Validity evidence based on response processes for an instrument assessing functional thinking*[1,2]

Paulina Araya[3]
Orcid: 0000-0001-6629-6906
Beltrán Pantoja[4]
Orcid: 0009-0004-6407-8859
Andrea Valenzuela[5]
Orcid: 0000-0003-4948-8564

Abstract

Various studies from the perspective of early algebra have employed written tests to assess students' functional thinking. However, there has been little research on the validity of the interpretations derived from these instruments. The purpose of this study was to validate the scores of an instrument designed to assess functional thinking in third-grade students by providing evidence based on response processes. A test and a retrospective cognitive interview were administered to 24 third-grade students from a school in Santiago, Chile. Responses from both the test and the interview were coded and compared to assess the agreement between the interpretations derived from each instrument. The results showed a strong and statistically significant correlation in overall performance and revealed that the item with the greatest divergence between the test and the interview involved describing the relationship between variables using natural language. In conclusion, the instrument offers a reliable means of capturing students' functional thinking, although it may underestimate their performance when expressing functional rules using natural language. Future research could use this instrument while incorporating brief interviews for students in the numerical-functional category, in order to more accurately capture their ability to generalize in natural language without relying on their writing skills.

* English version by Robin Brundell. The authors take full responsibility for the translation of the text, including titles of books/articles and the quotations originally published in Portuguese.

[1] Data availability: The dataset supporting the findings of this study is available in the Harvard Dataverse and can be accessed at https://doi.org/10.7910/DVN/4JALQ7

[2] Acknowledgments: This research was supported by the postdoctoral FONDECYT Grant No. 3220465, funded by the National Agency for Research and Development (ANID) of Chile and Millennium Nucleus for the Study of Early Mathematical Skill Development (MEMAT).

[3] Universidad Diego Portales, Santiago, Chile. Contact: paulina.araya7@mail.udp.cl

[4] Universidad de O'Higgins, Rancagua, Chile. Contact: beltran.pantoja@uoh.cl

[5] Pontificia Universidad Católica de Chile, Santiago, Chile. Contacto: apvalenzuela@uc.cl

https://doi.org/10.1590/S1678-4634202551287163en

Keywords

Assessment – Validity evidence – Early algebraic thinking – Functional thinking – Primary education

Introduction

Currently, various mathematics curricula have incorporated the teaching of algebra as an integral component from the early years of schooling (Kieran, 2022; Blanton *et al.*, 2019; NCTM, 2000). These approaches recommend that students begin by identifying general mathematical relationships and structures through age-appropriate situations, thereby laying the foundation for a progressive understanding of more advanced algebraic concepts, such as variable notation (Blanton *et al.*, 2019; Mason *et al.*, 2009). The early development of algebraic thinking is thus expected to help address low levels of mathematical achievement associated with the delayed introduction of algebra (Kaput, 2008; Moses & Cobb, 2001; Museus *et al.*, 2013).

One key area of algebraic thinking is *functional thinking*, which involves the generalization of covariational relationships, as well as expressing and justifying these relationships through different forms of representation (Brizuela *et al.*, 2015). Numerous studies from this perspective have consistently shown that, even in the early grades, students are capable of analyzing covarying quantities and gradually develop the ability to generalize these relationships using registers of varying complexity, eventually culminating in the expression of their generalizations through conventional symbolic language (Kieran, 2022; Radford, 2018).

In this context, various studies have examined changes in students' functional thinking before and after the implementation of educational interventions (Blanton *et al.*, 2019; Chimoni *et al.*, 2021; Ureña *et al.*, 2019). A central methodological aspect of these studies lies in how student learning is assessed. Two main forms of assessment have been used for this purpose: written instruments (tests) and interviews. Studies employing a qualitative approach have primarily relied on interviews (Blanton *et al.*, 2017; Brizuela *et al.*, 2015; Goñi-Cervera *et al.*, 2022; Ureña *et al.*, 2019), as they offer the advantage of allowing students to express their ideas without relying on their reading and writing skills (Larsson; Granhag, 2005), although they require considerable time for both administration and analysis.

On the other hand, studies that take a quantitative approach to assessing functional thinking (Blanton *et al.*, 2015, 2019; Chimoni *et al.*, 2021; Stylianou *et al.*, 2019) have predominantly relied on paper-and-pencil tests. This type of instrument offers the advantage of assessing a larger number of students in a shorter period of time, which is particularly relevant in studies involving large samples. It is important to note that, in general, the instruments used were developed by the researchers themselves, without a thorough analysis of their validity for measuring the intended construct.

A key aspect of developing assessment instruments is the collection of validity evidence; that is, evidence supporting the interpretations derived from the instrument (AERA; APA; NCME, 2014). One important source of such evidence is known as *response process evidence*, which involves examining the degree of alignment between the assumptions made by the test developers about how respondents will answer and the reasoning processes in which those respondents actually engage (AERA; APA; NCME; 2014; Willson; Miller, 2014). Despite its importance, this source of evidence has been less frequently explored in validation studies (Padilla; Leighton, 2017). Nonetheless, understanding it is essential for ensuring the validity of cognitive assessments.

With the aim of contributing response process validity evidence, this study sought to examine the extent to which a test designed to assess functional thinking captures this type of reasoning in third-grade students. Accordingly, the research questions guiding the study are: 1) To what extent is the functional thinking of third-grade students, as captured by a written test, consistent with that identified through a cognitive interview? and 2) Which items show the greatest discrepancies between the conclusions drawn from the test and those derived from the interview?

