Revista Iberoamericana para la Investigación y el Desarrollo Educativo
ISSN 2007 - 7467

https://doi.org/10.23913/ride.v14i28.1932

Scientific articles

Juicio evaluativo de docentes chilenos sobre la dificultad de los ítems PISA de matemáticas y el desempeño de los estudiantes

Evaluative judgment that Chilean teachers make about the difficulty of the PISA mathematics items and students' achievement

Julgamento avaliativo de professores chilenos sobre a dificuldade dos itens de matemática do PISA e o desempenho dos alunos

Verónica Villarroel Henríquez
Universidad San Sebastián, Chile
veronica.villarroel@uss.cl
https://orcid.org/0000-0002-3000-2248

María Angélica San Martín Toro
Universidad del Desarrollo, Chile
angelicasanmartin@gmail.com
https://orcid.org/0000-0003-4715-5782

Carmen Núñez Ramos
Universidad del Desarrollo, Chile
carmennunezramos@gmail.com
https://orcid.org/0000-0001-6223-4660

Carolina Hernández Opazo
Universidad del Desarrollo, Chile
carolinahernandezopazo@gmail.com
https://orcid.org/0000-0001-8640-3334

Isidora Castillo Rabanal
University College London, Reino Unido
isidora.rabanal.21@ucl.ac.uk
https://orcid.org/0000-0002-9941-1437

Revista Iberoamericana para la Investigación y el Desarrollo Educativo
ISSN 2007 - 7467

Alejandro Sánchez Oñate
Universidad del Desarrollo, Chile
alejandro.sanchez@udd.cl
https://orcid.org/0000-0003-0990-6004

Resumen

El objetivo de este estudio fue analizar el juicio evaluativo de los docentes de matemáticas sobre la dificultad de una muestra de ítems en los que los estudiantes mostraron alto y bajo desempeño en la Prueba PISA 2015 en Chile. Para ello, se elaboró una investigación con enfoque cuantitativo y un diseño exploratorio. Los participantes fueron 18 docentes de matemáticas, elegidos mediante un muestreo no probabilístico e intencional, que respondieron a una encuesta con preguntas de respuesta abierta y cerrada, la cual evaluaba cada ítem con base en cinco dimensiones: formulación del ítem, contenido, contextualización, habilidad y complejidad. Los resultados demuestran que el 40% de los docentes de matemáticas presentan dificultad para acertar con precisión respecto a los ítems en los que los estudiantes mostraron bajo desempeño. Sin embargo, tendieron a concordar en el 60% de las evaluaciones y a predecir en el 80% el desempeño de los alumnos en los ítems de alto desempeño. Según los docentes, los estudiantes tendrán más dificultades en ítems que miden una habilidad cognitiva superior, tienen baja familiaridad con el tipo de ítem y en los que la información se encuentra de forma implícita en el problema.

Palabras clave: competencias matemáticas, juicio evaluativo docente, prueba PISA, enseñanza media, desempeño académico, habilidades cognitivas.

Abstract

The aim of this study was to analyse the evaluative judgment of mathematics teachers regarding the difficulty of a sample of items on which students exhibited high and low performance in the PISA 2015 test in Chile. Through a quantitative research approach and an exploratory design, non-probabilistic and intentional sampling, 18 mathematics teachers participated responding to a survey with open and closed-ended questions, evaluating each item in relation to five dimensions (item formulation, content, contextualization, skill, and complexity). It was found that 40% of mathematics teachers have difficulty accurately pinpointing items where students demonstrated low performance. However, they tended to agree in 60% of the assessments and predict students' performance in high-performance items by 80%. According to teachers, students will face more

Revista Iberoamericana para la Investigación y el Desarrollo Educativo
ISSN 2007 - 7467

challenges in items measuring higher cognitive abilities, with low familiarity with the item type, and when information is implicitly presented in the problem.

Key words: Mathematical competencies, Teachers' evaluative judgement, PISA test, Highschool, Academic achievement, Cognitive abilities.

Resumo

O objetivo deste estudo foi analisar o julgamento avaliativo de professores de matemática sobre a dificuldade de uma amostra de itens em que os alunos apresentaram alto e baixo desempenho no teste PISA 2015 no Chile. Para isso, foi realizada uma pesquisa com abordagem quantitativa e delineamento exploratório. Os participantes foram 18 professores de matemática, escolhidos por meio de amostragem não probabilística e intencional, que responderam a uma pesquisa com questões de resposta aberta e fechada, que avaliou cada item com base em cinco dimensões: formulação do item, conteúdo, contextualização, habilidade e complexidade. Os resultados mostram que 40% dos professores de matemática têm dificuldade em responder com precisão os itens em que os alunos apresentaram desempenho ruim. Contudo, tenderam a concordar em 60% das avaliações e a prever o desempenho dos alunos em itens de alto desempenho em 80%. Segundo os professores, os alunos terão mais dificuldades em itens que medem maior capacidade cognitiva, têm baixa familiaridade com o tipo de item e em que a informação se encontra implicitamente no problema.

Palavras-chave: competências matemáticas, julgamento avaliativo de professores, teste PISA, ensino secundário, desempenho acadêmico, habilidades cognitivas.

