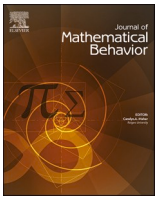
Journal of Mathematical Behavior 73 (2024) 101137

Contents lists available at ScienceDirect

Journal of Mathematical Behavior

journal homepage: www.elsevier.com/locate/jmathb



Validating a measure of graph selection and graph reasoning for dynamic situations

Courtney Donovan [1], Heather Lynn Johnson [*,2], Robert Knurek [3], Kristin A. Whitmore [4], Livvia Bechtold

School of Education and Human Development, University of Colorado Denver, United States

ARTICLE INFO

Keywords:
Mathematical reasoning
Graphs
Rasch model
College algebra
Assessment
Undergraduate education

ABSTRACT

Using a mixed methods approach, we report results from the evaluation and validation stages of a fully online Measure of Graph Selection and Reasoning for Dynamic Situations, implemented with undergraduate college algebra students across three U.S universities. The measure contains six items; each includes a video animation of a dynamic situation (e.g., a fishbowl filling with water), a declaration of understanding, four Cartesian graphs from which to select, and a text box for explanation. In the evaluation stage, we demonstrate usability and content validity, drawing on individual cognitive interviews (n = 31 students). In the validation stage (n = 673 students), we use Rasch modeling to evidence reliability and internal structure, establishing a continuum of item difficulty and confirming the viability of a partial credit scoring approach for graph selection. Rasch results provide statistical support that the theorized graph reasoning framework (Iconic, Motion, Variation, Covariation) from Johnson et al. (2020) forms a hierarchical scale.

1. Introduction

A growing number of mathematics education researchers have been working to create measures to assess students' mathematical reasoning and to test theoretical hypotheses (e.g., Kosko, 2019; Norton & Wilkins, 2009; Tzur et al., 2022). Via Rasch modeling (Rasch, 1980), mathematics education researchers have gathered statistical evidence to support theoretical progressions in multiplicative reasoning (Callingham & Siemon, 2021; Kosko, 2019; Tzur et al., 2022). With our study, we expand these efforts to include a Measure of Graph Selection and Reasoning for Dynamic Situations (MGSRDS). For decades, mathematics education researchers have theorized and studied students' reasoning about Cartesian graphs that represent relationships between attributes of objects in dynamic situations (e.g., Bell & Janvier, 1981; Carlson et al., 2002; Clement, 1989; Johnson et al., 2020; Kerslake, 1977; Leinhardt et al., 1990; Lee et al., 2020; Moore et al., 2019a; Thompson & Carlson, 2017). Our study extends this body of work by using Rasch modeling to test a theorized framework of students' graph reasoning from Johnson et al. (2020). Following Benson and Clark's (1982) classic instrument

* Correspondence to: School of Education and Human Development, University of Colorado Denver, Campus Box 106, P.O. Box 173364, Denver, CO 80217–3364, United States.
E-mail address: heather.johnson@ucdenver.edu (H.L. Johnson).
[1] https://orcid.org/0000–0001-5911–3294
[2] https://orcid.org/0000–0002-8865–2075
[3] https://orcid.org/0009–0004-9732–4403
[4] https://orcid.org/0009–0009-2025–7966

https://doi.org/10.1016/j.jmathb.2024.101137
Received 11 April 2023; Received in revised form 31 January 2024; Accepted 1 February 2024
Available online 8 March 2024

construction stages, we report on the evaluation and validation stages of the MGSRDS. For the evaluation stage, we report on cognitive interviews (n = 31), to examine usability and content validity. For the validation stage, we report results from two Rasch models, one for each of the constructs of graph selection and graph reasoning, collecting data from 673 undergraduate students enrolled in a college algebra course across three universities.

To begin, we provide our perspective on dynamic situations, graph selection, and graph reasoning. By "dynamic situation" we mean a situation that involves change and variation, such as a balloon being inflated with helium or a ball being tossed into the air. By "graph selection" we mean the process by which students choose a graph given a set of options to represent the dynamic situation. By "graph reasoning" we mean students' reasoning when they sketch, interpret, and/or select graphs, focusing on Cartesian graphs in particular. To theorize students' graph reasoning, we draw on the framework from Johnson et al. (2020), which puts forward four different forms of graph reasoning: Iconic, Motion, Variation, and Covariation. The Variation and Covariation constructs are rooted in Thompson's theory of quantitative reasoning (Thompson, 1994, 2011, 2022; Thompson & Carlson, 2017), which explains how people understand situations in terms of attributes that are possible to measure. The Iconic and Motion constructs are rooted in earlier research identifying students' challenges with Cartesian graphs (Clement, 1989; Kerslake, 1977; Leinhardt et al., 1990). To illustrate, consider a person, Nat, walking from a starting point to a large tree, then turning around and coming back. Now consider a Cartesian graph depicting Nat's total distance traveled on one axis and Nat's distance from the tree on the other axis. A student may expect a graph to look like the path Nat walked (Iconic), or to move back and forth, because Nat walked back and forth (Motion). Alternatively, a student may conceive of a graph representing only one of the distances (Variation), or a relationship between both distances (Covariation).

To assess students' graph selection and graph reasoning, we use the MGSRDS—a fully online tool housed in the Qualtrics platform that is accessible on mobile phones, tablets, or computers (Johnson, Olson et al., 2018, 2021). The multimedia MGSRDS contains six items, each spanning two screens. On the first screen there is a video of a dynamic situation, such as a toy car moving along a square track, and a question asking students whether they understand the situation. On the second screen, the video appears again, followed by four graphs from which to select, and a text box to explain the graph choice. We draw on students' graph choices and text responses on the MGSRDS as sources of data to evidence their graph selection and graph reasoning, respectively.

To evaluate the MGSRDS, we examine the tool's usability and content validity. Usability testing comes from the field of technology and design, as a means of examining the functionality of assessment items for the intended users (Riihiaho, 2018). We intend for the MGSRDS to be usable for college algebra students across devices (i.e., mobile phones, tablets, and computers). Content validity refers to evidence that assessment actually measures the intended constructs (DeVellis, 2003). For the MGSRDS, those constructs are graph reasoning and graph selection. To operationalize graph reasoning, we appeal to the Johnson et al. (2020) framework (Iconic, Motion, Variation, Covariation). For graph selection, we use a partial credit scoring approach (Incorrect, Partially Correct, Correct).

To validate the MGSRDS, we use Rasch models (Rasch, 1980). Rasch models are measurement models used for analyzing, calibrating, and creating measurements from categorical or ordinal data; they create a continuous scale from ordinal data using logits and confirm the hierarchical nature of the scale points (Wright, 1993). For example, with an agreement scale, it is typical to assume that the categories of strongly disagree, disagree, agree and strongly agree would appear in a certain order. By providing quantitative corroboration for the intended order of categories, a Rasch modeling approach can legitimize measures, such as the MGSRDS, by examining support for internal structure and scale validity (Bond & Fox, 2007).

In the evaluation stage, we examined usability and content validity for the MGSRDS, drawing on individual cognitive interviews (n = 31 students) and expert reviews. In the validation stage (n = 673 students), we investigated reliability and internal structure,

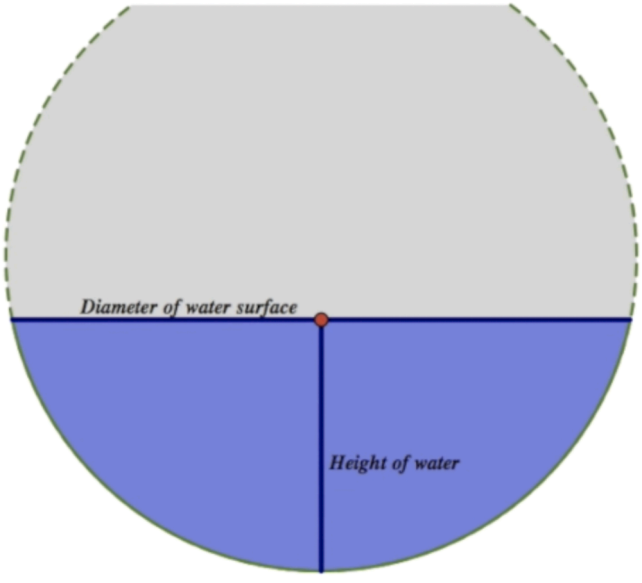


Fig. 1. Fishbowl dynamic situation.

examining whether the MGSRDS items provided a difficulty continuum for each of the constructs of graph reasoning and graph selection. Four research questions (RQs) guided our study:

1. Does the MGSRDS evidence usability and content validity?
2. Does the MGSRDS evidence reliability and internal structure for the constructs of graph reasoning and graph selection?
3. Do MGSRDS items provide a continuum for each construct (graph reasoning and graph selection), from most to least difficult?
4. Is there empirical evidence to support that the graph reasoning framework from Johnson et al. (2020) forms a hierarchical scale (Iconic, Motion, Variation, Covariation)?

To organize this paper, we begin by providing theoretical and conceptual background for students' graph reasoning. Then, we discuss details about the creation of the MGSRDS. In our methods and results, we separate the evaluation stage (RQ1) from the validation stage (RQ2–4). We conclude with contributions to research and practice.

2. Students' graph reasoning: Theoretical and conceptual background

The construct of quantity is a fundamental element of Thompson's (1994, 2011, 2022) theory of quantitative reasoning. Per Thompson's theory, a quantity is something more than a unit, such as feet, used as a label for a number (e.g., 5 feet). It is a person's conception of a measurable attribute of an object. For example, consider the dynamic situation of a fishbowl being filled with water at a constant rate (Fig. 1). Two attributes of the situation are the diameter of the water's surface and the height of the water. A student conceives of these attributes as quantities when they can think of them as measurable, which does not mean that they need to actually measure either attribute or assign numerical values to them.

In dynamic situations, attributes of objects undergo (or have the potential to undergo) change and variation. When a student conceives of some attribute as possible to measure and capable of varying, they are engaging in variational reasoning (Thompson & Carlson, 2017). For example, a student can conceive of the height of the water as a quantity that can vary as the fishbowl fills with water. Gross variation is an early level of variational reasoning; at this level, students conceive of the direction of change in an attribute. For example, in the fishbowl situation, the height of the water increases as the fishbowl fills with water. When a student conceives of change in two attributes varying simultaneously, they are engaging in covariational reasoning (Thompson & Carlson, 2017). For example, a student can conceive of the height of the water and the diameter of the water surface as varying together while the fishbowl fills with water. Gross coordination is an early level of covariational reasoning; at this level, students conceive of a loose connection between the direction of change in attributes. For example, the height of the water continues to increase while the diameter of the water surface increases, then decreases. With gross variation, students conceive of variation in individual attributes, and with gross coordination, students conceive of relationships between those attributes.