Early algebraic thinking

Early algebraic thinking refers to the type of reasoning engaged in by children aged 5 to 12 as they begin to make sense of the objects and ways of thinking they will later encounter in secondary school algebra (Kieran, 2022). Various studies have shown that students in this age range can successfully carry out core algebraic activities such as identifying variables, organizing data in tables, generalizing structures, and expressing generalizations through different forms of representation, including numbers, natural language, and both conventional and non-conventional symbols (Brizuela *et al.*, 2015; Cañadas *et al.*, 2019). According to Radford (2018), algebraic thinking is characterized by the ability to reason analytically about indeterminate quantities (variables or unknowns). In this regard, the use of alphanumeric symbolism is neither necessary nor sufficient for algebraic thinking, as other semiotic systems may also demonstrate that a student understands the structure of a sequence and can reason about it in a general and analytical way (Cañadas *et al.*, 2019; Kieran, 2022; Radford, 2018).

Research in early algebra generally identifies three major strands based on the domains involved: (i) equalities and equations, (ii) generalized arithmetic, and (iii) functional thinking (Blanton *et al.*, 2015). This study focuses on functional thinking, which entails analyzing and generalizing relationships between covarying quantities and expressing these relationships using a range of representations, such as tables, words, or symbols (Blanton; Kaput, 2011). The aim is not to formally introduce the concept of function, but rather to engage students in contextualized tasks involving covarying quantities that can be modeled through functions. These tasks typically require students to generalize a pattern and represent that generalization in multiple ways (Kieran, 2022; Pinto; Cañadas, 2021).

Levels of sophistication in algebraic thinking

Mastery of algebraic language is a gradual process that entails the progressive use of increasingly sophisticated forms of representation to express mathematical generalizations (Blanton *et al.*, 2015; Radford, 2018; Stephens *et al.*, 2017). Various authors have described the types of reasoning students use when analyzing patterns, organizing them according to levels of complexity and sophistication. Radford (2018) identifies three forms of algebraic thinking that emerge at different stages in students' learning trajectories: factual, contextual, and symbolic thinking. Factual thinking involves understanding the operational structure of a sequence and applying it to any term, though without explicitly referencing indeterminate quantities. Contextual thinking allows students to verbalize operational structures in a generalized way, making the indeterminate explicit through natural language. Finally, symbolic thinking reflects the ability to express operational structures using conventional alphanumeric symbols.

Blanton and Kaput (2011) identified three types of reasoning used in pattern analysis, ranging from the simplest to the most complex: (i) recursive thinking, which involves identifying variation within a single sequence; (ii) covariational thinking, which entails analyzing how two quantities change simultaneously, without necessarily identifying a direct relationship; and (iii) correspondence thinking, which involves recognizing a direct correlation between variables. Stephens *et al.* (2017) subsequently expanded Blanton and Kaput's (2011) framework by classifying students' responses based on the type of representation used and the completeness of the generalization. They organized responses by level of difficulty according to the form of representation used—numerical, symbolic, or natural language—and further distinguished whether the response made one or both variables explicit.

It is important to note that the framework proposed by Stephens *et al.* (2017), based on data from written tests, considers representing functional relationships in natural language to be more complex than doing so through symbolic representation. However, findings on whether natural language or symbolic notation is more accessible to students have been mixed (Kieran, 2022). Despite these discrepancies, both models agree that students working with covarying sequences typically begin with recursive and covariational approaches. A significant shift toward algebraic generalization occurs when they use correspondence strategies, initially represented through numerical expressions and progressively through more complex forms such as natural language and symbolic notation. Accordingly, a central issue in assessing functional thinking is determining whether students are first able to identify variables, organize data into tables, and recognize recursive and covariational patterns, and subsequently, whether they recognize correspondence by generalizing the situation using numerical, verbal, and symbolic forms.

Validity evidence based on response processes

Validity is an argument supported by empirical and theoretical evidence that justifies the interpretation of test scores for a specific intended use (AERA; APA; NCME, 2014). The evidence used to assess the validity of test scores can come from multiple sources, including test content, relationships with other variables, and response processes, among others. Together, these sources of evidence form an argument about whether the interpretation of scores is valid for the intended purpose. As such, collecting validity evidence is essential when making decisions about the use of a given test.

One source of validity evidence relates to the response process. This type of evidence aims to support the appropriate alignment between the construct being assessed and the nature of the response provided by the test taker. For example, if a test is intended to assess mathematical reasoning, it is important to ensure that the respondent is in fact applying that reasoning and not relying on a procedure that leads to the correct answer by chance (AERA; APA; NCME, 2014). This kind of evidence is generally obtained through the analysis of individual responses collected via cognitive interviews.

The cognitive interview is a technique that allows access to the mental processes of individuals being assessed as they respond to test items. This type of interview collects additional verbal information about a respondent's answer to a question. The information gathered helps determine whether the question is eliciting the mental process intended by the item designer (Padilla; Benítez, 2014).

The techniques used to conduct cognitive interviews include think-aloud protocols and retrospective designs (Caicedo; Zalazar-Jaime, 2018; Willis, 2019). In the think-aloud technique, participants verbalize their thoughts as they respond to the items. In contrast, retrospective designs involve having participants answer the test items under conditions similar to its intended administration and then take part in a cognitive interview, during which they describe the reasoning they used to formulate their responses. The advantages and disadvantages of each technique have been discussed in the literature (Conrad; Blair, 2009). One advantage of the retrospective interview is that participants complete the test without being influenced by the interviewer or distracted by the need to verbalize their thoughts, which can be an important considering for complex items that require multiple steps (Meadows, 2021).

Methodology

As a way to provide validity evidence based on response processes, this study aimed to analyze the extent to which a test designed to assess functional thinking is able to capture this type of reasoning in third-grade students. To this end, the analysis focused on the degree of alignment between the interpretations derived from the test and the reasoning described by students in a retrospective cognitive interview (Caicedo; Zalazar-Jaime, 2018). Both the students' written and interview responses were coded, and quantitative analyses were applied to assess the level of agreement between these two sources of evidence.

Regarding ethical considerations, informed and voluntary consent to participate in the study was obtained from the students, their parents/guardians, and the school principal. All procedures and consent forms were reviewed and approved by an institutional ethics committee.