Reception Date: November 2023 **Acceptance Date:** May 2024

Introduction

The development of mathematical skills is a relevant factor not only for the academic success of students (Claessens & Engel, 2013), but also for the intellectual development of children and young people, since it helps them think logically, reason in an orderly manner. and having a mind prepared for problem solving, generalization and abstraction (Benson-O'Connor *et al.*, 2019). In addition, it allows them to develop skills related to accuracy in results, understanding and use of symbols (Villegas-Zamora, 2019), as well as promoting cooperation skills, taking turns and self-regulation, even in the youngest students (Stipek *et al.*, 2012).

In short, regardless of the situation in the school environment, mathematical learning helps people to understand and participate in the world in an active and critical way and then make decisions in the face of events that require reasoning and numerical operations (Angier, 2019). In fact, mathematics is an essential tool for making decisions in the adult world (Menoyo Díaz, 2020; Trejo, 2020), which is put to the test, for example, when saving, investing, deciding on the pension system, understand what an interest rate entails or apply procedures to determine quantities when cooking or taking medicine, etc.

For these reasons, it can be assured that examining how students are learning mathematics and how familiar they are with applying it to solve everyday problems is also an indicator of their citizenship education (Andrade & Guzmán, 2018). To this end, national and international standardized tests can provide information of great interest that allows us to determine which skills should be enhanced, as well as which teaching methodologies and pedagogical strategies can be improved.

Based on this purpose, Canales and Maldonado (2018) studied the results of the 2011 eighth grade Education Quality Measurement System (SIMCE) and its complementary survey. These authors found that teachers contribute significantly to students' mathematics and language outcomes, especially those with more years of experience. Furthermore, they observed that teachers with less experience and pedagogical skills tend to be in charge of disadvantaged students, who are in a vulnerable socioeconomic situation, which further limits the possibilities for improvement of these students.

In accordance with this idea, Torres (2018) points out that although there are effective mathematics and language teachers in educational establishments with a low socioeconomic level, greater variability is observed in their effects on student performance. In fact, this variability is lower in establishments with a higher socioeconomic level, where there are fewer teachers classified as ineffective.

In an attempt to understand the performance of Chilean students in the Program for International Student Assessment (PISA), Villarroel *et al.* (2015) investigated the students' most frequent successes and errors in the different items of the 2009 test. The authors demonstrated that some factors that influenced student performance were the types of items, the students' familiarity with them and the complexity of the cognitive skills necessary to answer the questions. In addition, they observed that students make more errors in questions that measure highly complex cognitive skills, such as establishing relationships between data and procedures in mathematics.

Along the same lines, Valenzuela *et al.* (2015) analysed the systemic and individual variables that influence the improvement in student results in the PISA test in the period 2000-2009, and found that student attitudes explained 25% of the improvement in results. In turn, the learning strategies they used were not configured as significant in the explanation.

At an international level, in Turkey, Aydın and Özgeldi (2017) delved into the difficulties of mathematics pedagogy students in solving the items of the PISA 2012 test. To do this, they administered a 26-item exam to 52 future teachers, followed by 12 interviews, within the framework of a mixed strategy. The researchers reported that the teachers in training presented consistent difficulties in the items that combined conceptual, contextual, and applied knowledge of mathematics, since they offered limited conceptual explanations for the items and a fragmented contextualization.

For their part, Radišić and Baucal (2018) explored how teachers perceive Serbian students' thinking in relation to the mathematical content of two PISA 2012 items. They found that item familiarity is a facilitator of correct resolution, while decontextualization of the content would affect poor performance. However, the teachers agreed that the students only needed basic primary knowledge to answer the test items, although many were not able to specify the procedure that the students required to solve it. In this sense, the disagreement between the teachers' judgments to identify the element of the item that makes it more difficult was notable, since attributions were found about the complexity of the measured skill and the complexity of the instruction, or the construction of the question. Finally, the authors concluded that the inability of teachers to imagine the difficulties that students may present with an item can be linked to a weak cultivation of the teacher-student relationship and poor training in pedagogical skills.

These investigations focused on understanding the level of learning achieved by students in mathematics shed light on the educational policies of countries. In particular, the PISA test, developed by the Organization for Economic Cooperation and Development, measures the level of preparation of 15-year-old students to face the challenges they may encounter in the future (OECD, 2010, 2019).

The items of this instrument are based on situations and contexts close to the lives of students, who must face the challenge of solving problems using the knowledge learned (Caño & Luna, 2011). The level of complexity of these items can be classified into three categories: a) basic, where basic mathematical skills are evaluated, such as solving simple problems and interpreting numerical information, b) moderate, which addresses more complex situations that they require the application of mathematical concepts in real-world situations, and c) higher, which involves the

resolution of highly abstract and theoretical problems, which require a high degree of mathematical reasoning and the ability to address novel situations. These levels allow us to measure not only the understanding of mathematical concepts, but also the ability of students to apply their knowledge in diverse and challenging contexts (Villarroel *et al.*, 2015).

Now, in the specific case of Chile, the country has participated in the PISA evaluations since 2000, and although the results show sustained improvement, there is a significant percentage of students who do not achieve the minimum performance of skills necessary for life. The results, in fact, are below the average of OECD countries, although significantly above the average of other Latin American countries (OECD, 2019; Valenzuela *et al.*, 2009).