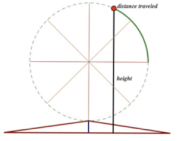
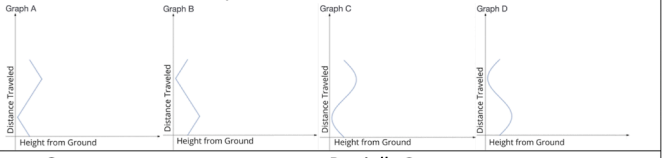
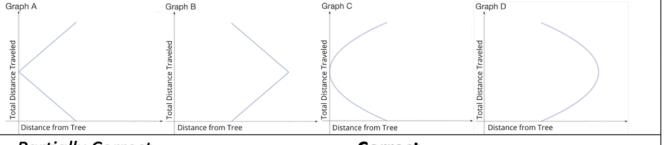
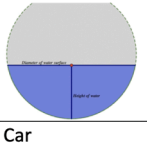
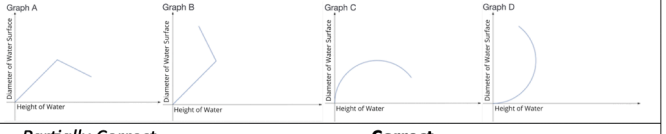
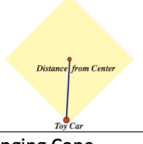
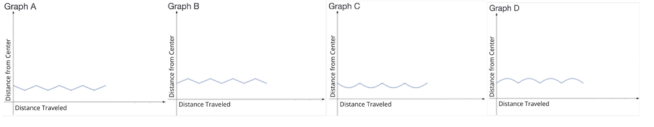
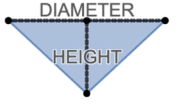
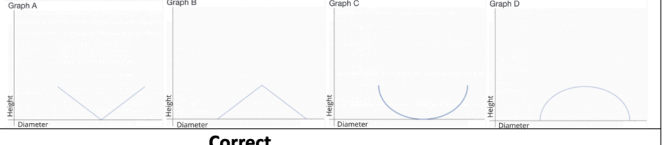
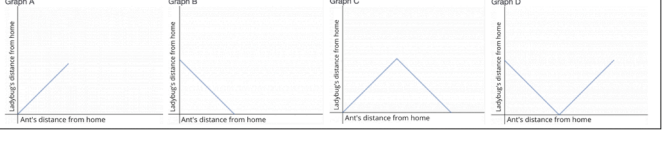
Students' reasoning intertwines with their conceptions of what graphs represent (Johnson et al., 2020). When engaging in covariational reasoning, students can conceive of graphs as representing relationships between varying quantities, such as the height and diameter in the fishbowl situation. When engaging in variational reasoning, students can conceive of graphs as representing a quantity varying along with experiential time (Thompson & Carlson, 2017). Students can use a single attribute, such as "diameter" to describe such a graph, or they may even wonder how it is possible to have a single graph that represents both height and diameter at the same time (see also Johnson et al., 2020). Per Thompson's (1994, 2011, 2022) theory both of these conceptions entail quantitative reasoning.

In contrast, students also can conceive of graphs in ways that are yet to evidence quantitative reasoning. Two such ways include iconic and motion conceptions of graphs. Students can conceive of graphs iconically, as resembling the characteristics of physical objects in a situation (Clement, 1989; Leinhardt et al., 1990). Students also may conceive of a graph in terms of the physical motion of objects in a dynamic situation (Bell & Janvier, 1981; Kerslake, 1977). For example, consider a graph representing a relationship between distance and time for a person's (Kim's) walk from home to school and back, up and down a large hill. Students with iconic conceptions expect such graphs to look like the hill depicted on Kim's walk (see also Kontorovich et al., 2019). Students with motion conceptions expect such graphs to show Kim's back and forth movement on the walk. Furthermore, it can be challenging for students to shift their motion conceptions of graphs, to form and interpret relationships between attributes represented in dynamic situations (Johnson et al., 2020).

Johnson et al. (2020) developed a framework categorizing students' graph reasoning into one of four forms: Iconic, Motion, Variation, and Covariation. The first two forms, Iconic and Motion, referred to students' graph reasoning in terms of physical features or observable movement in dynamic situations (Bell & Janvier, 1981; Clement, 1989; Kerslake, 1977; Leinhardt et al., 1990). The second two forms, Variation and Covariation, referred to students' graph reasoning that is at least at the levels of Gross Coordination or Gross Variation, per Thompson and Carlson's (2017) framework. Together, these four forms of graph reasoning described a range of students' conceptions of what graphs represent.

When sketching graphs, students' interpretations of the coordinate system itself can impact their graph reasoning (Paoletti et al., 2022b). Two such interpretations are spatial and quantitative (Lee et al., 2020). In a spatial interpretation, students conceive of a coordinate plane as representing a physical space either real or imagined, much like a map. In a quantitative interpretation, students conceive of a coordinate plane as a new space, disembedded from a physical situation, in which they can represent relationships between quantities. When students engage in iconic or motion reasoning, it suggests they are employing a spatial interpretation of a coordinate system (see also Paoletti et al., 2022b). In contrast, students' variational or covariational reasoning suggests a quantitative interpretation of a coordinate system.

Table 1
MGSRDS assessment items.

Assessment Item	Directions	Graph Choices
Ferris Wheel 	Select the graph that best represents a relationship between the Ferris wheel cart's **height from the ground** and the **distance traveled**, for one revolution of the Ferris wheel. Each graph shows the height from the ground on the horizontal axis and the distance traveled on the vertical axis. Explain why you chose the graph that you did.	*Partially Correct* **Correct** 
Nat & Tree 	Select the graph that best represents a relationship between Nat's **distance from the tree** and Nat's **total distance traveled**. Each graph shows Nat's distance from the tree on the horizontal axis and Nat's total distance traveled on the vertical axis. Explain why you chose the graph that you did.	**Correct** *Partially Correct* 
Fishbowl 	Select the graph that best represents a relationship between the **height of the water** and the **diameter of the water surface**. Each graph shows the height of the water on the horizontal axis and the diameter of the water surface on the vertical axis. Explain why you chose the graph that you did.	*Partially Correct* **Correct** 
Toy Car 	Select the graph that best represents a relationship between the toy car's **distance traveled around the track** and the toy car's **distance from the center of the track**. Each graph shows the toy car's distance traveled on the horizontal axis and toy car's distance from the center on the vertical axis. Explain why you chose the graph that you did.	*Partially Correct* **Correct** 
Changing Cone 	Select the graph that best represents a relationship between the cone's **diameter** and **height**. Each graph shows the cone's diameter on the horizontal axis and cone's height on the vertical axis. Explain why you chose the graph that you did.	*Partially Correct* **Correct** 
Ant & Ladybug 	Select the graph that best represents a relationship between the **ant's distance from home** and the **ladybug's distance from home**. Each graph shows the ant's distance from home on the horizontal axis and the ladybug's distance from home on the vertical axis. Explain why you chose the graph that you did.	**Correct** 

Note. From Johnson et al. (2024)

The framework from Johnson et al. (2020) makes a conceptual distinction compatible with the constructs of static and emergent shape thinking, put forward by Moore and Thompson (2015). Static shape thinking entails students conceiving of graphs as malleable, manipulable shapes. For example, students can think of a graph of a parabola as a physical object that can be moved up and down, left and right, or even stretched on a coordinate plane. Emergent shape thinking, on the other hand, entails students conceiving of graphs as in-progress traces created by covarying quantities, represented on each axis. For example, in the fishbowl situation, a student can conceive of a graph as representing an emerging trace of a relationship between the height of the water and the diameter of the water surface. Like static and emergent shape thinking, the Johnson et al. (2020) framework distinguishes between physical (Iconic, Motion) and quantitative (Variation, Covariation) forms of graph reasoning. Yet, students engaging in motion reasoning can think of a graph as an in-progress trace without conceiving of that graph as representing a relationship between covarying quantities. This suggests that it can be useful to distinguish students' conceptions of graphs as emergent traces from their conceptions of graphs as representing relationships between quantities (see also Paoletti et al., 2022b).

Employing a lens of quantitative and covariational reasoning, researchers have conducted interview-based, teaching experiment (Steffe & Thompson, 2000) studies investigating graph reasoning of middle and secondary students (e.g., Ellis & Grinstead, 2008; Johnson, 2012; Johnson et al., 2020; Tasova, 2022) and prospective mathematics teachers (e.g., Moore et al., 2013; Moore et al., 2019a, 2019b; Paoletti et al., 2018). An affordance of such studies, with participants rarely exceeding a few dozen, is the potential for detailed analysis to draw out nuanced claims related to students' reasoning. Yet, a challenge is to scale up studies, to make claims for larger sample sizes. One way to scale up is via online instruments such as graphing tasks. This has been useful for researchers investigating practicing teachers' graph reasoning (Moore et al., 2019a; Thompson & Carlson, 2017).

Two online instruments to investigate practicing teachers' graph reasoning included survey items (Moore et al., 2019a) and online animations (Thompson et al., 2017). Moore et al. (2019a) provided teachers (n = 45) with a static image of an unconventional graph (i.e., a graph of a "function" in which certain x values corresponded to multiple y values), along with a hypothetical student response to the graph (i.e., if x were a function of y, the graph could represent a "function"). Teachers then were to explain how they would respond to the student. By incorporating unconventional graphs, the researchers learned more about how teachers viewed the viability of student responses that broke from conventions. In a study of 487 practicing secondary mathematics teachers in the U.S. and Korea, Thompson et al. (2017) analyzed teachers' responses to a graphing task. First, teachers viewed an online animation in which the two variables' values changed concurrently. Each variable was represented by a dynamic segment varying along the axis of a Cartesian coordinate plane. One variable's direction of change remained constant (i.e., always increased), while the other variable's direction of change varied. Thompson et al. (2017) did find evidence of teachers' covariational reasoning, with higher levels of covariational reasoning per the framework from Thompson and Carlson (2017), being more prevalent among high school teachers from Korea than from the U.S. While these researchers implemented online instruments with practicing teachers, such instruments also could be viable for other populations, including undergraduate students.

3. MGSRDS: Background and description

To ground the design and development of MGSRDS items, we drew on results from qualitative studies led by Johnson, investigating secondary students' quantitative and covariational reasoning related to functions and graphs (e.g., Johnson, 2012, 2015; Johnson & McClintock, 2018; Johnson et al., 2020). In addition, we consulted empirical research exploring preservice and practicing teachers' conceptions of graphing conventions (e.g., Moore et al., 2014; Moore et al., 2019a), as well as theorized interpretations of students' conceptions of the coordinate plane (e.g., Lee et al., 2020; Paoletti et al., 2022b). Furthermore, we examined how other researchers developed and refined a multiple choice graphing item, based on the well-known bottle problem, to assess precalculus students' covariational reasoning (Carlson et al., 2010).

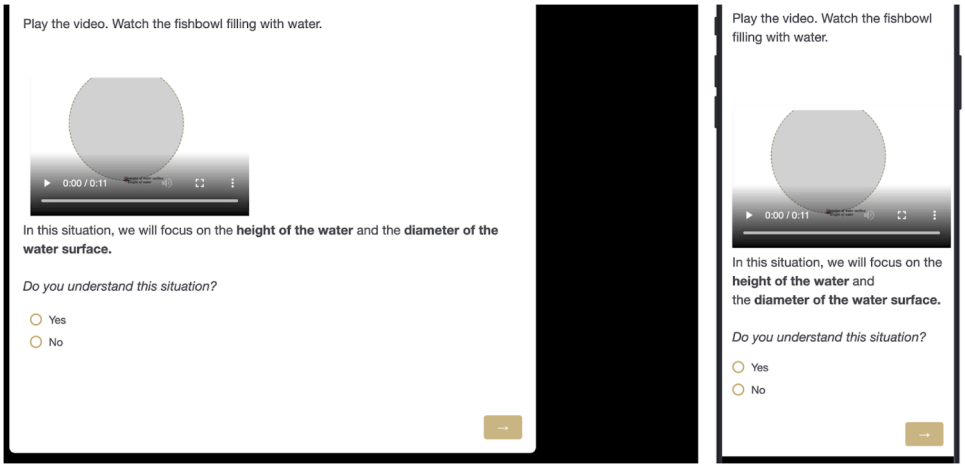


Fig. 2. Fishbowl item, optimized for computers/tablets (left) and mobile phones (right).

