and Applications in Early Grades Mathematics

Leanne R. Ketterlin-Geller, Muhammad Qadeer Haider & Jennifer McMurrer

To cite this article: Leanne R. Ketterlin-Geller, Muhammad Qadeer Haider & Jennifer McMurrer (16 Sep 2024): Using Evidence from Task-Based Interviews for Development and Validation of Classroom Assessments: Approaches and Applications in Early Grades Mathematics, Educational Assessment, DOI: 10.1080/10627197.2024.2398421

To link to this article: https://doi.org/10.1080/10627197.2024.2398421

EDUCATIONAL ASSESSMENT
https://doi.org/10.1080/10627197.2024.2398421

Routledge
Taylor & Francis Group

Using Evidence from Task-Based Interviews for Development and Validation of Classroom Assessments: Approaches and Applications in Early Grades Mathematics

Leanne R. Ketterlin-Geller[a], Muhammad Qadeer Haider[b], and Jennifer McMurrer[c]

[a]Southern Methodist University, Dallas, TX, USA; [b]University of Texas at Arlington, Arlington, TX, USA; [c]Gibson Consulting Group, Austin, TX, USA

ABSTRACT

This article illustrates and differentiates the unique role cognitive interviews and think-aloud interviews play in developing and validating assessments. Specifically, we describe the use of (a) cognitive interviews to gather empirical evidence to support claims about the intended construct being measured and (b) think-aloud interviews to gather evidence about the problem-solving processes students use while completing tasks assessing the intended construct. We illustrate their use in the context of a classroom assessment of an early mathematics construct – numeric relational reasoning – for kindergarten through Grade 2 students. This assessment is intended to provide teachers with data to guide their instructional decisions. We conducted 64 cognitive interviews with 32 students to collect evidence about students' understanding of the construct. We conducted 106 think-aloud interviews with 14 students to understand how the prototypical items elicited the intended construct. The task-based interview results iteratively informed assessment development and contributed important sources of validity evidence.

Cognitive interviews and think aloud interviews are two types of task-based interviews that may contribute empirical evidence to the development and validation of assessments. In general, task-based interviews are intended to elicit participants' cognitive processing while they are completing a carefully constructed task. Working one-on-one with a participant, an interviewer administers tasks while asking questions designed to better understand the participant's thinking. The specific task design and questioning techniques depend on the purpose of the interviews.

Cognitive interviews are used to investigate participants' cognitive processes as they relate to their underlying comprehension or understanding of the construct (Leighton, 2017). Results are used to understand how knowledge is represented, comprehended, and stored in long-term memory. Findings contribute empirical evidence to support claims about the definition of the construct being measured. In contrast, think-aloud interviews are designed to collect data about participants' problem-solving approaches while responding to a task (Ercikan & Pellegrino, 2017; Hubley & Zumbo, 2017). Results from think-aloud interviews contribute empirical evidence to evaluate whether the task elicits the intended cognitive processes that reflect understanding of the construct. Both types of interviews play an important and distinct role in developing and evaluating the validity of the uses and interpretations of an assessment, and yet, are often overlooked in assessment development and validation efforts.

The purpose of this article is two-fold. First, we describe and differentiate cognitive interviews and think-aloud interviews within the context of assessment development and validation. Referencing Figure 1, we provide an overview of the distinguishing features of these task-based interviewing

CONTACT Leanne R. Ketterlin-Geller lkgeller@smu.edu Southern Methodist University, P. O. Box 750114, Dallas, TX 75275

	Cognitive interviews	Think-Aloud interviews
Purpose within Instrument development	• Contributes evidence to define the construct including the conceptualizations (e.g., knowledge, skills, processes, strategies) that define the targeted content • Illustrates the depth and breadth of student thinking to help expand the definition of the construct • Provides evidence linking students' thinking with the content specified in learning progressions (e.g., entry and exit skills, intermediary phases, interconnections) • Capture students' misunderstandings or errors and how those impact their overall conceptualization of the content	• Contributes evidence to determine if an item model accurately capture students' knowing in the domain • Illustrates the reasoning strategies, understanding, and procedures students used when solving item models • Provides evidence about sources of construct-irrelevant variance and supports inferences about the accessibility of item models
Types of inferences	Contributes evidence to inferences made about the definition of the construct including the specification of the conceptualizations, the relationship between conceptualizations (e.g., compensatory, conjunctive, disjunctive), and the entry and exit skills.	Contributes evidence to inferences made about the alignment between the thinking that is elicited by the student when they respond to item models and the targeted content (e.g., knowledge, skills, processes, strategies)
Outcome within Instrument Development	Refine the construct to accurately identify the conceptualizations, relationship between the conceptualizations, and the entry and exit skills.	Refine the item models to maximize alignment with the construct and the elicited thinking, minimize sources of error or bias, and improve accessibility of the item models.
Research Questions	• What are students' conceptualizations (knowledge, skills, processes, strategies) of the content? • What prior knowledge do students elicit related to the content (e.g., entry skills)? • What connections with other concepts do they make (e.g., what type of relationships exists?)? • What conceptualizations indicate expertise commensurate with expectations for the targeted population (e.g., exist skills)? • At which grade do students demonstrate emerging and mastery level understandings on the content?	• What conceptualizations (knowledge, skills, processes, strategies) are elicited from the item models? Do these knowledge, skills, processes, and strategies align with the intended assessment target? • Is the item model comprehensible? • What aspects of the item model cause construct-irrelevant variance or are confusing for students? • What sources of bias may be included in the item model?
Characteristics of Interview Questions	• Open-ended questions • Probing questions are asked to understand students' thinking and their conceptualizations of the content • Allow students to broadly engage with concepts • Content related questions are not intended to be included on an assessment but may inform future assessment design decisions (e.g., it is not necessary to develop test specifications or scoring procedures)	Concurrent Questioning • Non-obtrusive prompts are posed to encourage the student to describe their thinking processes. • Observational data is recorded to monitor students' processes during problem solving. Retrospective Questioning • Questions are directly associated with the item model that is based on test specifications. Students' responses to the item models should align with the scoring procedures that will be applied in the operational assessment. • Probing questions are asked to understand rationale for the student's response to the item model itself.
Sample Interview Questions	• Can you show me what you are doing in your head? • How did you start thinking about this question? • What did you imagine when I asked you this question? Can show me in pictures, words, or numbers? • Tell me more about your answer. How did you decide this was the answer? • Can you tell me more about [conceptualization]? What else do you know? What questions do you have? • When have you seen this [manipulative/tool] before? Tell me what you know about it.	Retrospective • Describe how you solved the problem. What did you do first? • What was the problem asking you to do? • Were there any parts of the question that were confusing? • Were there any words that you didn't know or that were hard to understand? • How did you use the [manipulatives/tools] to answer this question? Have you used these [manipulatives/tools] before?

Figure 1. Distinctions between cognitive interviews and think-aloud interviews.

techniques, and summarize how the purpose and methodology contribute valuable data to inform inferences underlying the validity of the uses and interpretations of test results. Second, we illustrate the use of these task-based interviews within the context of an assessment development project intended to measure kindergarten through Grade 2 students' numeric relational reasoning skills. The classroom assessment developed through this project is intended to inform teachers' decisions about the design and delivery of instruction focused on the concepts underlying numeric relational reasoning. Because results are intended to inform young children's learning opportunities of a complex construct, validity evidence was needed to inform the definition of the construct and verify the cognitive processing elicited by the items, among other claims. As such, we employed cognitive interviews and think-aloud interviews to evaluate two key inferences, respectively: (a) The definition of the construct accurately represents students' mental representation of the mathematics concepts, and (b) the items elicit the intended mental representation of the construct.

Measuring mathematics constructs

As noted, we illustrate task-based interviewing techniques within the context of a classroom assessment of an early grades mathematics construct, numeric relational reasoning. This project provides an ideal context for juxtaposing the purpose and methodology of cognitive interviews and think-aloud interviews because of the complexity of accurately representing and assessing students' cognitive processes in mathematics. The National Mathematics Advisory Panel (2008) recognized two key cognitive mechanisms at play when students engage in mathematics. Namely, students develop mental representations of the content, and then use these representations to solve problems.

Mental representations as the basis for the construct definition

As students begin to learn mathematics concepts, they develop mental representations of the content, including types of knowledge forms, cognitive processes, and organization and structure of concepts. These mental representations are interrelated, and different theories of learning propose various mechanisms for their development (Mislevy, 2006). When designing an assessment, these mental representations are used to delineate the assessed construct. In assessment, the term *construct* refers to the knowledge, skills, processes, and abilities that are targeted by the assessment (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014). The construct should be defined with sufficient detail to clearly understand what aspects of students' mental representations are and are not targeted by the assessment (Pellegrino & Wilson, 2015). Some constructs may already be well defined in the literature; however, emerging constructs, such as numeric relational reasoning, may lack specificity and require additional elaboration. Defining the construct with such specificity promotes accurate interpretation of the students' scores or responses in relation to their mental representations of the content.