Given this scenario, the following questions were raised: are classroom teachers competent to predict the items in which their students will perform better and worse? What explanations do teachers offer regarding the items in which Chilean students have lower performance?

The literature review highlights the importance of understanding how teachers perceive the characteristics and difficulties of this type of evaluations, to understand the origin of student performance and have a direct impact on the design and implementation of innovative teaching and evaluation practices supported in specialized literature. This interest arises due to how significant the planning competence of the teaching and learning process is (Morales Salas, 2018), hence this research focuses on the perspective of mathematics teachers.

For this reason, and based on the conclusions of Villarroel *et al.* (2015), it can be indicated that the objective of this work is to analyse the capacity of middle school (or secondary) mathematics teachers to make an evaluative judgment in relation to the complexity of the PISA 2015 items in Chile, as well as estimate or predict the performance of Chilean students in them. In this regard, it is worth clarifying that this competence of teacher evaluative judgment is studied because it is the teachers who construct, apply, and interpret the evaluations in the classroom. The greater the distance and differences between the skills assessed in these tests and standardized measurements such as PISA, the less likely it is that students' performance in the latter will improve.

In short, the research question that guided this study was the following: is there agreement in the difficulty of the PISA 2015 items according to the teachers' evaluation and the levels of success and error that the Chilean students presented after their application?

Materials and method

The research was carried out using a non-experimental cross-sectional design, corresponding to a quantitative data expansion model (Onwuegbuzie and Collins, 2007). In addition, a quantitative data analysis was carried out by applying a questionnaire with closed and open response questions. Then, a content analysis of the opinions collected was carried out.

Sample

The sample was non-probabilistic, with intentional or convenience sampling, where participants were selected following specific criteria. Firstly, the inclusion criteria were the following: a) being high school mathematics teachers, b) having taught classes in the second year of high school in the province of Concepción, Chile, during the last two years, and c) the participating establishments had to have a SIMCE (Education Quality Measurement System) score within the national average for their agency. That is, public schools were in a score range between 240 to 250, subsidized private schools were between 260 to 290, and paid private schools were over 290 points.

In total, 18 mathematics teachers participated, of which 7 were women and 11 men. Regarding educational dependency, 7 belonged to public education, 6 to subsidized education and 5 to private education. The average age of the teachers was 47 years (SD = 4.3). Regarding teaching experience, it varied between 26 and 15 years, with an average of 23 years (SD = 5.0). All participants had a master's degree.

Procedure

Following the line of research by Villarroel *et al.* (2015), ten items in mathematics were selected: five in which Chilean students showed high performance (70% or more correct) (items: 12, 4, 16, 18, 20), and five in which they showed low performance (30% or less correct answers) (items: 11, 13, 15, 17, 19) in the PISA 2015 mathematics test (Table 1). Of the 10 questions, 2 were multiple choice (SM), 2 were complex multiple choice (SMC), 2 were closed-type constructed response (RCC), 2 were short-answer (RC), and 2 were open-type constructed response. (RCA).

These items were distributed in a varied manner in a *dossier-type document* that was delivered to the mathematics teachers in charge of evaluating them, which provided the item number and the correct answer option, in addition to an evaluation guideline for each item (the

following section will describe this instrument). The teachers evaluated blindly, that is, they did not know the performance of the Chilean students in these PISA questions.

Table 1. Identification of high and low success items according to test and percentage of success.

Math			
Performance level	Item type	*Item No.*	*% of success*
High Performance	YE	eleven	79.9
	SMC	13	27.7
	RCC	fifteen	54.9
	R.C.	17	52.1
	RCA	19	53.3
Low Performance	YE	12	29.9
	SMC	14	19.9
	RCC	16	26.7
	R.C.	18	7.60
	RCA	twenty	15.0

Source: self-made.

During the first approach to the establishments, the principals were interviewed to request their authorization through a letter to apply the study with the teachers. Both the school principal and the teachers signed an informed consent where they agreed to participate and authorized the use of the data collected for the research. Each teacher was informed of the objective of the study and was given the set of items. Teachers were informed that they would evaluate items that showed a higher degree of success and error in the PISA test, but the students' results were not specified in order to avoid bias when responding.

In the first part, teachers had to mark in a box their degree of agreement or disagreement with the indicator presented in each dimension. In the second, they evaluated the items that they evaluated in the first part using an X at one of the four proposed levels; then, they answered an open question related to the students' possible performance on each item. The guidelines were returned by teachers within 30 days of receipt.

Instrument

An evaluation guideline was designed for each item, which considered dimensions, indicators, and levels of agreement for each statement, in addition to two questions. This guideline was built from the literature consulted. Furthermore, it was assessed by judges, which yielded an intraclass correlation index (ICC= 0.85). The guideline consisted of five dimensions:

1. Item formulation: This dimension evaluated the wording of the item, the relevance of its vocabulary, the students' familiarity with the questions, and whether the information provided was useful to solve the problem in question. A higher score indicated that the item was better formulated.
2. Item content: It evaluated whether the item content was present in the school curriculum and whether it was treated primarily or not. A higher score implied that the content evaluated in PISA had been addressed in the Chilean school curriculum.
3. Contextualization of the item: It measured how contextualized and authentic the problem was, how realistic the item was and whether it showed a relevant problem and possible application to the student's daily life. A higher score indicated greater contextualization of the item.
4. Cognitive ability: This dimension evaluated the ability to reproduce, analyze and reason. A higher score indicated that higher-order cognitive skills, related to analysis and reasoning, were being measured.
5. Item complexity: This dimension sought to determine how complex the resolution of the problem would be for the students in terms of the content and form of the items. A higher score indicated greater complexity of the item.