The MGSRDS has six items; each contains a different dynamic situation: a car moving around a Ferris wheel (Ferris Wheel), a person walking back and forth from a stationary object (Nat & Tree), a fishbowl filling with water (Fishbowl), a toy car moving along a square track (Toy Car), a cone expanding and contracting (Changing Cone) and two insects crawling back and forth along a path (Ant & Ladybug). Each assessment item has two consecutive screens. Screen 1 includes a video animation of the situation and a question asking students whether they understand the situation. Screen 2 includes the same video animation, written directions for graph selection, four graphs from which to select, and a text box for explanation. Table 1 shows an image of the situation, directions for graph selection, and graph choices for the six MGSRDS items. In Table 1, we have marked the correct and partially correct graph choices.

To optimize for access across mobile phones, tablets, and computers, as shown in Fig. 2., we developed the MGSRDS in Qualtrics (https://www.qualtrics.com). We followed principles for effective educational multimedia, including user manipulation, cueing, and pacing (Plass et al., 2009). Students could manipulate the videos by dragging the timeline below the video (Fig. 2). Boldface type helped to cue students to the attributes on which to focus. To control pacing, students could click an arrow in the lower right corner, but students were forced to make a selection before proceeding to the next screen (e.g., select yes or no). Furthermore, Qualtrics allowed for assessment items to appear in random order. Hence, we eliminated the placement of an item as a potential variable impacting the item's difficulty level.

We conducted expert reviews in an ongoing process as the MGSRDS was being developed; this included item demonstrations, experts taking the assessment individually, and group discussions. We consulted four experts, one with measurement expertise, one with technology expertise, and two with covariational reasoning expertise. One covariational reasoning expert recommended that we consider tasks in which both attributes varied in their direction of change (e.g., both increased and decreased). Furthermore, our measurement expert recommended that we incorporate items that would lead to a spread of difficulty; hence we developed the Ant & Ladybug item, to address both recommendations. In addition, our technology expert questioned whether open-ended text responses might be more challenging on mobile phones, but also felt that students' facility with phones could make this a nonissue. Hence, we added a question about the device students were using, so we could examine if there were any differences across devices.

Across the items, there are correct, partially correct, and incorrect graph choices. Five of the items in Table 1 (Ferris Wheel, Nat & Tree, Fishbowl, Toy Car, Changing Cone) contain correct and partially correct graph choices. For each of these items, the direction of change in one attribute remains constant (increasing), while the direction of change in the other attribute varies. The partially correct graph represents the same gross covariation (Thompson & Carlson, 2017) in attributes as does the correct graph, but it does not represent an accurate relationship between values of each attribute. For example, for the Fishbowl item both the correct graph (C) and partially correct graph (A) represent the height continuing to increase while the diameter increases, then decreases. However, graph A does not represent an accurate relationship between the values for height and diameter as the fishbowl fills with water. At the widest part of the fishbowl, small changes in the height are associated with greater amounts of change in diameter than near the top or bottom of the bowl. For the Ant & Ladybug item, both attributes vary in their direction of change, as each insect's distance increases and decreases as it moves to (or from) home and back. This item has no partially correct graph choice.

Across the MGSRDS items, graph choices come in related pairs, and none of the graph choices includes numerical amounts, similar to the covariation item developed by Carlson et al. (2010). For five items (Ferris Wheel, Nat & Tree, Fishbowl, Toy Car, Changing Cone), graph choices A and B are piecewise linear, and choices C and D are nonlinear. For the Ant & Ladybug item, graph choices A and B are linear, and choices C and D are piecewise linear. Three items (Ferris Wheel, Nat & Tree, Fishbowl) include unconventional graphs, for which the variable with a constant direction of change is represented on the vertical axis (see also Moore et al., 2014; Moore et al., 2019a). For each of these items, physical features of the correct graph choices resemble physical features of the situation (e.g., The Ferris wheel is curved and the correct graph is curved). For the other three items (Toy Car, Changing Cone, Ant & Ladybug), physical features of the correct graph choices do not resemble physical features of the situation (e.g., The toy car's track has straight edges and the correct graph has curved portions.), and the graph choices are more conventional.

Throughout the items, we have accounted for ways in which motion or iconic reasoning may lead to incorrect graph choices. For motion reasoning, there are distractors that resemble the movement of objects in the animation. For example, on the Nat & Tree item, Nat walks towards the tree, going left to right, then back again. Starting at the x-intercept, the incorrect graphs C and D extend to the right and then the left. On the Ant & Ladybug item, the incorrect graphs C and D represent one insect moving from and to home, and to and from home, respectively. On the Ferris wheel item, the cart moves at a constant speed, which students may associate with linear graphs, and the incorrect graphs A and B are piecewise linear. For iconic reasoning, there are distractors that resemble the images in the animation. For example, on the Fishbowl item, the incorrect graph D looks like the right hand side of the bowl. On the Changing Cone item, the incorrect graph A looks like the image of the cone. On the Toy Car item, the incorrect graphs A and B are piecewise linear, with corners that resemble corners on the track.

In our considerations for motion and iconic reasoning, we have conjectured that students may expect physical aspects of the animation to have actual locations on the coordinate plane. For example, on the Nat & Tree item, students may expect the tree to appear in the right hand portion of the coordinate plane, similar to where it appears in the animation; incorrect graphs C and D reflect this. In our view, these considerations share some interconnections with students' spatial conceptions of the coordinate plane (see also Lee et al., 2020; Paoletti et al., 2022b). Furthermore, while we have associated particular graphs with motion or iconic reasoning, we view these not to be the only form of reasoning that may lead to the graph choice (see also Paoletti et al., 2022b). For further discussion of theoretical considerations underlying graph choices, see Johnson et al. (2024).

4. Methods

To develop the MGSRDS, we follow Benson and Clark's (1982) classic instrument construction stages: planning, construction,

evaluation, and validation. In the planning stage, researchers establish purposes for the measure, define constructs, and identify a target group, in conjunction with an extensive review of the literature. In the construction stage, researchers identify indicators of the constructs and decide how to write items, or build assessments, establishing scales for items, based on each construct. In the evaluation stage, researchers conduct expert reviews and cognitive interviews to establish usability and content validity. In the validation stage, researchers gather evidence of reliability, internal structure, item difficulty, and scale hierarchy. We report on the evaluation and validation stages, yielding a mixed methods approach integrating qualitative and quantitative analysis (Tashakkori & Creswell, 2007).

This study is part of a larger, National Science Foundation funded project spanning three U.S. universities. Our target population is early undergraduate students enrolled in college algebra, an early undergraduate mathematics course with a history of challenges. These include large percentages of adjunct faculty and graduate students who teach the course (Tunstall, 2018), copious amounts of content for instructors to cover (Gordon, 2008), and textbooks focused on skills and procedures (Mesa et al., 2012). Furthermore, large percentages of first generation to college students, students of color, and low income students enroll in courses such as college algebra (Chen, 2016). Functions and graphs are central to the content of college algebra. Course topics emphasize properties of different types of functions, including piecewise, linear, quadratic, polynomial, rational, exponential, and logarithmic (Sullivan, 2020). Students are to graph these functions, explore relationships between different types of functions (e.g., inverse, composition), and use these functions to model different phenomena. With our study, we work to uncover the graph reasoning of this population.

4.1. Evaluation stage

In this section, we describe our methods for data collection, qualitative analysis, and quantitative analysis for the individual cognitive interviews (n = 31). Through these methods, we address usability and content validity for the MGSRDS (RQ1).

4.1.1. Cognitive interviews: Data collection and qualitative analysis

We conducted the individual cognitive interviews over a period of one year. Students were invited to participate via an announcement in their college algebra course, with a link for students to schedule a time convenient for them. Students received a $15 gift card for their participation. Interviews were held via videoconference, with Johnson serving as the interviewer and a graduate research assistant (GRA) observing. Students used a variety of devices, allowing us to examine usability and consistency of responses across computer, tablet, and phone. At the start of each interview, Johnson told students that we were testing out an assessment to learn more about students' mathematical reasoning. During the interview, Johnson asked students to read directions out loud and to explain their thinking. She invited students to explain their thinking either as they went along or after they made a graph selection. To mimic the actual assessment experience as much as possible, Johnson asked minimal follow-up questions. In general, these were clarifying questions to ensure that we accurately represented students' graph selections and reasoning. At the end of the interview, Johnson asked students to share what the experience was like for them.

To analyze students' graph selection, we used a spreadsheet to code responses as Correct, Partially Correct, or Incorrect. To analyze students' graph reasoning, we used codes based on the framework from Johnson et al. (2020). The codes were: Covariation (COV), Variation (VAR), Motion (MO), Iconic (IC), and Limited Evidence (LE). We added the LE code to account for student responses that provided insufficient evidence for one of the other four codes.

We used a consensus coding process (Olson et al., 2016) for students' graph reasoning. To begin, each interview was transcribed by a GRA. GRAs familiarized themselves with the codes and descriptions shown in Table 2, along with sample responses illustrating each code. Then two GRAs coded independently. They examined students' work item by item, in the order completed on the assessment, identifying evidence of any of the forms of graph reasoning for each item. For each item, GRAs recorded a single code for students' graph reasoning, using the highest level of reasoning that a student evidenced. For instance, if a student showed evidence of iconic, motion, and variational reasoning on a particular item, only VAR was used as the final code for the item. Separately, Johnson coded all 31 transcripts, meaning that each transcript had three coders. Then, Johnson looked at her codes against the GRAs' codes. For 12 students, there were at least two coded items on which Johnson and the GRAs disagreed. Without telling the GRAs the specific disagreements by code or item, Johnson asked the GRAs to go back and take another look at those transcripts, feeling free to change codes or keep them the same. This resulted in five remaining codes (3% disagreement) needing to be discussed between Johnson and the

Table 2
Graph reasoning codes (adapted from Johnson et al., 2020).

Code	Description	Example of Student Answers
COV	Students coordinate amounts and/or directions of change in one quantity with direction(s) of change in another quantity.	The diameter keeps increasing while the height increases and then decreases.
VAR	Students conceive of variation in the direction of change and/or amounts of change in a single quantity.	Because the diameter kept getting bigger.
MO	Students conceive of a graph as a close approximation (or literal representation) of the motion of a physical object.	She goes and returns like in the graph I picked.
IC	Students conceive of a graph as a close approximation (or literal representation) of the shape of a physical object.	It is a square and will have edges in the graph not curves.
LE	Students state that they don't know, or that they chose the best graph, or they provide an off task response.	I guessed. Made the most sense. n/a

GRAs, who met to finalize those codes. This process yielded 100% interrater agreement for the set of 186 responses to each item (31 students * 6 responses/student).

4.1.2. Quantitative analysis: Usability and content validity

To examine if differences occurred across devices, two one-way analyses of variance (ANOVA) were conducted. The dependent variables were graph reasoning scores and graph selection scores. To get these scores, we quantitized (Sandelowski et al., 2009) qualitative codes. For graph selection, we used: Correct–2, Partially Correct–1, and Incorrect–0. For graph reasoning, we used: COV–4, VAR–3, MO–2, IC–1, and LE–0 (see Table 2 for descriptions of these codes). The independent variable was technology, used with three categories: computer, tablet, and phone. Assumptions of normality and homogeneity of variance were tested and met (Leech et al., 2015).