Gathering sufficient evidence to define the construct is often underemphasized during assessment development, with greater attention paid to designing items. Without sufficient evidence, test users run the risk of a "corruption of the interpretability of test scores" (Gorin, 2006, p. 33). Considerable variability exists in the evidence sources used in mathematics education to justify the definition of the construct. Examples include examining formal structures that define the discipline (e.g., theory, expert task analysis), observations of child development (e.g., classroom observations), indirect evaluation of the learning process (e.g., teacher surveys), and studies that track and monitor learning over time (e.g., longitudinal studies, in-depth case studies; Clements, 2007; Confrey, 2019; Duschl et al., 2011; Ketterlin-Geller et al., 2020; Penuel et al., 2014; Sarama & Clements, 2019). Although each source of evidence provides unique insights into the intended construct, few collect direct evidence of students' thinking.

In mathematics education, there is a growing trend to specify constructs as learning progressions or learning trajectories. Learning progressions describe how students' knowledge deepens as they advance their thinking within a domain (Pellegrino, 2014). As a representation of the construct,

learning progressions should be grounded in theoretical and empirical research on student learning. Graf et al. (2021) proposed a 15-step framework for validating learning progressions that extends from theory development to evaluating the efficacy of learning progression-informed instruction. Within their framework, the learning progression is defined through iterative cycles of experimentation, including conducting task-based interviews, teaching experiments, and psychometric modeling of field test data. At each step in the validation framework, data are used to examine inferences about the specificity of the learning. Direct evidence about student thinking supports these inferences.

Cognitive interviews provide direct evidence of students' cognitive processing that may help understand students' mental representations of the content, especially for emerging or hard-to-define constructs. Cognitive interviews (sometimes called cognitive laboratory interviews) are grounded in psychological interviewing methods as a way of investigating participants' cognitive processes and understanding how participants store and structure their knowledge in long-term memory (Leighton, 2017; Padilla & Leighton, 2017). During cognitive interviews, participants work on tasks that are often open ended and require multiple steps. The interviewer asks in-depth questions designed to better understand the participant's thinking (see Figure 1 for additional specification and examples). Often, responses are coded to illuminate anticipated or unanticipated mental representations of the content. Cognitive interviews have received limited attention in the research literature (Leighton, 2017).

Transforming students' mental representations into assessment items

The second cognitive mechanism students employ when engaging in mathematics occurs when students solve problems (National Mathematics Advisory Panel, 2008). When applying their knowledge to a problem scenario, students enlist information-processing mechanisms of cognition that bring forward their stored mental representations from long-term memory. Students then perform actions in their working memory that allow them to generate a response or provide a solution to the problem. Assessment designers create tasks that are intended to elicit the targeted mental representations with fidelity. Irrelevant processes that may impact students' ability to accurately demonstrate their understanding should be minimized.

Empirical evidence about students' response processes is warranted to evaluate the alignment between the intended construct and the thinking that is actually elicited by the tasks (Castillo-Diaz & Padilla, 2013; Ercikan & Pellegrino, 2017; Hubley & Zumbo, 2017; Peterson et al., 2017). Although response process data are identified by the Test Standards (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014) as an essential source of validity evidence, these data have historically been overlooked. Cizek et al. (2008) examined test reviews published in the *16th Mental Measurements Yearbook* by the Buros Institute of Mental Measurements and found that only 1.8% of the 283 reviews reported on response process data. However, with recent advances in computational abilities and the proliferation of computer-based tests, considerable attention has been placed on using advanced analytics and machine learning techniques to better understand students' responses processes (Oranje et al., 2017). These approaches vary widely, and range from examining item-level response times to studying the length and efficiency of students' action sequences to conducting direct observations of students' eye movements. These data may shed light on students' responding behaviors; however, deriving meaning from these data requires advanced modeling, specialized equipment, and may ask examinees to engage with complex tools (e.g., eye-tracking equipment; Oranje et al., 2017). With a low threshold for advanced technical skills and equipment, think-aloud interviews may be particularly useful and can be employed at various stages of task design, from initial prototyping through operational production.

Think-aloud interviews require participants to verbalize their cognitive processes involved in solving problems. As explicated in Figure 1, think-aloud interviews transform a problem-solver's covert thinking processes into overt and observable behavior so that the thinking processes can be documented and analyzed (van Someren et al., 1994). Following a structured and systematic protocol, the interviewer engages with participants to elicit the steps and strategies they used to complete

a problem-solving task, either during or immediately following the task (Padilla & Leighton, 2017). During task completion, nonintrusive prompting – often called concurrent questioning – is used to encourage the participant to verbalize what they are doing. Immediately after task completion, as part of retrospective questioning, the interviewer asks participants to reflect on the actions they performed to draw out the contents of their working memory. Insights about students' responding behaviors can inform inferences about alignment with the intended construct as well as task design features that support or hinder students' ability to demonstrate their knowledge, skills, or abilities. These data may be particularly useful when assessing examinees who may be unfamiliar with specific testing procedures or whose responding behaviors may be less predictable, such as young children.

Considerations when using cognitive interviews and think-aloud interviews

There are several notable considerations associated with using results from task-based interviews during assessment development and validation. It is important to recognize possible drawbacks at the outset so additional evidence may be collected to verify the findings. One primary concern is the ability of interviewees to respond to the interview questions with accuracy. In particular, young children may have difficulty recalling their thinking after they complete a task and/or accurately representing what was going through their minds (Leighton, 2017). Similarly, asking questions while they are completing a task may place an undue burden on their working memory that impacts how they work through the task. Two approaches are helpful to overcoming these issues. First, Peterson et al. (2017) recommend implementing practice opportunities for the interviewee. They suggest starting with a familiar task and asking the interviewee to describe their thought stream. Feedback can be provided to improve the specificity of their descriptions. Second, interviewees should be presented with tasks that are moderately difficult for them (Castillo-Diaz & Padilla, 2013). Tasks that are too difficult will cause a strain on their working memory and potentially impact their retrieval of information; tasks that are too easy rely on automatic processing so interviewees may not recognize their thinking processes.

Another concern associated with integrating findings from task-based interviews is the time and cost associated with collecting and analyzing data. To meaningfully incorporate the findings into construct definition and refining item models, sufficient time must be built into the assessment development process. Without sufficient time and resources to adequately analyze the data, the findings may be susceptible to a confirmation bias (Peterson et al., 2017).

Finally, concerns emerge about the generalizability of the findings from task-based interviews. Generalizability was previously considered by some qualitative methodologists as irrelevant to qualitative inquiry and controversial to others (Osbeck & Antczak, 2021). Maxwell (2021) argues the process of transferability (Lincoln & Guba, 1985) and external generalization to other settings, groups, or populations are critical in qualitative research to understanding how an outcome is attained within a given context. When using task-based interviews in assessment development, the intention is to provide substantive evidence underlying interviewees' thought processes. Obtaining varied perspectives that represent a range of students' characteristics may help illuminate or uncover unexpected outcomes (Peterson et al., 2017). These data should be used in conjunction with other sources of psychometric and quantitative evidence that may be generalizable to the broader population.

In the remainder of this article, we describe two studies in which we applied task-based interviews in the development and validation of a classroom assessment intended to measure an early mathematics construct, numeric relational reasoning. Because the purpose of this assessment is to inform teachers' instructional decisions, validity evidence is needed to inform the definition of the construct and verify the cognitive processing elicited by the items. Our intention is to highlight the unique value data from task-based interviews provide within the iterative assessment development and validation process; we are not proposing that these data supplant other sources of relevant evidence (e.g., psychometric analyses). First, we describe our initial efforts to define the construct. Next, in Study 1, we employed cognitive interviews to gather validity evidence about the construct definition. Then, in Study 2, we used think-aloud interviews to examine the alignment between the thinking elicited by

the item models and the intended construct. We explain how the results of each task-based interview informed the iterative development of the assessment and contributed meaningful validity evidence.