Each dimension had 16 indicators, which were answered considering five levels: (1) strongly disagree, (2) disagree, (3) neither agree nor disagree, (4) agree, and (5) strongly agree.

Finally, the evaluation guideline presents a question related to the evaluation of the students' performance, which was divided into four levels: (1) very bad, (2) bad, (3) good, and (4) very good. In addition, it included an open question related to the students' possible performance on the item.

Analysis of data

A descriptive analysis of the data was carried out according to sex, establishment, dependency, type of test, evaluation of each indicator and prediction of the level of student performance in each item. Then, the degree of agreement between teachers in relation to the evaluation of each item and its five dimensions was analysed, using the descriptive Kripendorff alpha statistic (Hayes and Krippendorff, 2007). Likewise, the evaluation of the five dimensions was examined through the average of the evaluations to determine the dimension best evaluated by them.

Subsequently, the agreement between the teachers' prediction of the students' performance in the analysed items was assessed. Since the variables did not follow a normal distribution, it was decided to use the non-parametric Kruskal-Wallis test.

Finally, content analysis was carried out on the attributions to the students' performance, accompanied by some prototypical excerpts for each category. Teachers were coded with an Arabic number, a letter to identify sex (*m* for women and *h* for men) and a letter for administrative dependency (*m* for public, *ps* for subsidized private and *pp* for paid private). For example, teacher number 1 corresponds to a man from a paid private school (teacher 1hpp).

Results

The results derived from the quantitative analysis are presented below, followed by the production of qualitative data.

Teachers' description of the difficulty of the items

Tables 2 and 3 present the average (M) of the evaluation carried out by the teachers on the mathematics items on a scale ranging from 1 (strongly disagree) to 5 (strongly agree). The indicators presented in the tables correspond to the description that best represents the statement proposed in the item evaluation guideline.

Table 2 reveals that the dimensions that explain the students' difficulties in responding appropriately to the items are content (M = 3.21), cognitive ability (M = 3.81) and complexity (M = 3.89), which shows differences significant with respect to the rest of the dimensions (chi square = 7.82; $p < .05$).

In the *content dimension*, the lowest indicator was application, while in the rest of the dimensions all indicators scored high (*analyse*, *integrate* and *plan stand out)*. Likewise, a low score

is observed in the *known indicator* of the *item formulation* dimension. In summary, a student would have more difficulty solving these questions if the type of item is little known, the content requires application, the cognitive skill assessed is related to analysis, and the item involves planning and integration of knowledge.

Table 2. Average evaluation of teachers in mathematics for items with low accuracy.

		Underperforming items (type)					
Dimension	Indicator	12 (YE)	14 (SMC)	16 (RCC)	18 (RC)	twenty (RCA)	*M*
Formulation of the item	Drafting	4.06	4.33	3.89	4.61	4.39	4.25
	Vocabulary	4.06	4.50	4.17	4.61	4.50	4.36
	Acquaintance	3.17	3.39	2.00	4.33	3.94	3.36
	Information	4.11	4.44	3.33	4.67	4.44	4.19
	M	3.84	4.16	3.34	4.55	4.31	4.04
Content	Curriculum	4.39	4.50	2.94	4.78	4.44	4.21
	Frequency	3.39	3.67	2.50	4.44	3.56	3.50
	Application	2.00	1.61	3.11	1.39	1.61	1.94
	M	3.26	3.26	2.85	3.53	3.20	3.21
Contextualization	Realistic	4.11	3.94	3.78	4.44	3.39	3.93
	Important	3.56	3.22	2.94	3.39	2.78	3.17
	Familiar	3.39	3.50	2.44	3.61	2.61	3.11
	M	3.68	3.55	3.05	3.81	2.92	3.40
Ability cognitive	play	2.67	3.11	3.33	3.83	3.44	3.27
	Analyze	4.44	4.50	4.39	3.89	4.11	4.26
	Reflect	4.39	4.44	3.83	2.94	3.94	3.90
	M	3.83	4.01	3.55	3.55	3.83	3.81
Complexity	To integrate	4.44	4.44	3.83	3.28	3.94	3.98
	To plan	4.17	4.28	3.83	3.56	3.94	3.95
	Complex	4.06	4.39	4.00	2.61	3.72	3.75
	M	4.22	4.37	3.88	3.15	3.86	3.89
	M total	3.78	3.89	3.39	3.77	3.67	3.70

Source: self made

Table 3 shows that, according to the mathematics teachers, the dimensions that most facilitate the resolution of the items are *item formulation* (M = 4.41) and *contextualization* (M = 3.98), since they obtained the highest averages and show significant differences with respect to the rest of the dimensions (chi square = 7.33; $p < .05$). Regarding the indicators that would facilitate the resolution of the items, the most mentioned were writing, vocabulary, acquaintance and information, corresponding to the *formulation dimension of the item;* as well as the realistic, important and familiar indicators of the *contextualization dimension* . In other words, a student would have less difficulty solving these items if the item formulation is clear, the vocabulary and item type are known, and the context is familiar.