To examine if response patterns in the interviewed sample (n = 31) matched responses patterns in the broader sample (n = 673), we conducted several correlations (Leech et al., 2015). For graph selection, we used the total number of items correct. For graph reasoning, we used responses with COV, VAR, and MO codes for each of the six items. The numbers of IC and LE codes in the interviewed sample were insufficient to conduct correlations (see Section 5.1.2 for further description on this). We then correlated frequencies of responses for each sample, setting all confidence levels at 95%.

4.2. Validation stage

In this section, we begin with a description of Rasch modeling, to situate our work in the validation stage. Then we describe our data collection, preparation, and analysis. Through these methods, we address reliability and internal structure for the MGSRDS (RQ2), create scales of item difficulty for graph reasoning and graph selection (RQ3), and provide quantitative corroboration that the graph reasoning framework from Johnson et al. (2020) (Iconic, Motion, Variation, Covariation) forms a hierarchical scale (RQ4).

4.2.1. Rasch modeling

Rasch models are ideal for establishing measures in educational settings because they follow a rigorous set of measurement rules that Rasch (1980) created to examine practical problems in real world settings where smaller sample sizes are typical. Linacre (2023) notes the goal of the Rasch model is not to fit the model to the data as other item response theory (IRT) models do. Instead, the goal is to examine how the data fits the established Rasch model, a theoretically and mathematically supported measurement model. Like the one-parameter IRT model, two independent parameters are estimated in Rasch models: person ability and item difficulty.

The Rasch model establishes a relationship between individual item properties on a particular instrument and the individuals that are responding to the instrument, while confirming a single underlying latent trait being measured (Boone et al., 2014; Glynn, 2012; Rasch, 1980). Rasch models assume that the latent trait is organized on a continuum and can determine a person's position on that continuum. Thus, the probability of a correct response or of endorsing any specific item is determined by both the item's difficulty and the respondent's ability (Boone et al., 2014; Glynn, 2012). For example, we would expect those with higher abilities to respond more positively to more difficult items. Rasch models are stricter measurement models than classical test theory models (e.g., confirmatory factor analysis) because they have established what model parameters should be for a measure to be considered "good." Therefore, the data must meet these model expectations.

The "ability" of a person is the probability they will endorse a specific item, so ability is really a function of the underlying latent trait and not necessarily a person's proficiency. For example, if an instrument is measuring a person's attitude, those with a higher ability, in Rasch terms, are those with a more positive attitude and thus more likely to endorse items that reflect a more positive attitude (Boone et al., 2014; Linacre, 2023). In the context of this study, higher ability for the graph selection construct translates to more correct graph choices, and higher ability for the graph reasoning construct translates to greater evidence of variational or covariational reasoning.

4.2.2. Data collection

Data collection took place during four consecutive semesters: Fall 2020, Spring 2021, Fall 2021, and Spring 2022. To facilitate data collection, instructors included an online module in their course learning management system, which described the study and gave students the option to opt out if they chose. Near the end of each semester, students (n = 673) completed the MGSRDS as part of their course, either during class or asynchronously, via a link that instructors shared in the course learning management system. Instructors did not know which students completed the MGSRDS, and there was no impact on students' course grades. At the start of the study, all three universities were classified as Hispanic Serving Institutions meaning that at least 25% of the undergraduate students identified as Hispanic or Latino. However, no student demographic information was collected.

4.2.3. Data preparation: Quantitizing qualitative codes

For graph selection and graph reasoning, we used the same codes and scores as the cognitive interview analysis (see Section 4.1.2). Similar to the qualitative analysis of the cognitive interviews (see Section 4.1.1), we used consensus coding (Olson et al., 2016) for graph reasoning, analyzing students' text responses explaining their reason for their graph selection. Again, GRAs led the coding process. GRAs coded individually, then met to calibrate their codes. As calibration decisions were made, we modified the coding rubric to add detail and examples that informed the qualitative coding. The qualitative analysis occurred in four rounds after each semester with three GRAs splitting the data so each response was coded by two GRAs. After individual coding, two GRAs met to discuss any disagreements. Eleven percent of the time they disagreed and had to bring a third GRA. This process necessarily resulted in 100%

interrater agreement for the set of 4038 responses (673 students * 6 responses/student).

While there was no missing data in this sample, data cleaning and considerations were needed. Six students (1% of the dataset) got all items correct, and another six students got all items incorrect. Because Rasch models account for floor and ceiling effects, and this was only 2% of the dataset, we did not remove these responses.

4.2.4. Data analysis

Because we examine two constructs, graph selection and graph reasoning, we created two Rasch models. The major assumption of Rasch models is that the data must be unidimensional (Linacre, 2023). Hence, we begin data analysis with principal components analysis of residuals to establish unidimensionality of each construct. If a measure explains 40% or more of the total raw variance, with the first contrast (which is equivalent to a second factor) having an eigenvalue of 2.0 or less, with less than 5% variance due to the first contrast, then there is sufficient evidence that the item set can be considered unidimensional (Linacre, 2023).

Using Winsteps software (Linacre, 2023), we examine two Rasch models, one for each of the constructs of graph selection and graph reasoning. We assess overall model fit with the standardized mean statistics (ZSTD), using both the infit (weighted fit statistic to control for extreme responses) and outfit (unweighted fit statistic), with values close to 0.0 indicating good fit. Item and person fit statistics are assessed with mean square (MNSQ) values. MNSQ is a statistic to test model fit that removes extreme observations, with MNSQ values between 0.5–1.5 being productive of measurement. While the model considers persons and items in the same manner mathematically, we expect items to "behave" more than persons. Persons (and items) with low MNSQ values overfit the model, meaning they are overly predictable. Persons with high MNSQ values underfit the model, meaning they create off-variable noise which degrades the measurement model and are more concerning. For the graph reasoning model, we have removed five students with MNSQ values over 2.0 and ZSTD values over 3.0. All students fit the graph selection model; hence, we have kept all students in that model.

Reliability refers to how consistently a student is responding, with Cronbach's alphas above 0.70 indicating a reliable scale (Tavakol & Dennick, 2011). Unique to IRT models is an added consideration of reliability looking at the separation of persons and items. Person separation explores the ability of items to identify levels of the measure across persons on a less-to-more continuum. Conversely, item separation explores the ability of persons to identify levels of the measure across the items on a less-to-more continuum (Bond & Fox, 2007). Practically speaking, we are looking for persons and items to be spread fairly evenly through the construct, so person and item separation values greater than 2.0 would indicate acceptable reliability (Linacre, 2023). Finally, we examine the invariance of items using differential item functioning (DIF). This statistic examines if items differ by groups. At 95% confidence, p values greater than 0.05 indicate no DIF, which means the item functions the same across groups. In this study, we have examined DIF across the three university sites to ensure items were functioning the same across student groups.

Rasch models present items and persons on the same scale to examine scale targeting and item difficulty using Wright maps (Linacre, 2023). In these maps, persons are represented as "#" on the left hand side of the vertical center line and items are represented on the right hand side of the vertical center line. Persons near the top of the left hand side demonstrate greater ability for each construct (and persons near the bottom demonstrate lower ability). Items near the top of the right hand side are "harder" (and items near the bottom are "easier"). Hence the Wright map provides a continuum of "easiest" to "hardest," thus creating a ruler for a construct. For graph selection, persons near the top demonstrate greater ability to select correct graphs, and items at the top are harder to get correct. For graph reasoning, persons near the top demonstrate greater ability to evidence covariational reasoning, and items at the top are those that were least likely for students to provide evidence of covariational reasoning (i.e., "harder").

Scale validation was vital, because we coded students' written responses based on the framework from Johnson et al. (2020), which prior to this study had not been quantitatively supported as a hierarchical scale of graph reasoning. The Rasch model tested if the order of the theorized categorical scale (Iconic, Motion, Variation, Covariation) was supported when changed into continuous logit values. If the theoretical ordering matched the mathematical model, we would expect to see category probability curves that indicated an even distribution of "hills" with clearly advancing steps in the theorized order with no evidence of misfit with category MNSQ infit values less than 2.0 (Linacre, 2023).

For scale validation, category probability curves depict the probability of a response in each scale category against the difficulty continuum shown in the corresponding Wright map. For example, if students are positioned with lower graph selection or graph reasoning ability as measured by the Wright map, it is expected that they would have a greater probability of providing incorrect responses or demonstrating iconic reasoning, respectively. If the category probability curves provide evidence that the constructs represent a hierarchical scale, they will follow observable patterns. The category probability curve representing the "lowest" construct on the scale will be monotonically decreasing, and consequently the "highest" will be monotonically increasing. The middle categories will increase, then decrease, with maxima falling between the lowest and highest category probability curves, ordered from left to right according to the theorized scale. When this pattern is not consistent, it provides evidence that the empirical ordering of categories is not consistent with the theorized ordering. In addition, Andrich Thresholds, which represent the probability of moving from one category to the next, should be ordered based on the theorized scale (Linacre, 2023).

5. Results

We begin by reporting on the evaluation stage, in which we address usability and content validity for the MGSRDS (RQ1). Then, we report on the validation stage, separating into three sections: Internal Structure and Reliability (RQ2), Item Difficulty (RQ3), and Scale Hierarchy (RQ4). Within each section, we report on both the graph reasoning and graph selection constructs.

5.1. Evaluation stage: Establishing usability and content validity for the MGSRDS

5.1.1. Usability of the MGSRDS

Students demonstrated that they understood the MGSRDS items. Across devices, students were able to view videos, select graphs, and enter text. Using analysis of variance at 95% confidence, we found no significant difference by device on students' reasoning and graph selection responses. Students spoke positively about the items and the overall assessment experience. The four responses shown in Table 3, in which students described what the experience was like for them, were representative of the broader set of student responses (all names are pseudonyms). Even when students experienced challenges, they described how they worked to make meaning, demonstrating the usability of the MGSRDS.

5.1.2. Content validity of the MGSRDS: Graph selection and graph reasoning

The interviews demonstrated that the graph choices were viable. Across the interviews, students selected all but three graph choices. For the graph choices that were not selected (Ferris Wheel Graph A, Nat & Tree Graph D, Toy Car Graph C), students considered them as potential options. For example, on the Ferris Wheel item, at first Darren said that he thought the answer could be Graph A or B, because "the cart is moving at a constant speed." However, he chose Graph D, because he experienced a similar situation during his algebra class, and the correct answer was nonlinear, like Graph D. On the Nat & Tree item, Jerry looked at the graph options and said "I really like D but at the same time I like B better." However, Jerry chose Graph A, saying "I realized when he walked further, the further he walked he got closer to the tree." On the Toy Car item, Emma eliminated graphs B and D, saying "it's going to start by decreasing and then increasing." To distinguish between Graphs A and C, Emma said "I'm going to pick graph A, because to me, it just makes more sense to move on a straight line when we're talking about a square and how it's going around the square track, as opposed to graph C, where it's curved lines." Because students' descriptions of their thinking demonstrated the viability of the three unselected options, we chose to keep all graph choices.

Students had a range of incorrect, partially correct, and correct graph choices. At one end of the spectrum, two students got all items either correct or partially correct (5 correct, 1 partially correct and 4 correct, 2 partially correct, respectively). At the other end of the spectrum, one student got zero items correct but four partially correct, and another student got one item correct and one item partially correct. For graph selection, the response patterns of interviewed students ($n = 31$) matched response patterns in the broader sample ($n = 673$). The correlation between interviewed and surveyed students' total items correct was 0.48, with most interviewed students getting 2–3 items correct and most surveyed students getting 1–2 correct.