Measuring numeric relational reasoning

Numeric relational reasoning is the ability to mentally analyze relationships between numbers or expressions using knowledge of properties of operations, decomposition, and known facts (Baroody et al., 2016; Carpenter et al., 2005). Numeric relational reasoning is closely related to number sense, the ability to work with numbers flexibly (Gersten & Chard, 1999). Numeric relational reasoning is not a rote procedure that can be followed step by step, but instead requires students to reason strategically using their knowledge of number relations (Whitacre et al., 2016). Previous research has linked strong numeric relational reasoning skills with greater gains in other numeracy skills as well as future mathematics concepts such as algebra (Carpenter et al., 2005). Because of this predictive relationship to other important outcomes, numeric relational reasoning is a vital concept for instruction in kindergarten through Grade 2.

As part of a larger project, we drafted an initial definition of numeric relational reasoning that specifies the knowledge forms, cognitive processes, and organization and structure of the concepts students use when analyzing relationships between numbers or expressions. Following an approach proposed by Ketterlin-Geller et al. (2013), we framed the definition as a learning progression to denote how students develop greater sophistication and complexity in their numeric relational reasoning within and across grades. First, we conducted a thorough review of the literature to begin outlining the mental representations of numeric relational reasoning. Then, we convened internationally and nationally renowned experts to refine the numeric relational reasoning learning progression, generate a hypothesized developmental sequence, and propose the foundational and targeted levels of proficiency for each grade. Finally, we conducted a survey of 274 early elementary teachers to investigate their perceptions about the developmental appropriateness of the progression (Ketterlin-Geller et al., 2020).

The hypothesized numeric relational reasoning learning progression is organized into three targeted learning goals (relations, composition and decomposition, and properties of operations) that represent coarse-grained categories of understanding. Within each targeted learning goal, three to four core concepts provide additional specificity. Each of these core concepts is further broken down into fine-grained subcomponents that represent the knowledge forms (knowledge, skills) and cognitive processes (reasoning, strategies) underlying numeric relational reasoning. This structure is illustrated in Figure 2. The organization and structure of the mental representations are specified in the progression of subcomponents – ordered from least to most complex – within and across core concepts. A total of 58 subcomponents were specified in the initial hypothesized numeric relational reasoning learning progression.

Because of the expansiveness of the learning progressions, in this article, we narrow our focus to one core concept within the targeted learning goal of Relations titled Foundations of Operations. Students understand that the next number in the counting sequence is 1 more than the preceding number for all numbers and vice versa. Students generalize this knowledge to greater quantities such as 2 and 10 more or less. With continued experiences – with and without mathematical tools (e.g., hundreds chart), students begin to develop flexibility with numbers, laying the groundwork for composing and decomposing numbers quickly and efficiently (Baroody et al., 2016). These experiences help students develop a mental number line and support conceptual understanding of operations so students can develop strategies for addition and subtraction that do not rely on memorization (Dyson et al., 2013). We selected this core concept as an example because of the significant revisions that were made based in-part on evidence from the cognitive interviews. The initial subcomponents for this core concept were:

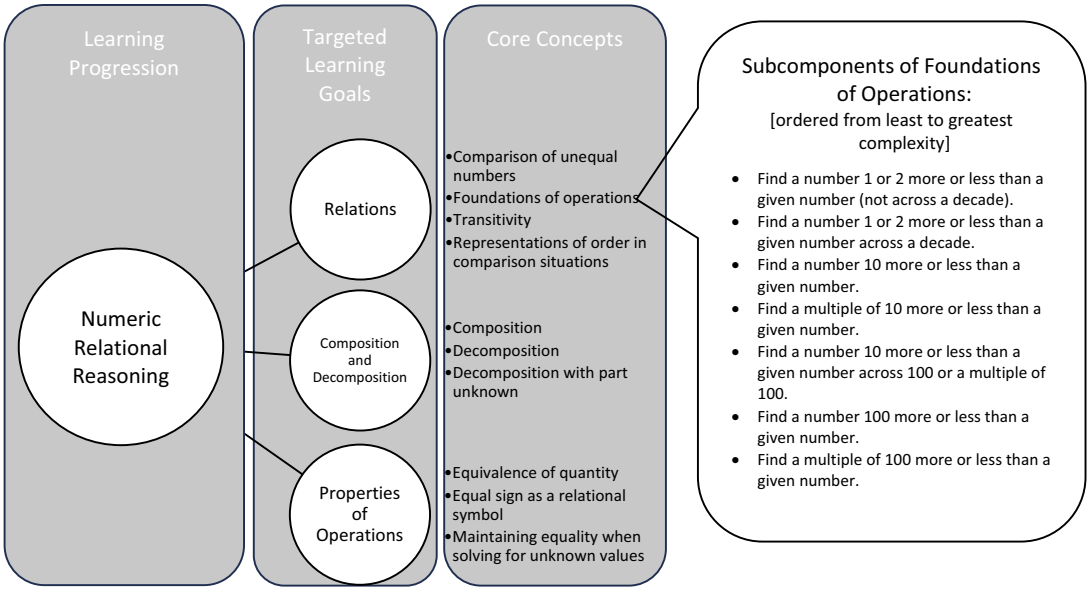


Figure 2. Structure of the numeric relational reasoning learning progression illustrating the subcomponents for the core concept of *foundations of operations* (adapted from Ketterlin-Geller, et al., 2022).

(1) Without counting, students use tools to find a unit more or less than a given number.
(2) Without counting, students mentally find a unit more or less than a given number.

In Study 1, we used cognitive interviews to gather validity evidence examining students' mental representations (e.g., knowledge forms, cognitive processes, and organization and structure of concepts) of numeric relational reasoning. Data collected from the cognitive interviews was used to refine the construct definition. After we developed items based on the finalized construct, in Study 2, we used think-aloud interviews to examine how students interacted with prototypical item models for the preoperational assessment. Data obtained from the think-aloud interviews informed the iterative design of the items by examining alignment with the intended construct and identifying item features that might introduce construct-irrelevant variance or obscure accurate measurement. As this process illustrates, data from task-based interviews contributed timely information to support assessment development decisions, while simultaneously substantiating our claims about the validity of the uses and interpretations of the results.

Study 1: Cognitive interviews

We conducted cognitive interviews to examine whether the learning progressions accurately represented students' mental representations of numeric relational reasoning, and asked:

(1) What are students' knowledge forms and cognitive processes related to numeric relational reasoning?
(2) What is the organization and structure of the knowledge forms and cognitive processes? Specifically, what is the sequencing (e.g., ordering from least to most complex) of the knowledge forms and cognitive processes within a core concept? At which grades do students demonstrate emerging and mastery levels of understanding?

Setting and participants for study 1

We interviewed 32 students in kindergarten through Grade 2 from three private parochial schools in a metropolitan area in a southern state. Students were purposefully selected to represent a range of prior knowledge and experience with early mathematics concepts, based on teachers' observations. Ten students from each grade level participated. Two additional students from Grade 3 also participated; their data are combined with the Grade 2 student data. Seventeen students were female. Other demographic data were not collected. Data from one student was excluded from analyses due to the limited amount of mathematical responses provided by the student.

Each student participated in two interviews for a total of 64 interviews. Each interview lasted 30–40 min. All interviews occurred over 5 weeks. Two trained researchers conducted interviews in person during the school day.

Instruments and procedures for study 1

The cognitive interview protocols included (a) mathematical tasks and (b) interview prompts (see Figure 3 for an example from the Foundations of Operations Core Concept). The mathematical tasks assessed students' thinking on a particular subcomponent of the numeric relational reasoning learning progression, and consisted of two parts: content questions (scored for correctness) and reasoning questions (qualitatively coded). We varied the number range of the items depending upon the child's grade level and readiness for the task. Tasks were sequenced from the least complex subcomponent on the learning progression to most complex. All students responded to all tasks regardless of the concept's hypothesized complexity or grade-level boundaries so we could collect empirical evidence to test these hypotheses.

The interview prompts were interspersed with the mathematical tasks to encourage students to describe and extend their thinking in a conversational manner. These open-ended prompts were often asked in response to the student's actions. Sample questions included, "How are you starting to think about [insert content]?" "I see that you are looking at [insert statement]. Tell me about what you are thinking." The cognitive interview protocol for the Subcomponent titled "*Without counting, students*

Mathematical task	Content question	Reasoning question	Interview prompts
The child is presented with a hundreds chart and counters.	Here is a number chart [Give child the hundreds chart].		Have you used a hundreds chart before?
	Can you show me where [number] is on the chart?		What do you know about hundreds charts? If you want to find any number, how can you use the hundreds chart? What are you starting to think about when you are looking at the hundreds chart?
	What number is 1 less than [number]?		Did you use the hundreds chart to find [answer]?
		How do you know [answer] is 1 less than [number]?	Where could I find 1 less than [number] on the hundreds chart? For any number on the hundreds chart, how can I find a number that is one less than a number?
	What number is 10 more than [number]?		Did you use the hundreds chart to find [answer]?
		How do you know [answer] is 10 more than [number]?	How does using the hundreds chart help you find numbers that are 10 more or 10 less than other numbers?