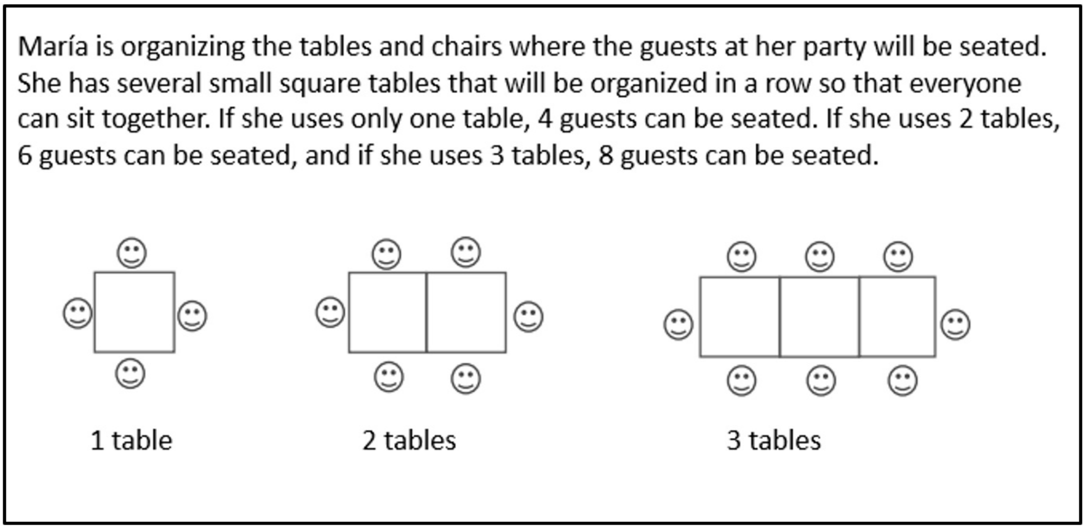
Participants

Twenty-four third-grade students from a public school in Santiago, Chile, participated in the study. These students had previously taken part in a six-week intervention designed to develop functional thinking. Once a week, they worked on a functional situation using various forms of generalization, including variable notation. The tasks in the intervention did not involve the context or function addressed in the test. Participants were selected through purposive non-probability sampling (Castillo; Padilla, 2012), with the objective of including both male and female students with varying performance levels. A sample of this size (N = 24) is considered adequate for collecting validity evidence based on response processes (Padilla; Benítez, 2014; Willis, 2019).

Instruments

The test was developed based on one of the contexts used by Blanton *et al.* (2015) (see Figure 1), from which all test items were derived. Four groups of items were constructed. The first group assessed arithmetic skills, specifically solving near-term cases and organizing the results in a table. The second group included items requiring students to generalize the functional relationship numerically, that is, to calculate the value of a distant term in the sequence. The third group consisted of a single item asking students to describe the relationship between variables using natural language. The fourth group focused on expressing the relationship using symbolic notation. The items, item groups, and targeted skills are detailed in Table 1. The full test is available in Appendix 1 on Dataverse[6].

Figure 1- Functional situation used in the test



María is organizing the tables and chairs where the guests at her party will be seated. She has several small square tables that will be organized in a row so that everyone can sit together. If she uses only one table, 4 guests can be seated. If she uses 2 tables, 6 guests can be seated, and if she uses 3 tables, 8 guests can be seated.

Source: Prepared by the authors.

[6] Annexes available on Dataverse: https://doi.org/10.7910/DVN/4JALQ7

Table 1- Test items by group, targeted skill, and prompts

Group (items on…)	Targeted skill	Item No.	Prompt
Arithmetic thinking	Apply the sequence structure to near terms	1	How many guests can be seated if 4 tables are put together?
		2	How many guests can be seated if 7 tables are put together?
	Identify variables	3A	Organize the number of tables and number of guests in the following chart (2×9 table is provided).
	Organize data in a table identifying the values	3B	
	Identify patterns in the table	4	Do you notice any pattern in the table? Describe it.
Numerical generalization	Generalize the sequence structure by applying it to distant terms	5	How many guests can be seated if 100 tables are put together?
		6	How many guests can be seated if 204 tables are put together?
Verbal generalization	Generalize the sequence structure using natural language	7	Describe how to calculate the number of guests that can be seated for any number of tables.
Generalization using symbolic language	Generalize the sequence structure using symbolic language	8	How would you describe the total number of guests that can be seated at any number of tables using variables (letters)?

Source: Prepared by the authors.

A script with specific questions was developed for each item on the test. The script followed what Willson and Miller (2014) describe as the role of the story teller, in which interviewees are asked to explain why they responded to each item as they did. For example: “In item 4, you wrote that 100 tables would seat 200 guests. Why did you write 200? How did you arrive at that number?” If the student’s answer clearly conveyed a line of reasoning, the interviewer moved on to the next item; otherwise, follow-up prompts were used, such as: “I don’t think I quite understand. Could you explain it again?”

To code the students’ responses for both the test and the interview, a rubric based on theoretical categories was developed. One advantage of this type of rubric is that each performance level is described qualitatively, facilitating the classification of responses into different performance levels (Förster *et al.*, 2017).

In this study, the rubric was applied in two stages. First, responses to each item were coded. For illustration, Table 2 presents part of the rubric showing the codes used for item 7.

In the second stage, the students’ performance was categorized for each item group (see Table 1) based on the codes assigned in the first stage. Each student was assigned a performance level—high, medium, or low—for each group. For example, a student who performed well on items 5 and 6, which comprise the numerical generalization group, was classified as having a high level. The criteria used to assign performance levels for each group are detailed in Appendix 2, available on Dataverse.