The interviews demonstrated that the graph reasoning codes were viable. All 31 students completed all 6 items. With one code per item, there were 186 possible graph reasoning codes. Interestingly, COV had the greatest number of codes (103), followed by VAR (41) and MO (35), with IC (3) and LE (4) having only a few instances. We attribute the lower number of IC and LE codes to the fact that interviewed students could talk about their thinking, and thus had more opportunities to show evidence of their graph reasoning. Table 4 shows the response of a representative student, Sophia, on the Fishbowl item, in which she demonstrated evidence of iconic reasoning, along with other forms of graph reasoning, and hence received a code of a higher level of graph reasoning, per the Johnson et al. (2020) framework. Sophia demonstrated evidence of iconic reasoning ("because the bowl is a curved shape, the graph should not have any sharp corners"), variational reasoning ("one of two graphs where the height of the water does not decrease"), and covariational reasoning ("As the height of the water rises the diameter does increase and then gets small again at the top of the bowl"). Hence we coded Sophia's response as COV.

For graph reasoning, response patterns of interviewed students ($n = 31$) demonstrated consistency with the broader sample ($n = 673$). We were able to analyze correlations across the COV, VAR, and MO codes. The correlation across the COV code between the two samples was $r = 0.91$ and the MO code correlation was $r = 0.95$. The VAR code correlation was $r = 0.40$ with most interviewed students coded for VAR on the Toy Car item while most surveyed students were coded VAR evenly across Toy Car, Ferris Wheel, and Nat & Tree items. The interviewed students generally chose more correct graphs and demonstrated greater evidence of covariational and variational reasoning than the broader sample, which we attribute to their confidence to take part in an optional interview along with their verbal statements providing more evidence for coders than most written responses from the broader sample.

5.2. Validation stage: Internal structure validity and reliability

5.2.1. Graph reasoning: Internal structure validity and reliability

Graph reasoning was unidimensional with an eigenvalue of the first contrast being 1.41 and the raw variance of the measure

Table 3
Four student responses describing what the experience was like for them.

Student	Response
Sierra	This one was nice. I know that the one with the ant and the ladybug, I feel like I took a lot of time on that trying to understand like where I was going. I was trying to measure only one of the, I think, the bugs. But I have to take into account both of them.
Maya	Okay. Especially concerning the like, differentiating between the wavy graphs and the straight graphs. That was a bit challenging for me and I'm not quite sure that the reasons I gave for distinguishing them were correct, but I just felt that those reasons made sense.
Xander	It was nice. I got to think about what the graph was, how the situation and the graph were connected with each other.
Karly	Pretty challenging. I have trouble reading graphs a lot. I'm not a good graph reader. Um, usually have to look at the video a few times to understand at first which is why I would look at the video like twice or maybe three times to get a better understanding.

Table 4
Interviewed student response to the Fishbowl item.

Item	Student	Response
Fishbowl	Sophia	Verbal response to interviewer: "I'm guessing a curved one. As the height of the water rises the diameter does increase and then gets small again at the top of the bowl. And the height of the water doesn't go back down, so I would say this one." [Chooses Graph C] Written response in text box: "This graph one of two graphs where the height of the water does not decrease. Also, because the bowl is a curved shape, the graph should not have any sharp corners."

explaining 55.3% (Table 5). Overall fit was further supported with ZSTD infit value of 0.16 and outfit value of 0.13. All items fit the model well with MNSQ values between 0.80–1.17 (Table 6). All items contributed to the construct with point-measure correlations ranging from 0.56- 0.77. At 95% confidence, no significant differential item functioning was found between the three universities. Person reliability was lower than expected at 0.67 with 1.44 separation (Table 5). This showed that most college algebra students grouped closely together with their graph reasoning ability, in Rasch (1980) modeling terms. It also indicated that students received a range of reasoning codes across the MGSRDS items. Item separation was excellent at 12.01 with high reliability (Table 5). Hence, we confirmed reliability and internal structure for the graph reasoning construct.

5.2.2. Graph selection: Reliability and internal structure

Graph selection was unidimensional with an eigenvalue of the first contrast at 1.36 and the raw variance of the measure explaining 34.3% (Table 5). Overall fit was supported with ZSTD infit value of 0.03 and outfit value of 0.06 (Table 5). Items fit well between 0.75–1.33 and loaded onto the construct well with point-measure correlations between 0.46–0.55 (Table 6). At 95% confidence, no significant differential item functioning was found between the three universities. Person reliability was lower than expected at 0.44 with only 0.98 separation (Table 5). This showed most students were grouped closely together on graph selection ability, in Rasch terms (Rasch, 1980). It also indicated that students selected a range of correct, incorrect, and partially correct graphs across the MGSRDS items. Item separation was excellent at 11.25 with a reliability of 0.99 (Table 5). Hence, we confirmed reliability and internal structure for the graph selection construct.

5.3. Validation stage: Item difficulty

5.3.1. Graph reasoning: Item difficulty

The Graph Reasoning Wright map (Fig. 3) shows that items were spread well, creating a continuum of graph reasoning difficulty. The Ant & Ladybug and Toy Car items, positioned near the top of the Wright map, were the least likely to elicit responses demonstrating evidence of covariational reasoning, hence the "hardest." The Fishbowl and Changing Cone, positioned near the bottom, were the most likely to elicit responses demonstrating evidence of covariational reasoning, hence the "easiest." Items were well targeted with most students spread along the upper range of 0–2 logits (see Table 6). Therefore, items demonstrated a range of graph reasoning difficulty, aligned with the Johnson et al. (2020) framework.

5.3.2. Graph selection: Item difficulty

The Graph Selection Wright map (Fig. 4) shows that items were spread well to create a continuum of the graph selection construct. The Ant & Ladybug, much closer to the top than any of the other items, was considerably harder to get correct. The Toy Car and Changing Cone items, positioned closer together, were the next most difficult, followed by the Fishbowl and Nat & Tree items. The

Table 5
Dimensionality, fit, and separation.

Index	Graph Selection	Graph Reasoning
Dimensionality – eigenvalue for 1st contrast	1.36	1.41
Mean ZSTD Infit	0.03	-0.16
SD ZSTD Infit	1.04	1.17
Mean ZSTD Outfit	0.06	-0.13
SD ZSTD Outfit	0.99	1.10
Model Person Separation	0.89	1.44
Model Person Root Mean Square Error	0.65	0.76
Model Reliability of Person Separation	0.44	0.67
Cronbach's Alpha	0.44	0.84
Model Item Separation	11.25	12.01
Model Reliability of Item Separation	0.99	0.99

Note: Values are for the graph selection scale with partial credit and the graph reasoning scale without the 0 (no evidence) code. ZSTD Infit is a t statistic testing model fit with sensitivity to midrange observations. ZSTD Outfit is a t statistic testing model fit with sensitivity to extreme responses. Person/Item Separation is the ratio of the true standard deviation to the error standard deviation. Person Root Mean Square Error is standard error of the measure inflated for misfit. Reliability of Person/Item Separation = Separation^2 / (1 + Separation^2).

Table 6
Item fit statistics.

Item#	Mean (SD)	Logit Position	*SE*	Infit MNSQ	Pt-Measure Correlation
Graph Selection					
Ant & Lady Bug	0.35 (.76)	1.24	0.07	1.33	0.50
Toy Car	0.79 (.73)	0.20	0.05	0.75	0.52
Changing Cone	0.87 (.78)	0.05	0.05	0.81	0.52
Fishbowl	1.10 (.88)	-0.36	0.05	0.93	0.55
Nat & Tree	1.14 (.93)	-0.43	0.05	1.20	0.46
Ferris Wheel	1.28 (.91)	-0.69	0.05	1.16	0.50
Graph Reasoning					
Toy Car	1.58 (1.26)	0.94	0.07	0.88	0.66
Ant & Lady Bug	1.58 (1.26)	0.89	0.07	1.10	0.56
Nat & Tree	1.90 (1.36)	0.25	0.06	0.80	0.70
Ferris Wheel	1.84 (1.51)	0.08	0.07	0.98	0.73
Changing Cone	2.17 (1.70)	-1.02	0.07	1.17	0.75
Fishbowl	2.29 (1.68)	-1.14	0.07	0.87	0.77

Note. Item logit position is the value seen in the Wright map which creates the construct's continuum. SE is standard error of the logit position. MNSQ is a statistic testing model fit that removes extreme observations with values between 0.5-1.5 being productive of measurement (Linacre, 2023). Pt-Measure correlation is the relationship between the individual item and the total measure.

Ferris Wheel item was the easiest for students to get correct. Items were well targeted to the students, with most students grouped on the lower end of the construct at 0–1 logit positions (see Table 6). Hence, most students were getting a few items correct but rarely got all six correct. This fit our design intentions, to create a measure that was neither too easy (most students get all items correct) nor too difficult (most students get no items correct).

5.4. Validation stage: Scale hierarchy

5.4.1. Graph reasoning: Scale hierarchy

First, we conducted Rasch modeling, using quantitized (Sandelowski et al., 2009) graph reasoning codes for the five-code scale shown in Table 2 (LE–0, IC–1, MO–2, VAR–3, COV–4). However, there were major problems. While categories were used fairly evenly ranging from 10%– 29% observed (Table 7), the categories were disordered, as shown by the category probability curves (Fig. 5) and Andrich Thresholds (Table 7). Even though the LE and COV curves followed expected patterns (monotonically decreasing and increasing, respectively), there are no distinct maxima moving from left to right for the middle categories. Furthermore, the point of equal probability between the IC and MO categories, the Andrich Threshold for the MO category, was less than the threshold for the IC category. This also happened for the COV and VAR categories, with the COV threshold being less than the VAR. We tried multiple category versions with collapsing categories but removing the LE code worked best. This decision also had strong conceptual grounding because the LE code was meant to capture all written responses that lacked evidence of one of the forms of reasoning in the Johnson et al. (2020) graph reasoning framework.

We conducted Rasch modeling again, this time with a four-code scale (IC–1, MO–2, VAR–3, COV–4). Removing responses receiving an LE code, we based this model only on those responses coded IC, MO, VAR, or COV. All categories demonstrated good fit and were used fairly evenly ranging from 14%– 35% observed (Table 7). While the VAR category was slightly muted as it did not peak as strongly (Fig. 6), the category probability curves showed that its ordering was distinct. Furthermore, the Andrich Thresholds demonstrated steps from one category to the next, according to the theorized scale (Table 7). Hence, we provided quantitative corroboration to support that the graph reasoning framework (Iconic, Motion, Variation, Covariation) from Johnson et al. (2020) forms a hierarchical scale.

5.4.2. Graph selection: Scale hierarchy

The graph selection scale (Incorrect–0, Partially Correct–1, Correct–2) worked well as a hierarchical scale. All categories demonstrated good fit and were used fairly evenly ranging from 21%– 44% observed (Table 7). Although the Partially Correct category was more muted than others (Fig. 7), the ordering was distinct. Because the Ant & Ladybug item did not have a Partially Correct option (see Table 1), we attempted a model with a dichotomous scale (Incorrect–0, Correct–1). All item ordering and model statistics were consistent with a dichotomous scale, but person reliability became extremely poor. This pointed to the need for a three-point scale, including Partially Correct, to differentiate students' graph selection abilities.

Because the Ant & Ladybug was the only item without a Partially Correct option, one might expect it to be a problematic item. A potential challenge could be that the limited graph selection options (Correct and Incorrect only) would contribute to less variance and therefore greater difficulty. However, the Ant & Ladybug item had a similar standard deviation to Toy Car and Changing Cone items (see Table 6). Furthermore, when we explored the dichotomous scale (Correct/Incorrect), we saw that the Ant & Ladybug item was the most difficult, regardless of the model. Theoretically, this made sense because the Ant & Ladybug item was designed to be the most difficult (see Section 3), and our modeling provided quantitative support for this.