Figure 3. Example cognitive interview protocol for the subcomponent "*without counting, students use tools to find a unit more or less than a given number*" in the Foundations Of Operation core concept.

use tools to find a unit more or less than a given number" in the Foundations of Operations Core Concept is presented in Figure 3.

Two trained interviewers conducted the cognitive interviews using the interview protocols. Simultaneously, two trained observers took notes about the students' responses and general comfort with the tasks. The cognitive interviews were divided roughly evenly between each interviewer – observer pair. The observer noted when the interviewer asked follow-up questions that extended beyond the interview protocol. All interview sessions were videotaped.

Qualitative analysis

Prior to analyses, a research team member watched the videos and inserted students' nonverbal expressions, actions, gestures, and any pertinent interviewer actions into the audio transcriptions. Another researcher verified that the accuracy of the insertions of students' nonverbal actions for 20% of the interviews.

We used a two-cycle coding process of a deductive, a priori schema followed by open coding to search for emergent themes. We developed a priori structural codes based on our research questions (DeCuir-Gunby et al., 2011). We used a systematic procedure to screen the data constantly (Corbin & Strauss, 2015) to find evidence for a priori codes. Our coding team consisted of a lead coder, a secondary coder, and another researcher who helped the coders reach consensus. We analyzed students' reasoning regardless of correctness of their responses to the content questions. Using explicit code definitions as guides, we assigned the most accurate code for each child's reasoning response. We maintained a codebook of a priori codes for each targeted learning goal (Relations, Composition and decomposition, and Properties of operations).

After completing a priori coding, the reconciled file was distributed to the two coders with a new codebook for open coding. Open coding was iterative, using axial coding to analyze data and find codes within participants' voices (Saldaña, 2015). First, the coders independently coded using NVivo software and developed preliminary codes. Then, the two coders and the researcher held an intensive discussion to achieve group consensus on codes using a process called dialogical intersubjectivity (Brinkman & Kvale, 2015; Harry et al., 2005; Sandelowski et al., 2007). Finally, at the end of each meeting, the coders recorded in the final codebook the code names, definitions, examples of student talk and gestures from the data, and any exclusion criteria.

Quantitative analyses

To address Research Question 2, we scored students' responses to the content questions for correctness. To evaluate the sequencing or ordering of the subcomponents from least complex to most complex, we used classification accuracy statistics: (a) false positives, (b) false discovery rate, and (c) the proportion of students who were correct on the hypothesized more complex components and incorrect on the hypothesized less complex components (subsequently referred to as *c-prop*). False positives indicate the proportion of students who correctly answered the hypothesized more difficult subcomponents out of the total number of students who incorrectly answered the hypothesized less difficult subcomponents. The false discovery rate is the proportion of students who answered the hypothesized less difficult subcomponents incorrectly out of the total number of students who correctly answered the hypothesized more difficult subcomponents. Higher values could indicate issues with the sequencing or ordering of the subcomponents for both proportions. However, both values are conditional probabilities and do not reflect the entire sample. To address this issue, we computed the proportion of students who provided correct responses on the hypothesized more difficult components and incorrect responses on the hypothesized less difficult components (c-prop). Similarly, higher values indicated issues with sequencing or ordering of the subcomponents.

Table 1. A priori codes for foundations of operations core concept.

Code	Description
Counting tool use strategy	When describing the reasoning underlying their response, student uses and/or references a counting tool. Tools include colored square tiles or linking cubes as counters, or finger use.
Written number line strategy	Student uses and/or references a written number line or open number line.
Mental number line strategy	Student uses and/or references a mental number line.
Hundreds chart strategy	Student uses and/or references a hundreds chart.
Calculation strategy	Student uses and/or references calculating or employing a calculation strategy.
Other tool use strategy	Student uses and/or references any other tool (exclusive of counting tool, number line, or calculation).

For each item, we calculated the difficulty index in classical test theory estimates (p-values). The p-value is the proportion of students who responded correctly to an item out of all responses. A higher p-value indicated a less difficult item, and a lower p-value indicated a more difficult item. We disaggregated these indices by grade level to assess whether subcomponents performed differently by grade level.

Findings for study 1

To address Research Question 1, students' responses were analyzed to better understand their knowledge forms and cognitive processes. A priori codes were established for the Foundations of Operations Core Concept based on a comprehensive review of prior research, and the codes were used to determine whether students' observed reasoning aligned with anticipated strategies. All a priori codes were listed with a rich description of the code for a clear understanding of these codes, for example, the code "Counting tool use strategy" was explained as "When describing the reasoning underlying their response, student uses and/or references a counting tool. Tools include colored square tiles or linking cubes as counters, or finger use." A priori codes for the subcomponent presented in this illustration are described in Table 1.

Next, open coding was used to develop a more comprehensive understanding of students' knowledge forms and cognitive processes that might extend beyond the anticipated approaches. For open codes, we developed list of codes, rich description of each code, and a supporting example for each code. For example, Count all strategy was described as Student counted all numbers on the chart to find the number 10 more or less than the starting number, with superficial use of the tool. The code was supported with an example from the data as Explaining how student knows a number is 10 more:

Table 2. Open codes with examples for foundations of operations core concept.

Code	Description	Example
Count all strategy	Student counted all numbers on the chart to find the number 10 more or less than the starting number, with superficial use of the tool.	Explaining how student knows a number is 10 more: "We just start counting: 1, 2, 3, 4, 5, 6, 7, 8, 9" (Grade 2 student)
Count on strategy	Student counted on from a starting number, saying each number to arrive at the number 10 more. Tool was used superficially.	Explaining why student thinks 10 is 1 more than 8: "If you count up from 1, it's 9. Then you skip 9, it's gonna be 10" (Grade 1 student)
Count down	Student counted backward from a starting number, saying each number to arrive at the number 10 less. Some children counted aloud, and others wrote the numbers to count down, writing first the 1s, then the 10s place. Limited tool use.	Explaining how student knows a number is 10 less: "Like, the numbers that are before 92 [writes out numbers, 1s and then 10s] 82" (Grade 2 student)
Value of digits	Student compared the value of digits within numbers without explicitly using place value. Student correctly aligned numbers in the 10s and 1s places when comparing without explaining place value or providing a unit value distinction between the two numbers.	Explaining how student knows a number is 10 less: "When it's 10 less than a number, usually the first number goes 1 down and the next number stays the same" (Grade 2 student)

Table 3. Difficulty indices overall and by grade level for foundations of operations.

	Overall		Kindergarten		Grade 1		Grade 2	
Subcomponent	*N*	p-value	*n*	p-value	*n*	p-value	*n*	p-value
1	16	.44	5	.00	5	.40	6	.83
2	15	.33	4	.00	5	.20	6	.67

'We just start counting: 1, 2, 3, 4, 5, 6, 7, 8, 9' (Grade 2 student). Using the process described above, the codes presented in Table 2 were observed. Samples of students' utterances are also included in Table 2.

To address Research Question 2, quantitative data from students' responses to the content questions on the mathematical tasks were analyzed to better understand the organization and structure of the numeric relational reasoning learning progressions, specifically focusing on the sequencing of subcomponents and ordering across grades. The *p*-values are presented in Table 3. Overall, 44% of the students responded correctly to the items assessing the first subcomponent, whereas 33% responded correctly to the items assessing the second subcomponent. Broken down by grade, no students in kindergarten responded correctly to the content prompts for either subcomponent. In Grade 1, 40% responded correctly to the content prompts for the first subcomponent, and 20% responded correctly to the content prompts for the second subcomponent. In Grade 2, 83% (5 out of 6) students responded correctly to items assessing the first subcomponent, and 67% (4 out of 6) responded correctly to the items assessing the second subcomponent.

No false positives were observed for the Foundations of Operations Core Concept, meaning that all students who responded incorrectly to Subcomponent 1 ($n = 8$) also responded incorrectly to Subcomponent 2. Similarly, the false discovery rate was 0.0, meaning that all students who responded correctly to Subcomponent 2 ($n = 5$) also responded correctly to Subcomponent 1. The c-prop was also 0.0, indicating that of all the students who responded to items for both subcomponents ($n = 15$), no students responded incorrectly to Subcomponent 1 and correctly to Subcomponent 2. These findings should be interpreted with caution because of the small sample sizes; however, they suggest that the hypothesized sequencing of the subcomponents may accurately reflect a progression from least to greatest complexity.