Table 2- Rubric codes for item 7: "Describe how to calculate the number of guests that can be seated at any number of tables"

Code	Descriptor
A	Includes the correct operation and correctly names both variables: "the number of tables is multiplied by two and then two is added to get the number of guests."
B	Includes the correct operation and names one variable: "the number of tables is multiplied by two and then two is added."
C	Describes the correct operation: "multiply by two and add two," or "add the same number twice and then add two."
D1	Identifies only the addition and includes at least one variable. Example: "two is added to the number of tables."
D2	Identifies only the multiplication and includes at least one variable. Example: "the number of tables is multiplied by two."
E	Describes an incomplete operation without mentioning variables. Example: "you multiply by two."
F	Provides an incorrect answer but consistent with the student's previous responses to large-number cases. For example, the student previously multiplied by four and now writes "you multiply by four."
G	Uses an irrelevant strategy. Example: "you draw the tables and the people and then count the people."
H	The explanation described by the student is not understandable.
J	Writes a correct example, with or without an accompanying explanation. For instance, "for example, if you gave 70 tables you have to do 70+70+2=142".
NR	No response.

Source: Prepared by the authors.

Administration of instruments

The test was administered by a project researcher one week after the intervention ended. At the start of the session, the researcher carefully read the questions aloud and instructed students to write down all the calculations they performed while completing the test. The administration lasted 60 minutes, during which all students were able to complete the entire test.

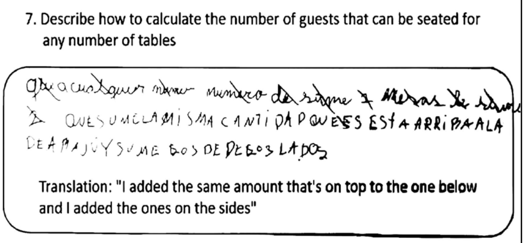
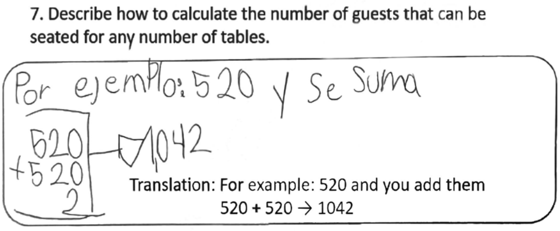
The retrospective interviews were conducted the following day in a room at the school equipped with a video camera. The interviewer called the students one by one and gave them their completed test to help them recall their responses during the interview.

Coding process

The rubric was applied independently by two coders, and their results were subsequently compared. To measure the degree of intercoder agreement, Cohen's Kappa coefficient was calculated for each item. The resulting values ranged from 0.742 to 1, indicating an acceptable level of agreement (Cerda; Villarroel, 2008). Any discrepancies were discussed and resolved through consensus to arrive at a single coding for each response.

To illustrate the coding process, Figure 2 shows the responses of two students to item 7. Student E02's response was coded as "C" (as described in Table 2), while student E29's response was coded as "J".

Figure 2- Responses from two students to item p7

(E02)	(E29)
 7. Describe how to calculate the number of guests that can be seated for any number of tables [illegible] a cualquier ~~[illegible]~~ numero de [illegible] 7 Mesas le [illegible] QUE SUME LA MISMA CANTIDAD QUE ES ESTA ARRIBA A LA DE ABAJO Y SUME DOS DE LOS LADOS Translation: "I added the same amount that's on top to the one below and I added the ones on the sides"	 7. Describe how to calculate the number of guests that can be seated for any number of tables. Por ejemplo: 520 y Se Suma 520 +520 2 1,042 Translation: For example: 520 and you add them 520 + 520 → 1042

Source: Study data.

The interviews were transcribed, and two coders independently assigned each response the rubric code that best reflected the reasoning expressed by the student. Cohen's Kappa coefficient was calculated to assess the level of agreement between coders for each item, with values ranging from 0.646 to 1.0, indicating an acceptable level of agreement. Any discrepancies were discussed and resolved.

To illustrate this coding process, excerpts from interviews with two students, E08 and E16, are presented, focusing on their responses to the question, "How did you arrive at your answer for item 5?" Although both students had written the same incorrect answer—102 guests—on their tests, their interview explanations revealed different lines of reasoning. The excerpts and the analysis are presented below:

> E08: I did make a mistake here.
>
> I: Let's see, tell me about it.
>
> E08: What happened is that I should have added 100 to 100 plus 2, but I got it wrong and wrote 102.
>
> I: Why would you need to add 100 to 100 plus 2?
>
> E08: Because at the top, like here, the same number of guests always appears, and there are always two more. So it's 100 plus 100 plus 2, which is 202.
>
> E16: I realized I couldn't do 100 tables because it would take me until tomorrow. So, I imagined it in my mind and knew the answer was 102, because it goes up by twos... so 100+2.

Analysis of the interview showed that, although E08 made a calculation error, the student correctly identified the underlying structure of the functional situation (code A). In contrast, E16 identified only one of the operations involved (code B1).

Analysis

Based on the coding process, the following variables were defined for each student: performance on each individual test item and interview item, performance across each group of test items, and performance across each group of interview items. To assess the degree of alignment between the codes assigned in the test and those from the interview, the percentage agreement was calculated for each item and each item group.

Subsequently, to examine the relationship between students' responses in the interview and on the test, Kendall's tau correlation coefficient (Kendall, 1938) was applied. This coefficient is appropriate for establishing associations between ordinal variables. Given that the variables in this study reflect ordinal performance categories derived from the assessment, Kendall's tau is particularly well suited for analyzing the associations between them.

To describe overall performance on the test, five categories were established based on the theoretical model employed (see Table 3). All students were classified into one of these categories according to their performance across the item groups, with the exception of two students whose performance did not align with the model. For this reason, a sixth category was added, termed irregular trajectory.