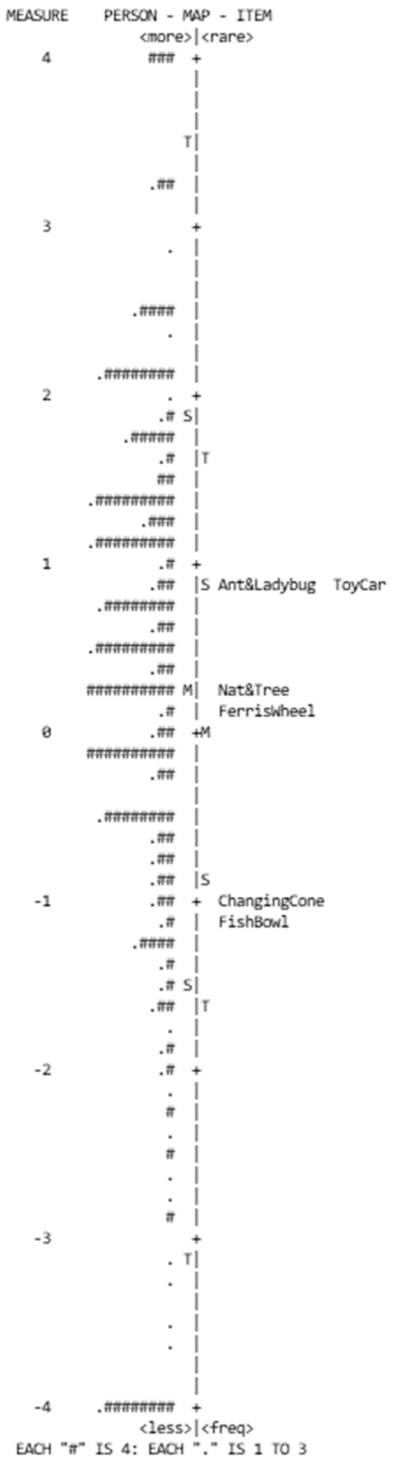


Fig. 3. Graph Reasoning Wright map.

6. Discussion and conclusion

We set out to examine the evaluation and validation stages of the MSGRDS. For the evaluation stage, we demonstrated that the MGSRDS evidences usability and content validity, via analysis of cognitive interviews and expert review (RQ1). For the validation stage, we used Rasch analysis to create models for students' graph selection and reasoning. We found that the MGSRDS evidences

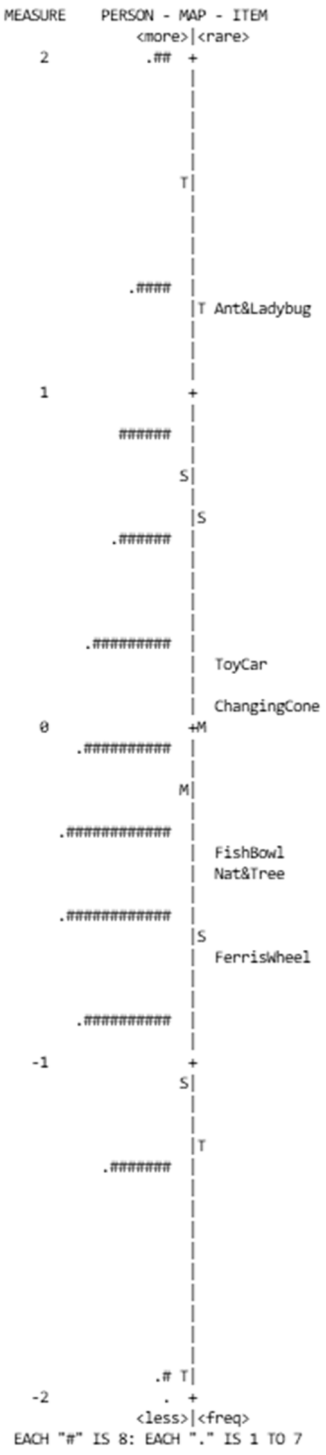


Fig. 4. Graph Selection Wright map.

reliability and internal structure for graph selection and graph reasoning (RQ2), with MGSRDS items providing a continuum for each construct, from most to least difficult (RQ3). Finally, we found empirical evidence to support that the theorized graph reasoning framework (Iconic, Motion, Variation, Covariation) from Johnson et al. (2020) forms a hierarchical scale (RQ4).

Table 7
Self-system process scale step structure.

Category **Graph Selection**	Observed Percentage	Observed Average	Infit MNSQ	Andrich Threshold
0 – Incorrect	44	-0.77	1.02	None
1 – Partially Correct	21	-0.09	0.88	-0.30
2 – Correct	36	0.55	1.02	0.30
Graph Reasoning – original 5 point scale				
0 – Limited Evidence	29	-0.93	1.19	None
1 – Iconic	10	-0.56	0.84	0.02
2 – Motion	25	-0.10	0.98	-1.15
3 – Variation	15	0.45	0.97	0.74
4 – Covariation	21	1.03	0.91	0.40
Graph Reasoning – modified 4 point scale				
1 – Iconic	14	-1.17	1.36	None
2 – Motion	35	-0.38	0.81	-1.95
3 – Variation	21	0.86	0.95	0.85
4 – Covariation	30	1.99	0.87	1.10

Note. Observed percentage is the percent of all responses for that category. Observed average is the average of the measure to produce the responses observed in the category. Infit MNSQ is the average of the infit MNSQs associated with responses in that category. Step Structure is the logit position where the conditional probability of being in either category is equal.

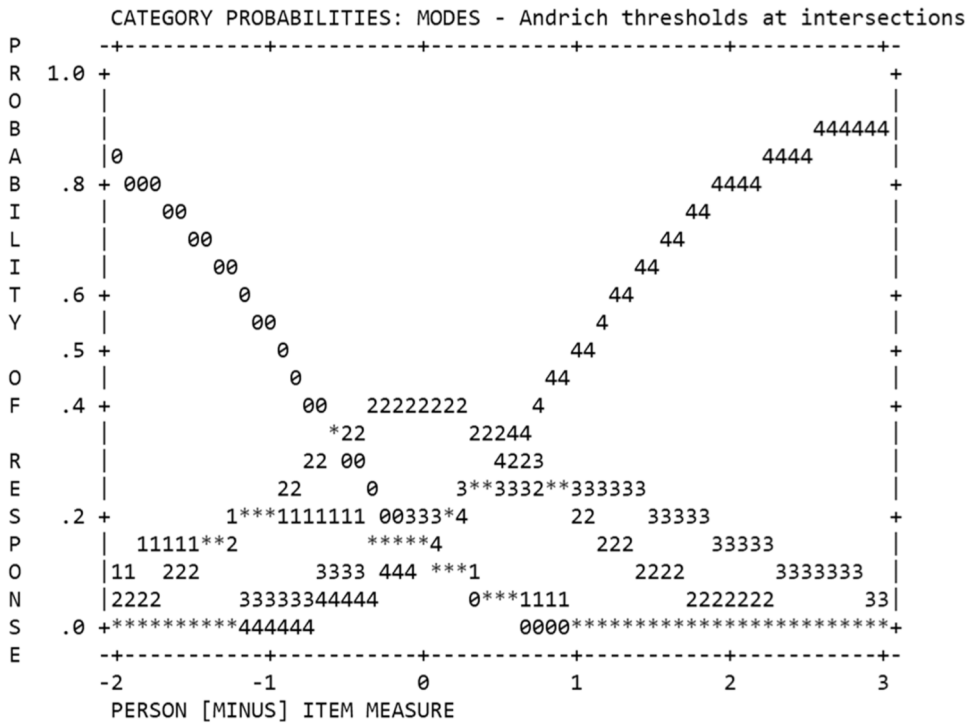


Fig. 5. Graph reasoning category probability curves with 5 point scale.

6.1. Item difficulty for the MGSRDS: Graph reasoning and graph selection

For graph reasoning, the item difficulty scale was: Fishbowl, Changing Cone, Ferris Wheel, Nat & Tree, Ant & Ladybug, Toy Car. This meant the Fishbowl item was the most likely to elicit evidence of students' covariational reasoning, while the Toy Car and Ant & Ladybug items were the least likely to do so (see Fig. 3). In our view, the nature of the movement in each of the animations was one aspect that could account for differences in item difficulty (see also Johnson et al., 2024). Both the Toy Car and the Ant & Ladybug items had an actor, represented by a single point, which "drove" the motion in the animations. As the toy car moved along the track and the insects walked along the paths, the distance attributes accrued with their movement. In contrast, the Fishbowl and the Changing Cone items did not have such an actor. The height and diameter varied as the fishbowl filled with water, and the height and diameter varied as the cone expanded and contracted. The Ferris Wheel and Nat & Tree items also had an actor in the animation, and they fell between the other two pairs of items on the Graph Reasoning Wright map (see Fig. 3).

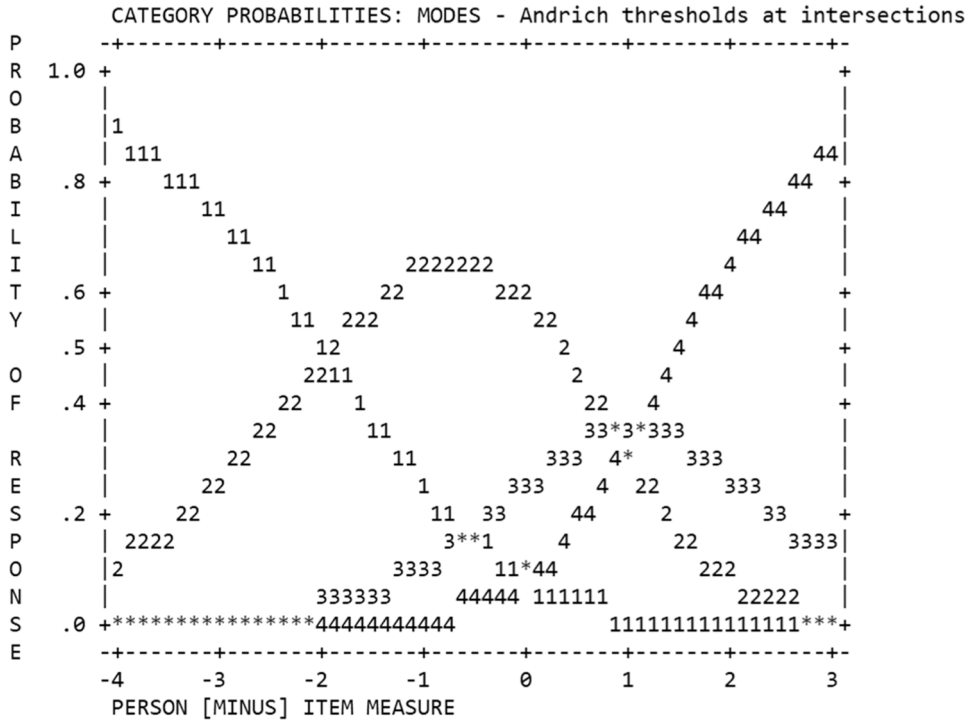


Fig. 6. Graph reasoning category probability curves with 4 point scale.