Discussion for study 1

The purpose of the cognitive interviews was to examine whether the learning progressions accurately represented students' mental representation of numeric relational reasoning, and make iterative refinements as needed. As such, both qualitative and quantitative data were reviewed by the research team and integrated with other sources of evidence. To facilitate this process, the research team compiled summary documents for each Core Concept that included evidence used to draft the initial hypothesized learning progressions (e.g., literature, input from experts), results from the teacher survey, and cognitive interview data. What follows is a description of how these data contributed to the examination of the Foundations of Operations Core Concept.

Qualitative analyses to address Research Question 1 indicated that, for the students in our study, the learning progressions for the Core Concept of Foundations of Operations did not fully represent their mental representations of numeric relational reasoning. In reviewing these data in conjunction with prior literature, we acknowledged that the two existing subcomponents were too broad and were not sensitive enough to capture variability in students' thinking. For example, to probe students' understanding of Subcomponent 1, they were presented with a hundreds chart and asked to find a given number. They were then asked to identify a number that was 1 more or less and 10 more or less than the original number. Based on our a priori codes, we anticipated that students would use the hundreds chart to identify the new number. However, students across all grades counted using a strategy of either "counting all" or "counting on" to find the new number. When asked about their strategies, a student in Grade 2 explained, "Because I counted 10 and ended up at 97, 10 more

[than 87], you count 10 out and see what number you end up on." This student did not use the mathematical tool in the anticipated way but instead relied on counting.

Relatedly, for Subcomponent 2, students were not presented with a tool (such as the hundreds chart) and were asked to find 1, 2, or 10 more or less than a given number. Students did not consistently employ mental strategies as hypothesized in this subcomponent. For example, students "counted on" or "counted down" from the starting number, saying each number until they arrived at the target number. They might have evoked a mental number line as they said the counting sequence, but they relied heavily on counting single units. When asked to explain how they found their answer using pictures, words, or numbers, at least two students compared the value of the digits within numbers without explicitly referencing place value.

Although based on a very small sample size, quantitative data addressing Research Question 2 indicated that the sequencing of the items progressed from least to most complex. Students' responding patterns suggest that items associated with Subcomponent 1 were less difficult than items associated with Subcomponent 2. However, because no students in Kindergarten and few students in Grade 1 responded correctly, items assessing the subcomponents may be too difficult for students in kindergarten and Grade 1.

These observations of students' thinking prompted us to revisit the two subcomponents associated with the Foundations of Operations Core Concept (See Figure 5 for a depiction of the revised learning progression). We reviewed theoretical evidence found in existing literature and input from the panel of experts we assembled to draft the initial hypothesized learning progression. Research on the emergency of numeracy skills notes the importance of students developing fluency applying the *number-after rule*, which states that the sum of n +1 is the number after n in the counting sequence (Baroody, 2006; Baroody et al., 2012). Building on this literature, our initial hypothesized learning progression excluded counting and required students to use tools that would facilitate mental retrieval of this rule (e.g., number line, hundreds chart). However, based on the qualitative data from the cognitive interviews, we observed students across the grades using counting strategies. Returning to the literature, Baroody (2006) emphasizes the importance of helping students understand number patterns and connect their knowledge of the counting sequence with n +1 problems. By counting, children are building and reinforcing these connections, which will ultimately become automatic with additional exposure and experience. As such, we removed the stipulation that students could not count from the learning progression.

Relatedly, the initial hypothesized learning progression specified the subcomponents at a grain size that may not be sensitive to students' incremental learning. Even though students in Kindergarten and Grade 1 were able to respond to the mathematical tasks, few children answered correctly. We revised the learning progression to include more subcomponents that reflect smaller increases in students' mental representations of the construct. Notably, for students in Kindergarten working within the number range of 0–5, the learning progression was revised to focus on applying the number-after rule within the decade, and no constraints were noted about using counting or manipulatives. In later grades, the learning progression builds on these understandings and makes an explicit connection with place value concepts. This supports students' generalization of the number-after rule to more advanced applications.

This process was applied for all of the core concepts in the numeric relational reasoning learning progression.

Study 2: Think-aloud interviews

The purpose of Study 2 was to evaluate the extent to which the prototypical items elicited the intended mental representations of the construct as defined by the revised numeric relational reasoning learning progression. We developed prototypical item models to assess each subcomponent of numeric

relational reasoning across kindergarten through Grade 2. We used think-aloud interviews to examine how students interacted with these item models, and asked:

(1) Did students understand the instructions and actions that were embedded in the mathematical tasks?
(2) Were there any components of the mathematical tasks that appeared to introduce construct irrelevant variance, such as cause confusion, introduce bias, or otherwise obscure accurate measurement of the intended construct?
(3) Did the mathematical tasks elicit the intended content knowledge and reasoning strategies in students' responses?

Setting and participants for study 2

We purposefully selected 14 students to participate in the think-aloud interviews. Five students were in kindergarten, four students were in Grade 1, and five students were in Grade 2. Nine (64%) students were male, and five (36%) were female; no students identified as nonbinary. Further, ten (71%) identified as White, three (22%) as multiracial, and one (7%) as Black. Additionally, six (43%) students identified as Hispanic. One student indicated speaking a language other than English at home.

Each student participated in an average of eight interviews for a total of 106. Each interview lasted 35–45 minutes. Students participated in multiple interviews so that researchers could comprehensively examine all the appropriate subcomponents within a core concept of numeric relational reasoning using multiple item model formats. We also collected responses from students that were representative of different number ranges, with a meaningful sample across the grade levels. Number ranges varied to provide students with tasks that were at the appropriate difficulty level (Castillo-Diaz & Padilla, 2013). All interviews occurred over 3 months.

Due to the COVID-19 pandemic, we conducted the student think-aloud interviews virtually via the videoconferencing platform Zoom. We coordinated with parents so that they – or another caretaker – would be present with the study participant to assist with technology and help with any physical materials that we mailed to participants' homes.

Instrument and procedures for study 2

The think-aloud interview protocols included (a) mathematical tasks and (b) interview prompts. The mathematical tasks were designed based on a review of prior research, data from the cognitive interviews, and an inventory of existing assessments (McMurrer et al., 2021). Each task included two content questions, a reasoning question, and an extension question to elicit the range of thinking outlined in the learning progression. We developed and tested two to three prototypical item models (labeled Model A, Model B, etc.) for each subcomponent of numeric relational reasoning to examine the impact of different item features on students' task-elicited thinking.

The interview protocol included concurrent and retrospective questions (Ericsson & Simon, 1996). Concurrent questions were interwoven within the mathematical tasks to encourage the participant to continue talking while they responded to the questions. The interviewer allowed the participant to think for a few seconds after unobtrusively asking each question. If the participant was silent for 5 seconds, the interviewer used the following prompts to encourage the student to express all of their thoughts out loud: "What are you thinking now?" "Any other thoughts?" "Tell me how you decided to give me that answer." The interviewers also asked follow-up questions based on the participants' responses.

After completing the mathematical tasks, the interviewers asked retrospective questions to glean more information about the participants' problem-solving process and reasoning:

(1) Describe how you solved the problem; what did you think about first?
(2) What was this problem asking you to do?

(3) Were there any parts of the question that were confusing?
(4) Were there any words that you didn't know or that were hard to understand?
(5) How did you use the [insert manipulative, e.g., counters] to answer this question?

To further clarify or seek deeper understanding of the participant's response, sometimes the interviewer asked non-scripted follow-up questions.

One of two trained interviewers – accompanied by one of two observers – conducted each video-recorded session. The observers documented the participants' responses in real time using a spreadsheet aligned with the interview protocol. This format allowed the observers to quickly document, analyze, and sort the think-aloud interview data efficiently across all the interviews.

Analyses

During the think-aloud interviews, two observers noted the verbatim answers from the participants and their use of the manipulatives in real time. Immediately after each interview, the observers made preliminary reflections about the functionality of the items. Using a similar approach as was described in Study 1, the observers modified and expanded the analysis structure by informally open coding their notes (Corbin & Strauss, 2015). First, they looked across the responses to determine whether students understood the question. Simultaneously, they looked for evidence of construct-irrelevant variance. Then, to evaluate whether the tasks elicited the intended mental representations (knowledge forms and cognitive processes), they analyzed the students' problem-solving processes and reasoning alongside the intended content and reasoning responses by case and then across cases for each subcomponent (Ericsson & Simon, 1996).