Table 3. Average teacher evaluation in mathematics for highly correct items

		High performance items (Type)			
Dimension	Indicator	twenty-one (YE)	23 (SMC)	26 (RCA)	*M*
Formulation of the item	Drafting	4.89	4.33	4.67	4.63
	Vocabulary	4.67	4.50	4.67	4.61
	Acquaintance	4.78	3.78	4.17	4.24
	Information	4.50	4.00	4.00	4.16
	M	4.71	4.15	4.37	4.41
Content	Curriculum	4.67	3.56	3.44	3.89
	Frequency	4.67	3.39	2.89	3.65
	Application	2.33	3.28	3.00	2.87
	M	3.89	3.41	3.11	3.47
Contextualization	Realistic	4.67	4.17	4.33	4.39
	Important	4.44	3.17	3.50	3.70
	Familiar	4.33	3.56	3.67	3.85
	M	4.48	3.63	3.83	3.98
Ability Cognitive	play	4.11	3.11	2.61	3.27
	Analyze	2.83	2.50	2.33	2.55
	Reflect	2.72	2.44	2.33	2.49
	M	3.22	2.68	2.42	2.77
Complexity	To integrate	2.50	2.89	2.78	2.72
	To plan	2.11	2.56	2.33	2.33
	Complex	2.06	2.78	2.44	2.42
	M	2.22	2.74	2.51	2.49
	M total	3.77	3.38	3.32	3.49

Source: self made

Agreement between teachers

Table 4 shows the evaluation carried out by the teachers for each item and its five dimensions: item formulation (Form), content (Conten), contextualization (Contex), cognitive ability (HCog) and complexity (Com). An average is presented to highlight the dimension best evaluated in the area of mathematics. Additionally, the degree of agreement is observed in relation to the evaluation of each item in the areas investigated.

Some agreement is evident between the teachers' evaluations, especially regarding the high-performance items, in which there is greater agreement. Regarding the dimensions, it is observed that in all areas the *item formulation dimension* has greater agreement among teachers, while *complexity* has a lower average agreement.

Table 4. Average evaluation of teachers in mathematics, according to item and dimension

Item	Guy	*Form*	*Contain*	*Contex*	*HCog*	*Com*	*α*	(95% CI)
High performance	YE	4.64	3.54	4.15	3.11	2.09	.60	(.55-.65)
	SMC	4.53	3.52	3.93	3.74	2.84	.37	(.29-.43)
	RCC	4.64	3.65	4.18	3.33	2.26	.57	(.51-.62)
	R.C.	4.54	3.41	3.13	3.17	2.28	.49	(.43-.55)
	RCA	4.08	3.46	3.59	3.31	2.43	.32	(.24-.39)
	M	4.48	3.51	3.79	3.33	2.38	.47	
Low performance	YE	3.85	3.26	3.69	3.83	4.22	.24	(.16-.32)
	SMC	4.16	3.26	3.55	4.02	4.37	.29	(.22-.36)
	RCC	3.34	2.85	3.05	3.85	3.89	.24	(.15-.32)
	R.C.	4.56	3.54	3.81	3.55	3.15	.Four. Five	(.39-.51)
	RCA	4.32	3.20	2.93	3.83	3.87	.27	(.20-.35)
	M	4.04	3.22	3.40	3.81	3.90	.29	
	M total	4.27	3.37	3.60	3.57	3.14		

Source: self made

Finally, the analysis of agreement between the teachers' assessment and the students' results by areas —following the criteria established in the study by Villarroel *et al.* (2015)—revealed that a significant degree of agreement is considered when the percentage is equal to or greater than 80%. The teachers evaluated the students' performance on four levels: (1) very bad, (2) bad, (3) good, (4) very good.

In this regard, it was found that teachers have a higher percentage of agreement in the highly correct items, since they managed to agree in 60% of the evaluations and predict 80% of the students' performance in the high-performance items in mathematics. However, 40% of mathematics teachers have difficulty in accurately answering the items where students showed poor performance in mathematics. In the low-performance items, they only reached 21.42% agreement.

Teaching differences according to dependency

Table 5 shows that, in the majority of the high-performance items in the mathematics area, no statistically significant differences were found in the assessment made by the judges of the different departments. Only in the RCA item was a statistically significant difference found, with a value of $\chi 2 = 6.011$ and $p = .05$. In the low accuracy items, no statistically significant differences were identified.

Table 5. Comparison of judges' assessment by school dependency

Item	Guy	Public (n=7)	Subsidized (n=6)	Private (n=5)	χ^2	*p*
High performance	YE	9.64	9.67	9.10	0.057	.972
	SMC	9.64	9.67	9.10	0.050	.975
	RCC	9.14	10.00	9.40	0.120	.942
	R.C.	8.14	10.17	10.60	1,029	.598
	RCA	6.50	12.50	10.10	6,011	.050
Low performance	YE	10.36	9.17	8.70	0.456	.796
	SMC	8.36	11.42	8.80	1,466	.480
	RCC	9.93	11.92	6.00	3,867	.145
	R.C.	8.64	13.17	6.30	5,506	.064
	RCA	7.86	11.00	10.00	1,550	.461

Source: self made

No significant differences were observed according to age, sex and teaching experience between the mathematics teachers' judgments in the different items and dimensions analyzed.