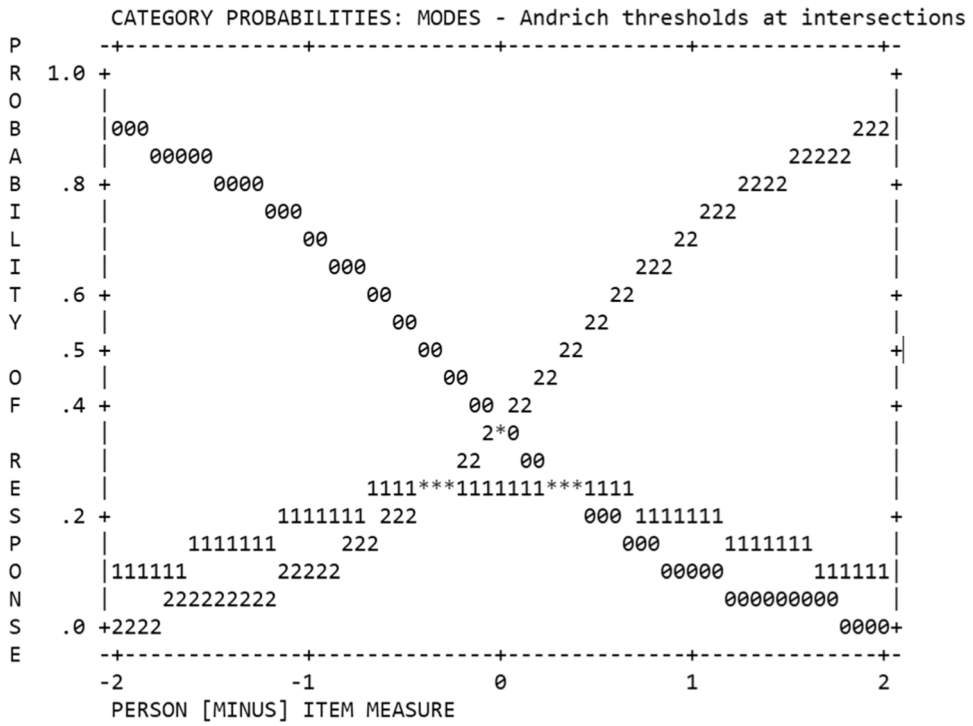


Fig. 7. Graph selection category probability curves with 3 point scale.

For graph selection, the item difficulty scale was: Ferris Wheel, Nat & Tree, Fishbowl, Changing Cone, Toy Car, Ant & Ladybug. This meant the Ferris Wheel item was the easiest for students to select the correct graph, while the Ant & Ladybug item was considerably harder than all others (see Fig. 4). The difficulty of the Ant & Ladybug item aligned with our theorizing; it was the only item in which both attributes varied in their direction of change. The three items that students were most likely to get correct (Ferris Wheel, Nat & Tree, Fishbowl) were also the three items that incorporated unconventional graphs (see Table 1). These graphs were unconventional because multiple values of the variable represented on the vertical axis mapped to the same value of the variable represented on the horizontal axis. Hence, they did not pass the "vertical line test," a standard technique in U.S. classrooms for determining whether a graph represented a function, with the assumption that "a function" meant that the variable on the vertical axis (typically y) was a function of the variable represented on the horizontal axis (typically x). Moore et al. (2014) advocated that breaking conventions could promote students' quantitative reasoning. The positioning of the items on the Graph Selection Wright map (see Fig. 4) indicated that unconventional graphs were viable for including as part of assessment items.

6.2. Limitations of the current study

The graph reasoning scale had a Cronbach's alpha of 0.84 indicating strong reliability, but the graph selection scale had a lower Cronbach's alpha at 0.44. Furthermore, the person separation for graph reasoning (alpha = 0.67 with 1.44 separation) and graph selection (alpha = 0.44 with 0.98 separation) were lower than we expected. However, the graph reasoning items separated students better than the graph selection items. When we tried a dichotomous graph selection scale (Incorrect–0, Correct–1), this lowered person separation even more, which meant that a partial credit category helped to differentiate student ability. In our view, the person separation values for graph selection and graph reasoning were lower due to inconsistent patterns across responses, because it was more likely for students to have a mix of graph selection and graph reasoning codes across their responses (i.e., only 2% of students got all items correct or all incorrect). Although higher reliability would be considered ideal for measurement statistics, in this situation it was not problematic for students to have a mix of codes across graph selection and graph reasoning. In future studies, researchers could examine effects on person separation values by extending to a broader student population (e.g., include secondary students and/or undergraduates at advanced stages of their studies).

Another limitation of our study was our use of students' text responses as a source of data for their graph reasoning. Namely, students might engage in graph reasoning not evidenced in their text responses, and their text responses may capture only a portion of their reasoning. As one might expect, students' verbal responses (n = 31) in the evaluation stage provided a richer source of data than students' text responses (n = 673) in the validation stage. We view this as a tension in the work to scale up smaller-scale studies of students' mathematical reasoning. By including space for students to explain their reasoning and by analyzing those text responses along with their graph selections, we problematize the notion that a particular graph choice will align solely with a particular form of reasoning. Put another way, we do not assume that there are items for which covariational reasoning is the only means by which students could select the correct graph, or in turn, that iconic reasoning is the only means by which students could select the incorrect graph. While students' engagement in covariational reasoning increases the likelihood of their correct graph selection (see Johnson et al., 2024), we recognize that students may select correct graphs for other reasons. By making our assumptions explicit, we work to communicate how we navigate the tensions in scaling up from smaller qualitative studies.

6.3. Contributions to research and practice

This study contributes to the work of mathematics education researchers using Rasch modeling to corroborate theoretical progressions in multiplicative reasoning (Callingham & Siemon, 2021; Kosko, 2019; Tzur et al., 2022), by extending to a new construct, graph reasoning. Our study design and method can offer a blueprint for researchers interested in quantitizing (Sandelowski et al., 2009) qualitative data to form a continuous scale. In future studies, research can test theoretical progressions for other forms of students' mathematical reasoning.

Our Rasch modeling clearly delineates iconic reasoning from motion reasoning along a graph reasoning scale. While Moore and Thompson's (2015) constructs of static and emergent shape thinking distinguish physical and quantitative forms of graph reasoning, respectively, there has been less attention to progressions in physical forms of graph reasoning, such as iconic and motion reasoning. To address this, we also consider Lee et al.'s (2020) distinction between students' spatial and quantitative conceptions of coordinate systems. Unlike iconic reasoning, motion reasoning resembles the "in progress" nature of emergent shape thinking. Yet, students engaging in motion reasoning seem to be operating with a spatial, rather than a quantitative, conception of a coordinate system. This description of motion reasoning shares similarities with one of the theoretical cases put forward by Paoletti et al. (2022b), in which they illustrate how a student may engage in emergent thinking within a spatial coordinate system. When students are engaging in motion reasoning, it is a promising time for teachers and researchers to foster their shifts to quantitative conceptions of coordinate systems. Yet, more needs to be known regarding how such shifts may occur. Extending from our study, researchers may explore conceptual mechanisms by which students advance along the graph reasoning framework (Iconic, Motion, Variation, Covariation) from Johnson et al. (2020). Furthermore, researchers can investigate connections between this graph reasoning framework and related theoretical constructs of static and emergent shape thinking, and spatial and quantitative interpretations of coordinate systems. In our view, a key question to investigate is the learning conditions under which students can shift from physical to quantitative forms of graph reasoning (see also Moore et al., 2019b).

An enduring challenge has been to scale up the results of interview-based studies of students' reasoning, to make claims for larger sample sizes (e.g., Norton & Wilkins, 2009; Tzur et al., 2022). Qualitative analysis of such interview studies is a time-intensive

endeavor. The MGSRDS is a useful assessment for teachers and researchers to use to diagnose students' graph reasoning and selection for dynamic situations. While we have used the MGSRDS to learn about the reasoning of the college algebra student population, researchers can use it to investigate the reasoning of other student populations. Such studies can provide an opportunity to corroborate findings we have reported, including whether the person separation for graph reasoning and graph selection would continue to be lower within and/or across different student populations. Furthermore, our study corroborates the results of a growing corpus of qualitative studies demonstrating that students can engage in covariational reasoning well before enrolling in advanced college math courses (e.g., Ellis et al., 2020; Johnson, 2012; Paoletti et al., 2022a).

In conclusion, we report a novel finding: we provide statistical evidence to support that a theorized graph reasoning framework from Johnson et al. (2020) forms a hierarchical scale. Educators can use this framework (Iconic, Motion, Variation, Covariation) to diagnose students' graph reasoning, in conjunction with the MGSRDS or with other tasks/assessments. This can support educators' knowledge regarding how students are drawing on their own sensibilities to make sense of relationships between attributes in dynamic situations. Future work can include the development of instructional materials to foster students' development in their graph reasoning.

CRediT authorship contribution statement

Bechtold Livvia: Data curation, Methodology. **Donovan Courtney:** Conceptualization, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Johnson Heather Lynn:** Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Knurek Robert:** Visualization, Writing – review & editing. **Whitmore Kristin A.:** Writing – review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Heather Lynn Johnson reports financial support was provided by National Science Foundation. Heather Lynn Johnson is a member of the editorial board for the Journal of Mathematical Behavior.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by the U.S. National Science Foundation, Division of Undergraduate Education [1709903,2013186]. Opinions, findings, and conclusions are those of the authors.

References

Bell, A., & Janvier, C. (1981). The interpretation of graphs representing situations. *For the Learning of Mathematics, 2*(1), 34–42. ⟨https://www.jstor.org/stable/4024074⟩.

Benson, J., & Clark, F. (1982). A guide for instrument development and validation. A guide for instrument development and validation. *The American Journal of Occupational Therapy: Official Publication of the American Occupational Therapy Association, 36*(12), 789–800. https://doi.org/10.5014/ajot.36.12.789

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer. https://doi.org/10.1007/978-94-007-6857-4

Callingham, R., & Siemon, D. (2021). Connecting multiplicative thinking and mathematical reasoning in the middle years. *The Journal of Mathematical Behavior, 61*, Article 100837. https://doi.org/10.1016/j.jmathb.2020.100837

Carlson, M., Jacobs, S., Coe, E., Larsen, S., & Hsu, E. (2002). Applying covariational reasoning while modeling dynamic events: A framework and a study. *Journal for Research in Mathematics Education, 33*(5), 352–378. https://doi.org/10.2307/4149958

Carlson, M., Oehrtman, M., & Engelke, N. (2010). The precalculus concept assessment: A tool for assessing students' reasoning abilities and understandings. *Cognition and Instruction, 28*(2), 113–145. https://doi.org/10.1080/07370001003676587

Chen, X. (2016). Remedial coursetaking at U.S. public 2-and 4-year institutions: Scope, experiences, and outcomes. Statistical analysis report. NCES 2016–405. U.S. Department of Education. Washington, DC: National Center for Education Statistics. https://eric.ed.gov/?id=ED568682.

Clement, J. (1989). The concept of variation and misconceptions in cartesian graphing. *Focus on Learning Problems in Mathematics, 11*(1-2), 77–87. ⟨https://eric.ed.gov/?id=EJ389508⟩.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Sage.