Concurrent with the observers' analyses, the two interviewers followed a similar coding process using memos they wrote immediately following each think-aloud interview. Together, the research team reviewed the observers' analyses and the interviewers' memos by subcomponent for each of the core concepts to collaboratively draft summaries of their collective analyses. These multi-phased individual observation and coding sessions, which were interwoven with frequent and extensive group discussions, allowed the research team to iteratively and meaningfully analyze the data.

Findings for study 2

Because of the extent of the data, we illustrate the findings for one subcomponent in the Core Concept focused on Foundations of Operations (described above). There are seven subcomponents for this Core Concept, so at least 14 prototypical item models were tested. In Figure 4, we present prototypical Item Models A and B for the subcomponent "find a number 1 or 2 more or less than a given number within a decade." In this example, the primary differences between Item Models A and B were the item format (multiple choice compared to constructed response) and the use of mathematical tools including visual materials and manipulatives in the reasoning questions. In Model A, we asked students to use counters to show and tell us what numbers are one less and one more than the given number. In Model B, shown in the shaded column in Figure 4, we asked students which numbers are one less and one more than the given number and we provided them with response options rather than a physical manipulative.

To address Research Question 1, we examined the retrospective questions to identify any aspects of the item models that might have been confusing. For both item models, all but one of the students noted that they understood what the problem was asking them to do and did not find any parts of the questions confusing. These students also completed the tasks. Conversely, one student pointed to the number line in the Model B extension question and said, "It's confusing."

We coded students' responses to evaluate possible sources of construct-irrelevant variance to address Research Question 2. No evidence of construct-irrelevant variance was observed for Item Model A; however, some students played with the counters as they were responding to the questions, which might cause a distraction. For Item Model B, construct-irrelevant variance was observed when

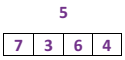
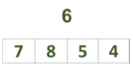
Subcomponent: Find a number 1 or 2 more/less than a given number within a decade.			
Question	Prototypical item Model A	Prototypical item Model B	Revised item model
Introductory question	"What is this number?" [wait] "Show me the number with counters." [Provide counters to student.] 	"What is the name of this number?" 	"Here is the number card and the same number of cubes. Name this number."  
Content Question 1	"Now tell me, what is one less than this number?"	"Which number is one less than 5?" 	"What number is one less than this number?"
Reasoning Question 1			"Show me with the [cubes] how you got your answer."
Content Question 2	"What is one more than this number?" 	"Which number is two more than 6?" 	"Here is the number card and the same number of cubes. Name this number. What number is one more than this number?"  
Reasoning Question 2	"Can you use the counters to show me how you got your answer?"	"How do you know that ___ is one more than 6?"	"Show me with the [cubes] how you got your answer."
Extension question	"Now use your counters to show me what two more than 6 would look like."	"Use this number line to show me how you know that ___ is one more than 6?" [provide number line]	

Figure 4. Example item models tested with kindergarten students (0–10 number range) on the core concept of foundations of operations during the Think-Aloud Interviews.

we asked students about displaying the answer choices. Students responded that the choices were not helpful, and one student said they "distracted me."

To address Research Question 3, we compared the alignment between the anticipated and the elicited knowledge and reasoning strategies. We observed that students either did not use or found the answer choices distracting in Model B when solving the problems. We also found that the Model A format elicited richer responses with students demonstrating their reasoning through their use of the physical manipulative. For example, a student explained, "I took 1 away to make it 1 less" while showing their work with the blocks.

Discussion for study 2

This subcomponent asks that students find a number 1 or 2 more or less than a given number within a decade. As noted at the conclusion of Study 1, students engage in these concepts as they work within the number ranges of 0–5, 0–10, and 0–19 (see Figure 5). The two item models we initially designed to assess this subcomponent varied in their item format and use of mathematical tools. These choices were based in part on the observations from Study 1, which indicated that students interacted with the content in both anticipated and unanticipated ways to demonstrate their knowledge and reasoning. In some instances, students used the mathematical tools provided, but many did not or only used them superficially. In designing the item models, we wanted to test whether providing a concrete representation (e.g., counters or other manipulatives) would help elicit the intended construct or whether students were able to demonstrate their knowledge without concrete representations.

Results from the think-aloud interviews suggest that concrete representations supported the participating students' engagement with the items and their ability to demonstrate their understanding. However, the choice of manipulatives was important to consider; the counters

<table>
<tr><td colspan="6">Foundations of Operations</td></tr>
<tr><td colspan="3">Kindergarten</td><td colspan="2">Grade 1</td><td>Grade 2</td></tr>
<tr><td colspan="6">Number Range</td></tr>
<tr><td>0-5</td><td>-10</td><td>-19</td><td colspan="2">-99</td><td>-999</td></tr>
<tr><td colspan="3">Find a number 1 or 2 more/less than a given number (not across a decade).</td><td colspan="3"></td></tr>
<tr><td colspan="2"></td><td colspan="2">Find a number 1 or 2 more/less than a given number across a decade.</td><td colspan="2"></td></tr>
<tr><td colspan="2"></td><td colspan="4">Find a number 10 more/less than a given number.</td></tr>
<tr><td colspan="3"></td><td colspan="2">Find a multiple of 10 more/less than a given number.</td><td></td></tr>
<tr><td colspan="5"></td><td>Find a number 10 more/less than a given number across 100 or a multiple of 100.</td></tr>
<tr><td colspan="5"></td><td>Find a number 100 more or less than a given number</td></tr>
<tr><td colspan="5"></td><td>Find a multiple of 100 more/less than a given number.</td></tr>
</table>

Figure 5. Revised version of Foundations Of Operations core concept.

were distracting for some students in our study. We revised the item to include grade-appropriate manipulatives that might be less distracting. We also ask students to use the manipulatives to show how they arrived at their answer, as was tested in Model A. In the revised item, the number is represented by both the mathematical symbol and a concrete representation of the quantity (see Figure 4). Presenting both representations may help students connect between different representations of the same relations among quantities (Venenciano et al., 2021). The revised item may better elicit students' conceptual understanding of these number relations.

Discussion

Developing assessments is an iterative process that involves collecting and integrating multiple sources of evidence at strategic decision points. As specified in the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014), appropriate sources of evidence are needed to justify assessment development decisions as well as evaluate the validity of the intended uses and interpretations of the results. It is important to align the sources of evidence with the underlying propositions claiming the information is trustworthy and meaningful for the given uses and interpretations. In this manuscript, we assert that data from task-based interviews – namely, cognitive interviews and think-aloud interviews – are viable and appropriate sources of evidence for two key inferences:

(1) The definition of the construct accurately represents students' mental representations.

(2) The items elicit the intended mental representations.

Providing validity evidence backing these two inferences is especially important when assessing hard-to-define constructs (such as numeric relational reasoning) with examinees who may have less experience engaging in assessment tasks (such as students in kindergarten through Grade 2).

To provide a comprehensive evaluation of the validity of the uses and interpretations of the classroom assessment reported in this manuscript, additional sources of evidence beyond those collected from task-based interviews are warranted. We briefly noted the additional sources of evidence we collected that extend beyond the purpose of this manuscript, such as theoretical information from prior research, extensive input from experts in the field, and data from surveys of teachers' enacted practices (Ketterlin-Geller et al., 2020). As previously noted, Graf et al. (2021) emphasized the cyclical and multi-phased nature of validating learning progressions-based assessments based on multiple sources of validity evidence. Aligned with this process, next steps for the work described in this manuscript include conducting a pilot test to examine item statistics and response patterns to better understand the sufficiency of the proposed scoring procedures. These data will be used to revise the items and scoring procedures for the content and reasoning questions. Subsequent studies will be conducted to further empirically recover the hypothesized learning progressions. As this manuscript is intended to illustrate, data from various sources may contribute meaningful evidence to evaluate inferences underlying assessment development and validation, and task-based interviews represent an under-utilized but important resource for test developers.