Table 3- Criteria for assigning performance levels based on item group results

Performance categories	Item groups			
	Arithmetic thinking	Numerical generalization	Verbal generalization	Symbolic generalization
Symbolic functional thinking	High or medium	High	High	High
Verbal functional thinking	High or medium	High	High	Medium or low
Numerical functional thinking	High or medium	High	Medium or low	Medium or low
Arithmetic thinking	High or medium	Medium or low	Medium or low	Medium or low
Low performance	Low	Medium or low	Medium or low	Medium or low

Source: Prepared by the authors.

The performance categories, including the irregular trajectory category, were converted into ordinal variables for the purposes of correlation analysis. The categories were ranked from highest to the lowest level of sophistication as follows: symbolic functional thinking, verbal functional thinking, numerical functional thinking, irregular trajectory, arithmetic thinking, and low performance. The irregular trajectory category was placed above arithmetic thinking because all cases within it corresponded to students who demonstrated high performance on the numerical generalization items but showed only medium or low performance on all other item groups.

Results

The following section presents the percentages of agreement between the codes assigned based on the interview data and those assigned based on the test responses for each item (see Table 4). Agreement levels exceeded 90% for items p1, p2, p3a, p3b, p5, and p8, while the lowest level of agreement (58.5%) was observed for item 7.

Table 4- Percentage of agreement between the codes assigned based on the test and the interview for each item

Question No.	1	2	3A	3B	4	5	6	7	8
Agreement	95.8%	95.8%	95.8%	95.8%	87.5%	91.7%	83.3%	58.3%	91.7%

Source: Study data.

For the items assessing arithmetic skills (p1 to p4), there was higher agreement between students' written responses on the test and what they expressed during the interview.

For items p5 and p6 (which involve generalizing the functional situation to distant terms), agreement between the test and interview was 91.7% and 83.3%, respectively. In these cases, students who answered correctly did engage in the expected reasoning, that is, generalizing the structure of the functional situation (double plus two) to specific numbers. The greatest discrepancies between the test and interview coding occurred for item p6. These corresponded to students who arrived at incorrect results through correct procedures but did not include their calculations in their written answers. For example, student E14 stated, "I added 204 to 204 plus 2 and got 500." Since several students made a calculation error when adding 408 + 2, a future version of the test should consider modifying item p6, by replacing the value 204 with 203, for example, to reduce this type of error.

Item p8 also showed a high degree of agreement (91%) between the test and the interview, suggesting that most students wrote algebraic expressions they considered appropriate to represent the observed relationship.

The item with the lowest level of agreement (58.5%) between the written test and the interview was p7. This item asked students to describe the relationship between the variables in their own words. Discrepancies in coding primarily arose because students' responses in the interviews were generally more complete. For example, student E06 wrote "add 2" on the test but said in the interview, "I wanted to explain it, but couldn't find the words, so I just wrote 'add 2'," and then provided an explanation that demonstrated an understanding of the structure of the functional situation. Another example is E01, who wrote the relationship without naming the variables on the test (code C in Table 2), but identified them clearly in the interview and was therefore assigned code A.

Table 5 presents the percentage of agreement between item groups based on the test and interview results. Agreement exceeded 83% in all groups. Once again, the group containing item p7 had the lowest level of agreement. Overall, the grouping of items and the definition of high, medium, and low performance levels made it possible to consolidate different performances whose codes reflected similar levels of achievement, leading to improved agreement percentages. In the previous example, the performance of student E01 was coded as high both on the test (code C) and in the interview (code A).

Table 5- Percentage agreement between codes assigned from the test and the interview for each item group

Item groups	Arithmetic items	Numerical generalization items	Natural language generalization items	Symbolic generalization items
Agreement	87.5%	91.7%	83.3%	91.7%

Source: Study data.

Table 6 presents the results of the correlation analysis between the test and the interview, both for each item group and for the overall performance category on the test. A strong and statistically significant correlation was found across the item groups, indicating that, despite the discrepancies previously described, the performance levels assigned to each group tended to follow the same pattern. Moreover, the overall performance category showed a high and significant correlation (τ_b=0.913; p<0.001, N=24), suggesting that the conclusions drawn from the test regarding students' functional thinking are largely consistent with those derived from the interviews.

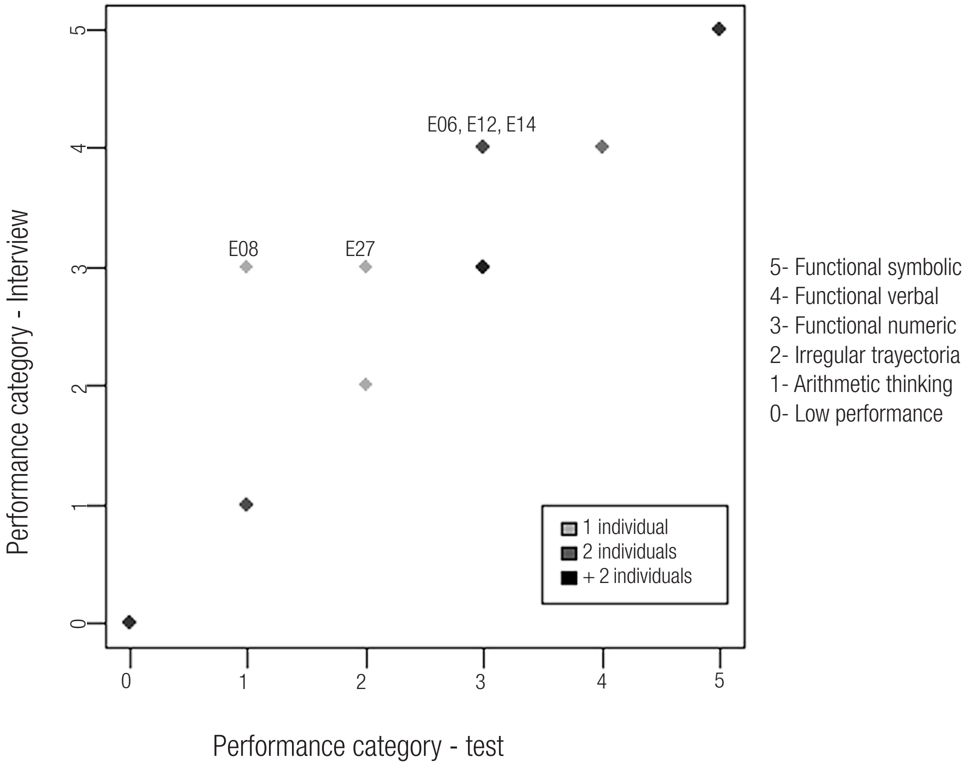
Table 6- Correlation between test and interview by item group and performance category