Content analysis

The content analysis was carried out considering the teachers' assessment of the students' performance in mathematics in each of the items, their agreement with the results obtained in PISA and the attributions expressed by the teachers regarding the students' performance. This analysis was divided into high and low performance items to determine the presence or absence of agreement between teachers, along with the reasons they gave for this.

In relation to the agreement between the teachers' perception and the students' results in this area, it was observed that, of the ten items evaluated, five corresponded to the high-performance category and five to low performance. Of the high-performance items, four showed greater agreement between the teachers' assessment and the students' good results. In these items, the teachers based their attributions of high performance on the notion that the cognitive ability required basic reasoning for its resolution and that the content was frequently worked on.

> *Only the item requires remembering the order of decimal numbers, a situation that is worked on from fifth grade. Therefore, this item is frequently practiced in the national curriculum. Therefore, the results of this question should be optimal* (teacher 26hps).

In item 13, a group of teachers described the same reasons for high performance mentioned above. On the other hand, those who rate the item as difficult for students argue that they are not familiar with it and that they are required to analyse variables, which requires greater reading comprehension and higher-order cognitive ability.

> *The topic is very familiar in the student's life, but the resolution of the problem is very complex and affects their reading comprehension* (teacher 41 hm).

Regarding low performance items, in three of them there is agreement between the teachers' perception and the low results obtained by the students in mathematics. In items 12, 14 and 20, the teachers considered that the reasons for the students' poor performance would be given by the complexity they present. In addition, the wording of the item would also affect poor performance due to length and wording, which makes reading comprehension difficult.

> *The problem contains a lot of information, which can make it difficult to understand. You have to work with more than one, you have to relate the information to what is requested, which may not be clear* (teacher 7mpp).

In item 16, some teachers attribute the same reasons for poor performance given in items 12, 14 and 20, namely, cognitive ability, content, complexity, and item formulation. Added to this is the lack of information and the contextualization of the item since the situation presented is not familiar to the student.

> *Working with a "new" numbering system for students is complex. Furthermore, it is not clearly explained how to write a fraction in this system, therefore, it is possible that they do not understand it and respond incorrectly* (teacher 44mm).

Item 18 was rated in the low difficulty category, and the teachers considered that the students performed well, as they argued that it demands basic skills. However, there is no agreement with the student results obtained.

> *It is very familiar and "similar" to what was worked on* (teacher 2hpp).

On the other hand, teachers who rated this item with low performance expressed that the difficulty comes from the integration of different variables at the same time, that is, it requires greater cognitive skills and management of content that is not addressed or exercised in its entirety in classes.

The answer is not in the context presented, since the student must resort to prior knowledge. No procedure to follow is indicated. There is no obvious formula to apply. The student is required to extract from their own knowledge the strategy to follow to achieve the solution. It is necessary to recognize the variables involved (side length), surface area of a square area (teacher 45hm).

Discussion

Educational assessment plays a crucial role in the teaching and learning processes by allowing teachers to know the impact of their pedagogical practices in the classroom and adjust them as necessary to improve student learning. This study has provided an understanding of mathematics teachers' evaluative judgment on the difficulty of items in the PISA test, as well as the agreement of their opinions with the results obtained by students. These findings contribute to the promotion of an evaluative culture that recognizes error as a learning opportunity in the classroom. Furthermore, the importance of the active participation of teachers as key agents in improving education is highlighted.

Regarding the quality of teachers' evaluative judgment, it is concluded that, in general, they have difficulties in recognizing the complexity of the items, especially those in which students show low performance, in contrast to what was observed in the French context. (Le Hebel *et al.*, 2019). However, teachers are able to better predict the level of difficulty of the items where students perform better. The same occurs with the agreement between teachers, which is greater when the items with the best student performance are evaluated.

Mathematics teachers tend to achieve better prediction of item difficulty when they focus on how familiar their structure is to the student. The key indicator in this sense is the one titled "known" within the *problem formulation dimension.* This finding is related to what was described by Villarroel *et al.* (2015), where it is suggested that the most challenging items for students are those with complex multiple responses, which are not common in national educational practice.

In relation to the dimensions evaluated, from the perspective of mathematics teachers, those that seem to favour student performance are the formulation of the item, particularly when the indicators associated with writing and vocabulary are high.

On the other hand, the difficulty of the items seems to be more linked to the teacher's evaluation of their complexity, especially when it involves the integration of variables, planning and subdivision of tasks. Although more research is still needed on context familiarity and its impact on student performance, it is plausible to claim that a familiar environment can facilitate

mathematical problem solving. However, it is still difficult for students to abstract and transfer the mathematical structure of the problem at hand, as Almuna-Salgado (2017) points out.

Regarding the opportunities for improvement of this study, it is crucial to strengthen the sample of teachers who served as judges of the complexity of the items for future research. In this sense, it is important to keep in mind that the teaching of mathematics has experienced significant changes, such as the increase in virtual instruction and the development of pedagogical currents that focus on the application of mathematics in real contexts and in everyday life, among other aspects (Cantoral, 2020).