Application of task-based interviews to the instantiated example

To illustrate the application of cognitive interviews and think-aloud interviews during assessment development and validation, we described their role within a larger assessment-development project to design a classroom assessment for an early mathematics construct, numeric relational reasoning. Not only did these data contribute valuable insights to support our assessment development efforts, but concurrently, they contributed validity evidence needed to substantiate our claims that the scores provide trustworthy and meaningful information for teachers. To examine the inference that the definition of the construct accurately represents students' mental representations, in Study 1, qualitative and quantitative analyses of the cognitive interview data provided important insights into students' mental representations of numeric relational reasoning that were combined with additional sources of evidence to refine the initial hypothesized learning progression. Findings from Study 1 contributed to four primary revisions across the Core Concepts of the numeric relational reasoning learning progression intended to improve the accuracy of the construct definition:

(1) Using multiple sources of evidence, we refined the description of the knowledge forms and cognitive processes. Across the numeric relational reasoning learning progression, we added nine (15%) subcomponents that were previously unspecified. We removed or combined 12 (20%) subcomponents that were redundant or not observed in the data. The final learning progression included three targeted learning goals, 10 core concepts, and 55 subcomponents (a net reduction of three subcomponents).
(2) Combining quantitative data with other sources of input, we changed the sequence or order of the subcomponents within core concepts. From the original learning progression, we reordered 27 (46%) of the subcomponents to more accurately reflect a progression from least to most complex.
(3) Based on qualitative analyses of the cognitive interview data, students' knowledge about one subcomponent was evoked when responding to questions addressing another core concept. To address this discrepancy, we carefully examined the theoretical rationale for the initial association. Five (8%) subcomponents were moved because of these combined data.

(4) Because all students responded to interview questions for all subcomponents regardless of the hypothesized grade-level boundary, we were able to examine the appropriateness of the grade-level boundaries for each subcomponent. We adjusted the grade-level boundaries for 24 (41%) subcomponents.

Findings from the cognitive interviews led to direct improvements in the definition of the construct that more accurately reflect students' mental representations. Specific to the Foundations of Operations Core Concept, the revised Core Concept specifies in greater detail how students' knowledge and understanding progress from knowing that the next number in the counting sequence is 1 more than the preceding number to generalized applications that build transferable skills to support strategies for operations. These adjustments were subsequently shared with an independent panel of experts to verify the appropriateness of these modifications. Using the revised numeric relational reasoning learning progression, we drafted prototypical items as models for the operational assessment. Then, we conducted think-aloud interviews with these prototypes to examine the alignment between the construct and the task-elicited knowledge forms and cognitive processing for students.

To examine the inference that items elicited the intended mental representations, the data generated in Study 2 as part of the think-aloud interviews provided important insights into the functionality of the prototypical item models for the preoperational assessment and informed the iterative design of the numeric relational reasoning items. We learned about students' understanding of the mathematical tasks and whether the tasks elicited the intended mental representations of the construct. During the interviews, we administered at least two prototypical item models for each subcomponent of the construct. Drawing from these data, the observers and interviewers provided evidence-based recommendations about which item models (or a hybrid) functioned better to elicit cognitive processing that resembles the hypothesized reasoning strategies for the students. Changes to the items can be classified into one of four categories:

(1) One of the two tested prototypical item models – A or B – clearly functioned better than the other, as evidenced by students' understanding of the task, no apparent construct-irrelevant variance, and alignment with the hypothesized knowledge forms and cognitive processes underlying the subcomponent. In these instances ($n = 8$), we selected the better functioning item model for the operational test without substantive changes.
(2) One of the prototypical item models functioned better than the other, but evidence indicated that students misunderstood one component of the task, there were possible sources of construct-irrelevant variance, or the task-elicited response did not clearly align with the hypothesized subcomponent. In these instances, we selected the better item model but revised the problematic components. This occurred for 27 (49%) of the subcomponents. If significant changes were made, the new prototypical item model was retested through another set of think-aloud interviews.
(3) For four (7%) subcomponents, some aspects of both prototypical item models functioned well. For the final item model, we selected the best components of each item model without making substantive changes (e.g., content questions from Model A and reasoning questions from Model B).
(4) Some aspects of both prototypical item models functioned better than others, but evidence indicated that students misunderstood one or more components of each task, there were possible sources of construct-irrelevant variance, or the task-elicited response did not clearly align with the hypothesized subcomponent. For these 16 (29%) subcomponents, we selected the better functioning components across item models but revised the problematic components. If significant changes were made, the new prototypical item model was retested through think-aloud interviews.

Note that coding of these categorizations was completed by one to two members of the research team; a third member conducted a systematic verification of 25% of the categorizations that were only made by one team member.

We also identified aspects of the items (e.g., instructions, manipulatives) that might cause confusion, introduce bias, or otherwise obscure accurate measurement of the constructs. We adjusted the prototypical items to ameliorate these interferences. The revised item models were shared with an expert panel for a final review before implementing in the operational assessment.

Limitations

In both Study 1 and Study 2, the sample sizes were small. The qualitative data generated rich descriptions of students' thinking over the course of 202 task-based interviews; however, the small samples of students (32 for cognitive interviews and 14 for the think-aloud interviews) limit the generalizability of the results. Yet, these qualitative data – which are situationally- and culturally-embedded in the classroom context – are critical to understanding young students' conceptualizations and how they reason about the intended content. Although a larger, geographically dispersed sample may have appeared to improve the generalizability of the findings, it may have come at the cost of these important insights.

Another limitation cause by the COVID-19 pandemic is that the think-aloud interviews were conducted virtually via Zoom. The research team mailed materials to the participants before the interviews, and parents/caregivers helped to distribute the materials during the sessions. In some instances, the think-aloud interview sessions took longer due to the time the interviewer spent sharing instructions with parents/caregivers.

Application of Task-Based Interviews to developing and validating mathematics assessments

As the instantiated example illustrated, direct evidence about student thinking can be collected through task-based interviews. When applied to assessment development and validation, these data help build an evidentiary basis underlying the inferences that (1) the construct definition is an accurate representation of students' mental representations of the content and (2) the items elicit thinking that is representative and relevant of the construct. Although multiple sources of evidence are needed to support these inferences, as this manuscript demonstrates, task-based interviews contribute meaningful evidence beyond psychometric data.

In mathematics education, learning progressions are emerging as a common approach for specifying mathematics constructs (Confrey, 2019), and with this increased emphasis, research is evolving to guide efforts to validate these hypothesized progressions. Recently applied methods to validate learning progressions include conducting teaching experiments (c.f., Crawford, 2022), psychometric modeling (c.f., Attali & Arieli-Attali, 2019; Clements, 2007), surveying teachers (c.f., Ketterlin-Geller et al., 2020), and other approaches. However, similar to the process for evaluating test validity (c.f., American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014), evidence collected to validate learning progressions should be relevant and provide a sound scientific basis for the assertions about students' mental representations. Cognitive interviews directly elicit students' mental representations through structured questioning and careful analyses. Although insufficient as the sole source of evidence, the value of cognitive interviews cannot be overlooked. In this manuscript, we document how cognitive interview data provided an evidentiary basis for iteratively refining the construct definition until evidence converged indicating that the construct accurately represented students' mental representations of numeric relational reasoning.

Once sufficient evidence corroborates the construct definition, items are written to elicit the mental representations specified by the construct. Evidence about students' response processing can support inferences that the task-elicited thinking aligns with the construct. Multiple sources of data, including cognitive-psychometric modeling, eye-tracking studies, monitoring response times, and examining

other data from log files, can provide evidence about students' thinking processes as they engage with tasks (Hubley & Zumbo, 2017; Oranje et al., 2017). Methods and approaches for using these data are continuing to emerge as technological advances facilitate their use. This manuscript emphasizes the value of data collected from think-aloud interviews for evaluating students' response processes. Conducting think-aloud interviews requires qualitative methodological expertise, but the threshold for advanced technical skills and equipment is low for both the assessment developers and interviewees, and findings can be immediately used to refine items. Moreover, quantitative and psychometric data may help identify problematic items, but data from think-aloud interviews can illuminate *why* these items are problematic (Peterson et al., 2017). As such, when combined with other data, think-aloud interviews represent a useful tool for evaluating task-elicited cognitive processes. This manuscript documents how think-aloud interview data contributed validity evidence to justify our claim that the classroom assessment items accurately elicit the intended construct of numeric relational reasoning.

In sum, this article emphasizes the unique role data from two task-based interviews play in assessment development and validation. These data are intended to complement and not supplant other sources of data needed during assessment development and validation, such as rigorous psychometric analyses from field test data. Together, multiple sources of data support inferences that form a network of claims about score meaning and use.

Acknowledgments

We sincerely thank the participants who engaged in these studies as well as the researchers involved at various stages of this project including Dr. Lindsey Perry, Cassandra Hatfield, Dr. Eloise Kuehnert, Dr. Anthony Sparks, Josh Geller, and Tina Barton.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Science Foundation under [Grant No. 1721100]. Any opinions, findings, and conclusions or recommendations expressed in this materials are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Ethical approval

Research conducted under this project was approved by the Institutional Review Board of [Southern Methodist University]. All participants provided informed consent.

Informed consent

All participants provided informed consent.

References

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

Attali, Y., & Arieli-Attali, M. (2019). *Validating classifications for learning progressions: Framework and implications*. ETS Research Report Series.