	Item group (items on…)				Performance category
	Arithmetic thinking	Numerical generalization	Generalization in natural language	Generalization in symbolic language	
Correlation	0.891***	0.871***	0.868***	0.920***	0.913***

Note: Kendall's tau was used to estimate correlations.
***p<0.001.
Source: Study data.

Finally, the scatterplot in Figure 3 illustrates the correlation between the overall performance categories derived from the test and the interview. Most cases (19 out of 24) fall along the diagonal, indicating agreement between the two instruments. In cases where there was no agreement, the interview placed students in a higher performance category than the test, suggesting that the test tended to underestimate the performance of these five students.

Figure 3- Scatterplot of performance categories assigned based on the test and the interview



Source: Study data.

Regarding the cases that do not fall along the diagonal, three students (E06, E12, and E14) are concentrated at the same point. In all three cases, the students were initially categorized as functional numeric based on the test, but the interview revealed that they belonged in the functional verbal category. These students were unable to respond to item p7 on the written test but answered it correctly during the interview. The remaining two cases arose from less common circumstances. Student E08 was moved from the arithmetic thinking category to functional numeric, as their incorrect responses to the numerical generalization items on the test were shown in the interview to result solely from calculation errors. Meanwhile, E27 had been classified under irregular trajectory on the test for not responding to items p1 and p2 and only producing drawings. However, in the interview, they demonstrated knowledge of the results and were therefore reclassified as functional numeric.

Discussion

Tests designed to assess functional thinking are often based on the assumption that students, when responding to various items, will engage in reasoning processes such as identifying patterns and generalizing relationships between variables. Examining the extent to which students actually employ these forms of reasoning constitutes a crucial type of evidence. This study aimed to contribute evidence of validity based on response processes (AERA; APA; NCME, 2014) by analyzing the extent to which interpretations derived from a test designed to assess the functional thinking of third-grade students were consistent with the reasoning expressed by those students during cognitive interviews. The agreement analysis between the test and the interview helped identify the items with the greatest discrepancies. Overall, a high and statistically significant correlation was found (, ; N=24) between the level of functional thinking assigned based on the test and that identified during the interview, supporting the conclusion that the test effectively elicits the mental processes it is intended to measure.

In line with other scholars such as Radford (2018), Stephens *et al.* (2017), and Blanton et al. (2015), the instrument developed in this study enabled the identification of a discernible trajectory in the development of students' functional thinking. Along this trajectory, performance categories can be distinguished according to varying levels of sophistication. The arithmetic thinking category included students who were only able to work with near terms and record some data in a table. The numerical functional category comprised students who managed to establish relationships between variables and applied them to distant terms, although they did not formulate general rules. In the verbal functional category, students generalized not only through numerical representations but also using indeterminate expressions in spoken or written natural language. The symbolic functional category identified students who generalized using all representational registers, including variable notation.

Unlike Stephens *et al.* (2017), the results of this study did not reveal instances of students using symbolic representations without also demonstrating the ability to express the relationship in words, whether written or spoken. That is, students who successfully generalized using variable notation were also able to describe the relationship in natural language, even if some experienced difficulties articulating their ideas in writing. In line with authors such as Radford (2018) and Pinto and Cañadas (2021), our findings suggest that students tend to express relationships between variables in natural language before acquiring the ability to use symbolic representations.

A detailed, item-by-item analysis of the alignment between the written test and the interview revealed that the most significant discrepancies occurred in item p7, in which students were asked to describe the relationship between the variables in their own words. It was found that only 58% of students conveyed in the written test what they actually understood, while the remaining 42% provided more complete responses during the interview. This suggests that formulating a generalized functional rule in writing may pose a challenge for many students. These findings are consistent with studies that highlight the difficulties that younger students face in expressing mathematical ideas in written form (Blackstock-Bernstein *et al.*, 2022; Hughes *et al.*, 2020).

Considering this aspect may be important when drawing conclusions based on this type of item. For example, in the study by Stephens *et al.* (2017), which used written instruments, students were reported to find it easier to express generalizations using variable notation than in words. The findings of our study suggest that the difficulty in generalizing using natural language may be more closely linked to the written format than to the use of natural language itself.

Lastly, the analysis of the overall performance category assignments showed that for 79% of the students (19 out of 24), the test and interview yielded the same classification. In the remaining 21% of cases, the test underestimated the level assigned based on the interview. These discrepancies were primarily observed among students categorized as functional numeric. For these students, the interview provided a more accurate understanding of their ability to generalize functional rules using natural language. These results suggest that the application of the instrument could be improved by incorporating brief follow-up interviews focused on item p7 and on students placed in the functional numeric category. Such a methodological adjustment could enhance the assessment of functional thinking without relying too heavily on students' writing skills.

Conclusions

This study provides validity evidence based on response processes that supports the interpretations derived from a written test designed to assess functional thinking in third-grade students (ages 9-10). Overall, the instrument under analysis—comprising the test and an accompanying rubric—proved effective in reliably capturing the students' functional thinking. One noteworthy finding was that the item showing the greatest discrepancy between the test and the interview involved articulating the functional relationship using natural language. This suggests that future research could consider incorporating complementary tools, such as brief interviews, to examine whether students classified in the functional numeric thinking category are able to express their generalizations in natural language, regardless of their writing skills.