Ellis, A., Ely, R., Singleton, B., & Tasova, H. (2020). Scaling-continuous variation: Supporting students' algebraic reasoning. *Educational Studies in Mathematics, 104*(1), 87–103. https://doi.org/10.1007/s10649-020-09951-6

Ellis, A. B., & Grinstead, P. (2008). Hidden lessons: How a focus on slope-like properties of quadratic functions encouraged unexpected generalizations. *The Journal of Mathematical Behavior, 27*(4), 277–296. https://doi.org/10.1016/j.jmathb.2008.11.002

Glynn, S. M. (2012). International assessment: A Rasch model and teachers' evaluation of TIMSS science achievement items. *Journal of Research in Science Teaching, 49*(10), 1321–1344. https://doi.org/10.1002/tea.21059

Gordon, S. P. (2008). What's wrong with college algebra? *PRIMUS, 18*(6), 516–541. https://doi.org/10.1080/10511970701598752

Johnson, H. L. (2012). Reasoning about variation in the intensity of change in covarying quantities involved in rate of change. *The Journal of Mathematical Behavior, 31*(3), 313–330. https://doi.org/10.1016/j.jmathb.2012.01.001

Johnson, H. L. (2015). Together yet separate: Students' associating amounts of change in quantities involved in rate of change. *Educational Studies in Mathematics, 89*(1), 89–110. https://doi.org/10.1007/s10649-014-9590-y

Johnson, H. L., Donovan, C., Knurek, R., Whitmore, K. A., & Bechtold, L. (2024). Proposing and testing a model relating students' graph selection and graph reasoning for dynamic situations. *Educational Studies in Mathematics*. https://doi.org/10.1007/s10649–024-10299–4

Johnson, H. L., Kalir, J., Olson, G., Gardner, A., Smith, A., & Wang, X. (2018). Networking theories to design a fully online assessment of students' covariational reasoning. In T. E. Hodges, G. J. Roy, & A. M. Tyminski (Eds.), *Proceedings of the 40th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 1343–1346). Greenville, SC: University of South Carolina & Clemson University. ⟨http://www.pmena.org/pmenaproceedings/PMENA%2040%202018%20Proceedings.pdf⟩.

Johnson, H. L., & McClintock, E. (2018). A link between students' discernment of variation in unidirectional change and their use of quantitative variational reasoning. *Educational Studies in Mathematics, 97*(3), 299–316. https://doi.org/10.1007/s10649-017-9799-7

Johnson, H. L., McClintock, E. D., & Gardner, A. (2020). Opportunities for reasoning: Digital task design to promote students' conceptions of graphs as representing relationships between quantities. *Digital Experiences in Mathematics Education, 6*(3), 340–366. https://doi.org/10.1007/s40751-020-00061-9

Johnson, H. L., Olson, G., Smith, A., Gardner, A., Wang, X., & Donovan, C. (2021). Validating an assessment of students' covariational reasoning. In D. Olanoff, K. Johnson, & S. Spitzer (Eds.), *Proceedings of the 43rd Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 63–67). Philadelphia, PA. ⟨http://www.pmena.org/pmenaproceedings/PMENA%2043%202021%20Proceedings.pdf⟩.

Kerslake, D. (1977). The understanding of graphs. *Mathematics in School, 6*(2), 22–25. ⟨https://www.jstor.org/stable/30212405⟩.

Kontorovich, I., Herbert, R., & Yoon, C. (2019). Students resolve a commognitive conflict between colloquial and calculus discourses on steepness. In J. Monaghan, E. Nardi, & T. Dreyfus (Eds.), *Calculus in upper secondary and beginning university mathematics - Conference proceedings* (pp. 119–122). Kristiansand, Norway: MatRIC. ⟨https://www.researchgate.net/publication/335797977⟩.

Kosko, K. W. (2019). A multiplicative reasoning assessment for fourth and fifth grade students. *Studies in Educational Evaluation, 60*, 32–42. https://doi.org/10.1016/j.stueduc.2018.11.003

Lee, H. Y., Hardison, H. L., & Paoletti, T. (2020). Two uses of coordinate systems [1]. *For the Learning of Mathematics, 40*(2), 32–37. ⟨https://www.jstor.org/stable/27091157⟩.

Leech, N. L., Barrett, K. C., & Morgan, G. A. (2015). *IBM SPSS for intermediate statistics: Use and interpretation (5th ed.)*. Routledge,. https://doi.org/10.4324/9781410616739

Leinhardt, G., Zaslavsky, O., & Stein, M. K. (1990). Functions, graphs, and graphing: Tasks, learning, and teaching. *Review of Educational Research, 60*(1), 1–64. https://doi.org/10.3102/00346543060001001

Linacre, J.M. (2023). A user's guide to Winsteps ministep: Rasch-model computer programs (Version 5.4.2). https://www.winsteps.com/a/Winsteps-Manual.pdf.

Mesa, V., Suh, H., Blake, T., & Whittemore, T. (2012). Examples in college algebra textbooks: Opportunities for students' learning. *PRIMUS, 23*(1), 76–105. https://doi.org/10.1080/10511970.2012.667515

Moore, K. C., Paoletti, T., & Musgrave, S. (2013). Covariational reasoning and invariance among coordinate systems. *The Journal of Mathematical Behavior, 32*(3), 461–473. https://doi.org/10.1016/j.jmathb.2013.05.002

Moore, K. C., Silverman, J., Paoletti, T., & LaForest, K. (2014). Breaking conventions to support quantitative reasoning. *Mathematics Teacher Educator, 2*(2), 141–157. https://doi.org/10.5951/mathteaceduc.2.2.0141

Moore, K. C., Silverman, J., Paoletti, T., Liss, D., & Musgrave, S. (2019a). Conventions, habits, and U.S. teachers' meanings for graphs. *The Journal of Mathematical Behavior, 53*, 179–195. https://doi.org/10.1016/j.jmathb.2018.08.002

Moore, K. C., Stevens, I. E., Paoletti, T., Hobson, N. L. F., & Liang, B. (2019b). Pre-service teachers' figurative and operative graphing actions. *The Journal of Mathematical Behavior, 56*, Article 100692. https://doi.org/10.1016/j.jmathb.2019.01.008

Moore, K. C., & Thompson, P. W. (2015). Shape thinking and students' graphing activity. In T. Fukawa-Connelly, N. E. Infante, K. Keene, & M. Zandieh (Eds.), *Proceedings of the 18th Meeting of the MAA Special Interest Group on Research in Undergraduate Mathematics Education* (pp. 782–789). Pittsburgh, PA: RUME. ⟨http://sigmaa.maa.org/rume/RUME18v2.pdf⟩.

Norton, A., & Wilkins, J. L. M. (2009). A quantitative analysis of children's splitting operations and fraction schemes. *The Journal of Mathematical Behavior, 28*(2), 150–161. https://doi.org/10.1016/j.jmathb.2009.06.002

Olson, J., McAllister, C., Grinnell, L., Gehrke Walters, K., & Appunn, F. (2016). Applying constant comparative method with multiple investigators and inter-coder reliability. *The Qualitative Report, 21*(1), 26–42. https://doi.org/10.46743/2160-3715/2016.2447

Paoletti, T., Gantt, A. L., & Vishnubhotla, M. (2022a). Constructing a system of covariational relationships: Two contrasting cases. *Educational Studies in Mathematics, 110*(3), 413–433. https://doi.org/10.1007/s10649-021-10134-0

Paoletti, T., Hardison, H. L., & Lee, H. Y. (2022b). Students' static and emergent graphical shape thinking in spatial and quantitative coordinate systems. *For the Learning of Mathematics, 42*(2), 48–50. ⟨https://eric.ed.gov/?id=ED606704⟩.

Paoletti, T., Stevens, I. E., Hobson, N. L. F., Moore, K. C., & LaForest, K. R. (2018). Inverse function: Pre-service teachers' techniques and meanings. *Educational Studies in Mathematics, 97*(1), 93–109. https://doi.org/10.1007/s10649-017-9787-y

Plass, J. L., Homer, B. D., & Hayward, E. O. (2009). Design factors for educationally effective animations and simulations. *Journal of Computing in Higher Education, 21*(1), 31–61. https://doi.org/10.1007/s12528-009-9011-x

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests (Expanded Edition)*. University of Chicago Press.

Riihiaho, S. (2018). Usability testing. In K. L. Norman, & J. Kirakowski (Eds.), *The Wiley handbook of human computer interaction* (pp. 255–275). John Wiley & Sons. https://doi.org/10.1002/9781118976005.ch14.

Sandelowski, M., Voils, C. I., & Knafl, G. (2009). On quantitizing. *Journal of Mixed Methods Research, 3*(3), 208–222. https://doi.org/10.1177/1558689809334210

Steffe, L. P., & Thompson, P. W. (2000). Teaching experiment methodology: Underlying principles and essential elements. In R. A. Lesh, & A. E. Kelly (Eds.), *Research design in mathematics and science education* (pp. 267–306). Lawrence Erlbaum Associates.

Sullivan, M., III (2020). *Algebra & trigonometry: Enhanced with graphing technology* (8th ed.). Pearson.

Tashakkori, A., & Creswell, J. W. (2007). Editorial: The new era of mixed methods. *Journal of Mixed Methods Research, 1*(1), 3–7. https://doi.org/10.1177/2345678906293042

Tasova, H. I. (2022). Student reasoning in dynamic situations: Spatial proximity reasoning. In A. E. Lischka, E. B. Dyer, R. S. Jones, J. N. Lovett, J. Strayer, & S. Drown (Eds.), *Proceedings of the 44th Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 341–345). Middle Tennessee State University. ⟨http://www.pmena.org/pmenaproceedings/PMENA%2044%202022%20Proceedings.pdf⟩.

Tavakol, M., & Dennick, R. (2011). Making sense of cronbach's alpha. *International Journal of Medical Education, 2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

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. 179–234). State University of New York Press. ⟨https://psycnet.apa.org/record/1994-98042-004⟩.

Thompson, P. W. (2011). Quantitative reasoning and mathematical modeling. In L. L. Hatfield, S. Chamberlain, & S. Belbase (Eds.), *New perspectives and directions for collaborative research in mathematics education. WISDOMe Monographs* (Vol. 1, pp. 33–57). Laramie, WY: University of Wyoming. ⟨https://www.researchgate.net/publication/264119207_Quantitative_reasoning_and_mathematical_modeling⟩.

Thompson, P. W. (2022). Quantitative reasoning as an educational lens. In G. Karagöz Akar, İ.Ö. Zembat, S. Arslan, & P. W. Thompson (Eds.), *Quantitative reasoning in mathematics and science education* (Vol. 21, pp. 1–16). Springer International Publishing. https://doi.org/10.1007/978-3-031-14553-7_1.

Thompson, P. W., & Carlson, M. P. (2017). Variation, covariation, and functions: Foundational ways of thinking mathematically. In J. Cai (Ed.), *Compendium for research in mathematics education* (pp. 421–456). Reston, VA: National Council of Teachers of Mathematics. ⟨https://www.researchgate.net/publication/302581485_Variation_covariation_and_functions_Foundational_ways_of_thinking_mathematically⟩.

Thompson, P. W., Hatfield, N. J., Yoon, H., Joshua, S., & Byerley, C. (2017). Covariational reasoning among U.S. and South Korean secondary mathematics teachers. *The Journal of Mathematical Behavior, 48*(Supplement C), 95–111. https://doi.org/10.1016/j.jmathb.2017.08.001

Tunstall, S. L. (2018). College algebra: Past, present, and future. *PRIMUS, 28*(7), 627–640. https://doi.org/10.1080/10511970.2017.1388315

Tzur, R., Johnson, H. L., Davis, A., Hodkowski, N., Harrington, C., Wei, B., & Norton, A. (2022). A stage-sensitive, written assessment of multiplicative double counting for grades 3-8. *Studies in Educational Evaluation, 74*, Article 101152. https://doi.org/10.1016/j.stueduc.2022.101152

Wright, B. D. (1993). *Rasch Measurement transactions: Logits?, 7*(2), 288. ⟨https://rasch.org/rmt/rmt72e.htm⟩.