Baroody, A. J. (2006). Why children have difficulties mastering the basic number combinations and how to help them. *Teaching Children Mathematics*, *13*(1), 22–31. https://doi.org/10.5951/TCM.13.1.0022
Baroody, A. J., Eiland, M. D., Purpura, D. J., & Reid, E. E. (2012). Fostering at-risk kindergarten children's number sense. *Cognition & Instruction*, *30*(4), 435–470. https://doi.org/10.1080/07370008.2012.720152
Baroody, A. J., Purpura, D. J., Eiland, M. D., Reid, E. E., & Paliwal, V. (2016). Does fostering reasoning strategies for relatively difficult basic combinations promote transfer by K-3 students? *Journal of Educational Psychology*, *108*(4), 576–591. https://doi.org/10.1037/edu0000067
Brinkman, S., & Kvale, S. (2015). *Interviews: Learning the craft of qualitative research interviewing*. SAGE.
Carpenter, T. P., Levi, L., Franke, M. L., & Zeringue, J. K. (2005). Algebra in elementary school: Developing relational thinking. *ZDM: The International Journal on Mathematics Education*, *37*(1), 53–59. https://doi.org/10.1007/BF02655897
Castillo-Diaz, 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. https://doi.org/10.1007/s11205-012-0184-8
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. https://doi.org/10.1177/0013164407310130
Clements, D. H. (2007). Curriculum research: Toward a framework for "research-based curricula. *Journal for Research in Mathematics Education*, *38*(1), 35–70. https://www.jstor.org/stable/30034927
Confrey, J. (2019, April). *Future of education and skills 2030: Curriculum analysis—A synthesis of research on learning Trajectories/Progressions in mathematics (EDU/EDPC[2018]44/ANN3)*. Organisation for Economic Co-operation and Development. https://www.oecd.org/education/2030/A-Synthesis-of-Research-on-Learning-Trajectories-Progressions-in-Mathematics.pdf
Corbin, J., & Strauss, A. (2015). *Basics of qualitative research: Techniques and procedures for developing grounded theor*^{y} (4th ed.). SAGE.
Crawford, A. (2022). Exploring a diverse learner's equipartitioning learning trajectory. *Investigations in Mathematics Learning*, *14*(4), 288–304. https://doi.org/10.1080/19477503.2022.2139090
DeCuir-Gunby, J. T., Marshall, P. L., & McCulloch, A. W. (2011). Developing and using a codebook for the analysis of interview data: An example from a professional development research project. *Field Methods*, *23*(2), 136–155. https://doi.org/10.1177/1525822X10388468
Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A review and analysis. *Studies in Science Education*, *47*(2), 123–182. https://doi.org/10.1080/03057267.2011.604476
Dyson, N. I., Jordan, N. C., & Glutting, J. (2013). A number sense intervention for low-income kindergartners at risk for mathematics difficulties. *Journal of Learning Disabilities*, *46*(2), 166–181. https://doi.org/10.1177/0022219411410233
Ercikan, K., & Pellegrino, J. W. (2017). Validation of score meaning using examinee response processes for the next generation of assessments. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 1–8). Routledge. https://doi.org/10.4324/9781315708591-1
Ericsson, K. A., & Simon, H. A. (1996). *Protocol analysis: Verbal reports as data* (Rev. ed.). MIT Press.
Gersten, R., & Chard, D. (1999). Number sense: Rethinking arithmetic instruction for students with mathematical disabilities. *The Journal of Special Education*, *33*(1), 18–28. https://doi.org/10.1177/002246699903300102
Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement Issues & Practice*, *25*(4), 21–35. https://doi.org/10.1111/j.1745-3992.2006.00076.x
Graf, E. A., van Rijn, P. W., & Eames, C. L. (2021). A cycle for validating a learning progression illustrated with an example from the concept of function. *Journal of Mathematical Behavior*, *62*, 100836. Article 100836. https://doi.org/10.1016/j.jmathb.2020.100836
Harry, B., Sturges, K. M., & Klingner, J. K. (2005). Mapping the process: An exemplar of process and challenge in grounded theory analysis. *Educational Researcher*, *34*(2), 3–13. https://doi.org/10.3102/0013189X034002003
Hubley, A. M., & Zumbo, B. D. (2017). Response processes in the context of validity: Setting the stage. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 1–12). Springer International Publishing.
Ketterlin-Geller, L. R. (2022). Designing and Implementing Assessments to Support Learning for Students who are Experiencing Difficulties Learning Mathematics. In Y. Ping, R. Tzur, & H. Thouless (Eds.), *Enabling Mathematics Learning of Struggling Students: International Perspectives*. Springer.
Ketterlin-Geller, L. R., Yovanoff, P., Jung, E., Liu, K., & Geller, J. (2013). Construct definition using cognitively based evidence: A framework for practice. *Educational Assessment*, *18*(2), 122–146. https://doi.org/10.1080/10627197.2013.790207
Ketterlin-Geller, L. R., Zannou, Y., Sparks, A., & Perry, L. (2020). Empirical recovery of learning progressions through the lens of educators. *Journal of Mathematical Behavior*, *60*, 100805. https://doi.org/10.1016/j.jmathb.2020.100805
Leighton, J. P. (2017). *Using think-aloud interviews and cognitive labs in educational research*. Oxford University Press.
Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. SAGE Publications. https://doi.org/10.1016/0147-1767(85)90062-8
Maxwell, J. A. (2021). Why qualitative methods are necessary for generalization. *Qualitative Psychology*, *8*(1), 111–118. https://doi.org/10.1037/qup0000173

McMurrer, J., Haider, M. Q., Hatfield, C., Holder, S., & Ketterlin-Geller, L. R. (2021). *Numeric relational reasoning: Think aloud interviews for assessment item development (Technical Report No. 21-07) [Technical Report]*. Southern Methodist University, Research in Mathematics Education.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed. pp. 257–306). American Council on Education.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the national mathematics advisory panel*. U.S. Department of Education.

Oranje, A., Gorin, J., Jia, Y., & Kerr, D. (2017). Collecting, analyzing, and interpreting response time, eye-tracking, and log data. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 25–38). Routledge. https://doi.org/10.4324/9781315708591-4

Osbeck, L. M., & Antczak, S. L. (2021). Generalizability and qualitative research: A new look at an ongoing controversy. *Qualitative Psychology*, *8*(1), 62–68. https://doi.org/10.1037/qup0000194

Padilla, J. L., & Leighton, J. P. (2017). Cognitive interviewing and think aloud methods. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 211–228). Springer International Publishing.

Pellegrino, J. W. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología Educativa*, *20*(2), 65–77. https://doi.org/10.1016/j.pse.2014.11.002

Pellegrino, J. W., & Wilson, W. (2015). Assessment of complex cognition: Commentary on the design and validation of assessments. *Theory into Practice*, *54*(3), 263–273. https://doi.org/10.1080/00405841.2015.1044377

Penuel, W. R., Confrey, J., Maloney, A., & Rupp, A. A. (2014). Design decisions in developing learning trajectories-based assessments in mathematics: A case study. *Journal of the Learning Sciences*, *23*(1), 47–95. https://doi.org/10.1080/10508406.2013.866118

Peterson, C. H., Peterson, N. A., & Powell, K. G. (2017). Cognitive interviewing for item development: Validity evidence based on content and response processes. *Measurement and Evaluation in Counseling and Development*, *50*(4), 217–223. https://doi.org/10.1080/07481756.2017.1339564

Saldaña, J. (2015). *The coding manual for qualitative researchers* (3rd ed.). SAGE.

Sandelowski, M., Voils, C. I., & Barroso, J. (2007). Comparability work and the management of difference in research synthesis studies. *Social Science & Medicine*, *64*(1), 236–247. https://doi.org/10.1016/j.socscimed.2006.08.041

Sarama, J., & Clements, D. H. (2019). Learning trajectories in early mathematics education. In D. Siemon (Ed.), *Researching and using progressions (trajectories) in mathematics education* (pp. 32–55). Brill.

van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modelling cognitive processes*. Academic Press.

Venenciano, L. C. H., Yagi, S. L., & Zenigami, F. K. (2021). The development of relational thinking: A study of measure up first-grade students' thinking and their symbolic understandings. *Educational Studies in Mathematics*, *106*(3), 413–428. https://doi.org/10.1007/s10649-020-10014-z

Whitacre, I., Schoen, R. C., Champagne, Z., & Goddard, A. (2016). Relational thinking: What's the difference? *Teaching Children Mathematics*, *23*(5), 303–309. https://doi.org/10.5951/teacchilmath.23.5.0302