While these findings are significant, it is important to note that the evidence was obtained from a primary school grade in a public school serving a low socioeconomic context. To assess its applicability in other settings, additional validity analyses would be needed with students from diverse socioeconomic backgrounds. Such an expansion would not only help to explore the generalizability of the findings, but also to determine whether difficulties with written language are present in other sociodemographic groups. Furthermore, it should be noted that response processes represent only one of several possible sources of validity evidence; therefore, future studies could investigate other sources to strengthen the validity of the interpretations derived from the test.

References

AERA; APA; NCME. **Standards for Educational & Psychological Testing**. Washington, DC: American Educational Research Association, 2014.

BLACKSTOCK-BERNSTEIN, Anne; WOODBRIDGE, Amy y BAILEY, Alison. Comparing elementary students' explanatory language across oral and written modes. **The Elementary School Journal,** Chicago, v.122, n. 3, p. 315-340, 2022. https://doi.org/10.1086/718077

BLANTON, Maria; BRIZUELA, Barbara; GARDINER, Angela Murphy; SAWREY, Katie; NEWMAN-OWENS, Ashley. Progression in first-grade children's thinking about variable and variable notation in functional relationships. **Educational Studies in Mathematics**, v. 95, p. 181-202, 2017. https://doi.org/10.1007/s10649-016-9745-0

BLANTON, Maria; KAPUT, James. Functional thinking as a route into algebra in the elementary grades. *In*: CAI, Jinfa; KNUTH, Eric (ed.). **Early algebraization**. Berlin: Springer, 2011. p. 5-23.

BLANTON, Maria; STEPHENS, Ana; KNUTH, Eric; GARDINER, Angela Murphy; ISLER, Isil; KIM, Jee-Seon. The development of children's algebraic thinking: The impact of a comprehensive early algebra intervention in third grade. **Journal for Research in Mathematics Education**, Reston, v. 46, n. 1, p. 39-87, 2015. https://doi.org/10.5951/jresematheduc.46.1.0039

BLANTON, Maria; STROUD, Rena; STEPHENS, Ana; GARDINER, Angela Murphy; STYLIANOU, Despina; KNUTH, Eric; ISLER-BAYKAL, Isil; STRACHOTA, Susanne. Does early algebra matter? The effectiveness of an early algebra intervention in grades 3 to 5. **American Educational Research Journal**, Washington, DC., v. 56, n. 5, p. 1930-1972, 2019. https://doi.org/10.3102/0002831219832301

BRIZUELA, Barbara; BLANTON, Maria; SAWREY, Katharine; NEWMAN-OWENS, Ashley; GARDINER, Angela Murphy. Children's use of variables and variable notation to represent their algebraic ideas. **Mathematical Thinking and Learning**, Cham, v. 17, n. 1, p. 34-63, 2015. https://doi.org/10.1080/10986065.2015.981939

CAICEDO, Estefanía; ZALAZAR-JAIME, Mauricio Federico. Entrevistas cognitivas: revisión, directrices de uso y aplicación en investigaciones psicológicas. **Avaliação Psicológica**, São Paulo, v. 17, n. 3, p. 362-370, 2018. https://doi.org/10.15689/ap.2018.1703.14883.09

CAÑADAS, María; BLANTON, Maria; BRIZUELA, Barbara. Special issue on early algebraic thinking. **Journal for the Study of Education and Development**, Madrid, v. 42, n. 3, p. 469-478, 2019. https://doi.org/10.1080/02103702.2019.1638569

CASTILLO, Miguel; PADILLA, José Luís. How cognitive interviewing can provide validity evidence of the response processes to scale items. **Social Indicators Research**, Dordrecht, v. 114, p. 963-975, 2012. https://doi.org/10.1007/s11205-012-0184-8.

CERDA L, Jaime; VILLARROEL DEL P, Luis. Evaluación de la concordancia inter-observador en investigación pediátrica: Coeficiente de Kappa. **Revista Chilena de Pediatría**, Santiago de Chile, v. 79, n. 1, p. 54-58, 2008.

CHIMONI, María; PITTA-PANTAZI, Demetra; CHRISTOU, Constantinos. The impact of two different types of instructional tasks on students' development of early algebraic thinking (El impacto de dos tipos diferentes de tareas instruccionales en el desarrollo del pensamiento algebraico temprano de los estudiantes). **Journal for the Study of Education and Development**, Madrid, v. 44, n. 3, p. 503-552, 2021. https://doi.org/10.1080/02103702.2020.1778280

CONRAD, Frederick G.; BLAIR, Johnny. Sources of error in cognitive interviews. **The Public Opinion Quarterly**, Oxford, v. 73, n. 1, p. 32-55, 2009. https://doi.org/10.1093/poq/nfp013

FÖRSTER, Carla; ZEPEDA, Sandra; NUÑEZ, Claudio. Instrumentos para la evaluación de aprendizajes, ¿con qué evaluar? *In*: FÖRSTER, Carla E. (ed.). **El poder de la evaluación en el aula**: mejores decisiones para promover aprendizaje. Santiago de Chile: Universidad Católica de Chile, 2017. p. 177-230.

GOÑI-CERVERA, Juncal; CAÑADAS, María; POLO-BLANCO, Irene. Generalizations in students with autism spectrum disorder: an exploratory study of strategies. **ZDM–Mathematics Education**, Berlin, v. 54, n. 6, p. 1333-1347, 2022. https://doi.org/10.1007/s11858-022-01415-w

HUGHES, Elizabeth; RICCOMINI, Paul; LEE, Joo-Young. Investigating written expressions of mathematical reasoning for students with learning disabilities. **The Journal of Mathematical Behavior**, v. 58, n. 2, 2020. https://doi.org/10.1016/j.jmathb.2020.100775

KAPUT, James. What is algebra? What is algebraic reasoning? *In*: KAPUT, James; CARRAHER, Davis; BLANTON, Maria (ed.). **Algebra in the early grades**. New York: Taylor and Francis, 2008. p. 5-17.