Therefore, the questions of future assessments, both those of PISA and those created by each teacher, must adapt to these new conditions, as well as the pedagogical practices implemented in the classroom.

Finally, a strong aspect of the study lies in the use of a research design that incorporates open questions, which has allowed us to delve deeper into the teachers' evaluations and attributions regarding the items and the students' performance.

It is concluded, therefore, that it is necessary to advance the research on educational evaluation, especially with regards to the analytical categories used in order to develop alternative strategies for both initial and continuous teacher training in the field of learning evaluation. Furthermore, it would be interesting to replicate this methodology to examine performance against other standardized assessments, both nationally and internationally, and with other study populations.

Conclusions

The results up to this point suggest that to improve student performance it is crucial to facilitate familiarity and contextualization of the items, both for students and teachers. Furthermore, it is evident that the latter face difficulties in identifying the level of skills evaluated in the items, since they tend to think that their students will have lower performance in items of greater complexity that measure cognitive analysis skills.

Therefore, we must work with teachers to recognize and apply skills, especially higher-level skills. It is also necessary to introduce this type of items, which measure high-order skills, more frequently in classroom evaluations.

It is possible that the difficulty in recognizing the cognitive abilities measured by the items is due to a functional approach to learning, in which knowledge is measured in a more rote and literal way. Therefore, the challenge of designing and proposing realistic problematic situations

where the different levels of cognitive complexity are transversally incorporated, in line with the results of the reviewed literature, is seen.

Future lines of research

The challenge for mathematics teachers to identify items of greatest difficulty for students is a crucial area that requires continued analysis. Teachers' ability to estimate the level of difficulty of test items, as well as their ability to predict students' success or failure on them, reflects their competence in the field of learning assessment. If there are difficulties in this aspect, it is likely due to a deficit in this specific teaching competence, as suggested by studies that indicate that classroom evaluation is the most lacking area in pedagogical practice, as evidenced in performance evaluations carried out on teachers in the public school system in Chile in recent years (Manzi *et al*., 2011).

For this reason, it is essential to continue conducting research with teachers, especially in the field of assessment, to better understand the assessment culture that influences their pedagogical practices and beliefs about assessment in the field of mathematics. Furthermore, it would be pertinent to delve deeper into the students' own evaluation of the difficulty of the PISA items. This would allow information to be obtained about their perception of success or failure, degree of familiarity with the different items, and level of motivation regarding the form and content of the evaluation. Integrating this information into classroom planning and assessment could significantly enrich teaching practice.

References

Andrade, T. y Guzmán, I. (2018). Educación matemática y formación ciudadana: un estudio que confronta la matemática escolar, el currículo y las prácticas docentes. *Revista Paradigma*, *39*(1), 319-331. http://funes.uniandes.edu.co/16281/1/Andrade2018Educaci%C3%B3n.pdf

Angier, C. (2019). How do beginning mathematics teachers in Scotland understand their role in education for global citizenship? In F. Curtis (ed.), *British Society for Research into Learning Mathematics*. BSRLM.

Almuna-Salgado, F. J. (2017). The role of context and context familiarity on mathematics problems. *Revista Latinoamericana de Investigación en Matemática Educativa, 20*(3), 265–292. https://doi.org/10.12802/relime.17.2031

Aydın, U. and Özgeldi, M. (2017). The PISA tasks: Unveiling prospective elementary mathematics teachers' difficulties with contextual, conceptual, and procedural knowledge. *Scandinavian Journal of Educational Research*. https//:doi.org/10.1080/00313831.2017.1324906

Benson-O'Connor, C. D., McDaniel, C. and Carr, J. (2019). Bringing math to life: Provide students opportunities to connect their lives to math. *Networks: An Online Journal for Teacher Research, 21*(2). https://eric.ed.gov/?id=EJ1224221

Caño, A. y Luna, F. (2011). *PISA: comprensión lectora. Marco y análisis de los ítems.* Instituto Vasco de Evaluación e Investigación Educativa. http.//www.isei-ivei.net/cast/pub/itemsliberados

Canales, A. & Maldonado, L. (2018). Teacher quality and student achievement in Chile: Linking teachers' contribution and observable characteristics. *International Journal of Educational Development, 60*, 33–50. https://ideas.repec.org/a/eee/injoed/v60y2018icp33-50.html

Cantoral, R. (2020). La matemática educativa en tiempos de crisis, cambio y complejidad. *Revista Latinoamericana de Investigación en Matemática Educativa, 23*(2), 143-146. https://doi.org/10.12802/relime.20.2320

Claessens, A. and Engel, M. (2013). How important is where you start? Early mathematics knowledge and later school success. *Teachers College Record, 115*(6), 1–29. https://doi.org/10.1177/016146811311500603

Hayes, A. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Method and Measures, 1*(1), 77-89. https://doi.org/10.1080/19312450709336664

Le Hebel, F., Tiberghien, A., Montpied, P. and Fontanieu, V. (2019). Teacher prediction of student difficulties while solving a science inquiry task: Example of PISA science ítems.

International Journal of Science Education, 41(11), 1517-1540. https//:doi.org/10.1080/09500693.2019.1615150

Manzi, J., González, R. y Sun, Y. (eds.) (2011). *La evaluación docente en Chile.* Facultad de Ciencias Sociales. Escuela de Psicología, PUC. https://www.mideuc.cl/web19/wp-content/uploads/Libro-Ev-Docente-en-Chile-FINAL-2011-07-20.pdf

Menoyo Díaz, M. D. P. (2020). Educar la mirada científica del alumnado de secundaria en el marco de los objetivos de desarrollo sostenible, educar para una ciudadanía global en un momento de cambio educativo. *Modelling in Science Education and Learning*, *13*(2), 21-42. https://dialnet.unirioja.es/servlet/articulo?codigo=7561444

Morales Salas, R. E. (2018). La planeación de la enseñanza-aprendizaje, competencia que fortalice el perfil docente. *Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, *8*(16), 311-334. https://doi.org/10.23913/ride.v8i16.343

Onwuegbuzie, A. J. and Collins, K. M. (2007). A typology of mixed methods sampling designs in social science research. *The Qualitative Report, 12*(2), 281-316. https://files.eric.ed.gov/fulltext/EJ800183.pdf

Organización para la Cooperación y el Desarrollo Económico (OCDE) (2010). *Resumen Resultados PISA 2009 Chile*. OECD/PISA.

Organización para la Cooperación y el Desarrollo Económico (OCDE) (2019). *PISA 2018 Results (volume I): What Students Know and Can Do*. OECD Publishing. https://doi.org/10.1787/5f07c754-en

Radišić, J. and Baucal, A. (2018). Teachers' reflection on PISA items and why they are so hard for students in Serbia. *European Journal of Psychology of Education, 33*, 445-466. https//:doi.org/10.1007/s10212-018-0366-0

Stipek, D., Schoenfeld, A. and Gomby, D. (2012). Math matters, even for little kids. *Education Week, 31*(26), 27-29. https://www.edweek.org/teaching-learning/opinion-math-matters-even-for-little-kids/2012/03

Torres, R. (2018). Tackling inequality? Teacher effects and the socioeconomic gap in educational achievement. Evidence from Chile. *School Effectiveness and School Improvement, 29*(3), 383–417. https://doi.org/10.1080/09243453.2018.1443143

Trejo, A. L. (2020). ¿Cómo vincular la enseñanza de las matemáticas con el desarrollo social sostenible y la socio-formación? *Revista de Ciencias Sociales y Humanidades*, *5*(24), 49-61. https://dialnet.unirioja.es/servlet/articulo?codigo=8274318

Valenzuela, J. P., Bellei, C., Sevilla, A., y Osses, A. (2009). ¿Qué explica las diferencias de resultados PISA Matemática entre Chile y algunos países de la OCDE y América Latina? En L. Cariola, G. Cares, y E. Lagos (eds.), *¿Qué nos dice PISA sobre la educación de los jóvenes en Chile? Nuevos análisis y perspectivas sobre los resultados en PISA 2006* (pp. 105-148). Ministerio de Educación, Unidad de Curriculum y Evaluación.

Valenzuela, J. P., Gómez, G. and Sotomaryor, C. (2015). The role of reading engagement in improving national achievement: An analysis of Chile's 2000–2009 PISA results. *International Journal of Educational Development*, *40*, 28-39. https//:doi.org/10.1016/j.ijedudev.2014.11.011

Villarroel, V., García, C., Melipillán, R., Achondo, E. and Sánchez, A. (2015). Aprender del error es un acierto. Las dificultades que enfrentan los estudiantes chilenos en la Prueba PISA. *Estudios Pedagógicos*, *41*(1), 293-310. https//:doi.org/10.4067/S0718-07052015000100017

Villegas-Zamora, D. A. (2019). La importancia de la estadística aplicada para la toma de decisiones en Marketing. *Revista Investigación y Negocios*, *12*(20), 31-44. http://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S2521-27372019000200004&lng=es&tlng=es

Contribution Role	Author(s)
Conceptualization	Verónica Villarroel Henríquez, María Angélica San Martín Toro, Carmen Núñez Ramos, Carolina Hernández Opazo
Methodology	Verónica Villarroel Henríquez, María Angélica San Martín Toro, Carmen Núñez Ramos, Carolina Hernández Opazo
Software	Does not apply
Validation	Does not apply
Formal Analysis	María Angélica San Martín Toro, Carmen Núñez Ramos, Carolina Hernández Opazo, Alejandro Sánchez Oñate
Investigation	María Angélica San Martín Toro, Carmen Núñez Ramos, Carolina Hernández Opazo
Resources	Veronica Villarroel Henríquez
Data curation	María Angélica San Martín Toro, Carmen Núñez Ramos, Carolina Hernández Opazo, Alejandro Sánchez Oñate
Writing - Preparation of the original draft	Isidora Castillo Rabanal, María Angélica San Martín Toro, Carmen Núñez Ramos, Carolina Hernández Opazo
Writing - Review and editing	Verónica Villarroel Henríquez, Isidora Castillo Rabanal, María Angélica San Martín Toro, Carmen Núñez Ramos, Carolina Hernández Opazo.
Display	Isidora Castillo Rabanal, María Angélica San Martín Toro, Carmen Núñez Ramos, Carolina Hernández Opazo, Alejandro Sánchez Oñate
Supervision	Veronica Villarroel Henríquez
Project management	Veronica Villarroel Henríquez
Fund acquisition	Does not apply