KENDALL, Maurice. A new measure of rank correlation. **Biometrika**, Oxford, v. 30, n 1/2, p. 81-93, 1938. https://doi.org/10.2307/2332226

KIERAN, Carolyn. The multi-dimensionality of early algebraic thinking: background, overarching dimensions, and new directions. **ZDM Mathematics Education**, Berlin, v. 54, n. 6, p. 1131-1150, 2022. https://doi.org/10.1007/s11858-022-01435-6

LARSSON, Anneli; GRANHAG, Pär Anders. Interviewing children with the cognitive interview: assessing the reliability of statements based on observed and imagined events. **Scandinavian journal of psychology**, Stockholm, v. 46, n. 1, p. 49-57. 2005. https://doi.org/10.1111/j.1467-9450.2005.00434.x

MASON, John; STEPHENS, Max; WATSON, Anne. Appreciating mathematical structure for all. **Mathematics Education Research Journal**, Sydney, v. 21, n. 2, p. 10-32, 2009. https://doi.org/10.1007/BF03217543

MEADOWS, Keith. Cognitive Interviewing Methodologies. **Clinical Nursing Research**, Thousand Oaks, v. 30, n. 4, p.375-379. 2021. https://doi.org/10.1177/10547738211014099

MOSES, Robert y COBB, Charles. **Radical equations**: math literacy and civil rights. Boston: Beacon Press. 2001.

MUSEUS, Samuel; PALMER, Robert; DAVIS, Ryan; MARAMBA, Dina. Special issue: racial and ethnic minority students' success in STEM education. **ASHE Higher Education Report,** Hoboken, v. 36, n. 6, p. 1-140. 2013. https://doi.org/10.1353/rhe.2013.0046

NCTM. National Council of Teachers of Mathematics. **Principles and standards for school mathematics**. Reston: NCTM. 2000.

PADILLA, José Luís; BENÍTEZ, Isabel. Validity evidence based on response processes. **Psicothema**, Oviedo, v. 26, n. 1, p. 136–144, 2014. https://doi.org/10.7334/psicothema2013.259

PADILLA, José Luís; LEIGHTON, Jacqueline. Cognitive interviewing and think aloud methods. *In*: ZUMBO, Bruno; HUBLEY, Anita (ed.). **Understanding and investigating response processes in validation research**. [*S. l.*]: Springer International/Springer Nature, 2017. p. 211-228. https://doi.org/10.1007/978-3-319-56129-5_12

PINTO, Eder; CAÑADAS, María. Generalizations of third and fifth graders within a functional approach to early algebra. **Mathematics Education Research Journal**, Sydney, v. 33, n. 1, p. 113-134. 2021. https://doi.org/10.1007/s13394-019-00300-2

RADFORD, Luis. The emergence of symbolic algebraic thinking in primary school. *In*: KIERAN, Carolyn (ed.). **Teaching and learning algebraic thinking with 5- to 12-year-olds**. Cham: Springer, 2018. p. 3-25. ICME-13 Monographs. https://doi.org/10.1007/978-3-319-68351-5_1

STEPHENS, Ana; FONGER, Nicole; STRACHOTA, Susanne; ISLER, Isil; BLANTON, Maria; KNUTH, Eric; GARDINER, Angela. A learning progression for elementary students' functional thinking. **Mathematical Thinking and Learning**, Cham, v. 19. n. 3, p. 143-166, 2017. https://doi.org/10.1080/10986065.2017.1328636

STYLIANOU, Despina; STROUD, Rena; CASSIDY, Michael; KNUTH, Eric; STEPHENS, Ana; GARDINER, Angela; DEMERS, Lindsay. Putting early algebra in the hands of elementary school teachers: examining fidelity of implementation and its relation to student performance. El álgebra temprana en manos del docente de primaria: un análisis de la fidelidad de ejecución y su relación con el rendimiento de los escolares. **Journal for the Study of Education and Development**, Madrid, v. 42, n. 3, p. 523-569, 2019. https://doi.org/10.1080/02103702.2019.1604021

UREÑA, Jason; RAMÍREZ, Rafael; MOLINA, Marta. Representaciones de la generalización de una relación funcional y el vínculo con la mediación del entrevistador. **Journal for the Study of Education and Development**, Madrid, v. 42, n. 3, p. 591-614, 2019. https://doi.org/10.1080/02103702.2019.1604020

WILLIS, Gordon. **Cognitive interviewing**: A tool for improving questionnaire design. Thousand Oaks: Sage, 2019.

WILLSON, Stephanie; MILLER, Kristen. Data collection. *In*: MILLER, Kristen; WILLSON, Stephanie; CHEPP, Valeria; PADILLA, José-Luís (ed.). **Cognitive interviewing methodology**. Hoboken: Wiley, 2014. p. 15-33.

Received on May 30, 2024
Reviewed on October 08, 2024
Approved on November 25, 2024

Responsible editor: Prof. Tatiana Hochgreb-Haegele

Paulina Araya is an assistant professor at the Faculty of Education, Universidad Diego Portales. She holds a Bachelor's degree in Exact Sciences and is a certified mathematics and physics teacher from Universidad de Chile, and earned her doctorate in education from Universidad Diego Portales and Universidad Alberto Hurtado.

Beltrán Pantoja is a mathematics professor at Universidad Alberto Hurtado and holds a master's degree in education with a specialization in learning assessment from Pontificia Universidad Católica de Chile.

Andrea Valenzuela is an adjunct professor at the Faculty of Education, Pontificia Universidad Católica de Chile. She is an early childhood educator and holds a master's degree in education with a specialization in learning assessment from the same university.