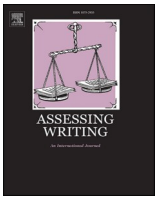
Assessing Writing 62 (2024) 100898

Contents lists available at ScienceDirect

Assessing Writing

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



Effects of a genre and topic knowledge activation device on a standardized writing test performance

Natalia Ávila Reyes [a,*], Diego Carrasco [a], Rosario Escribano [a], María Jesús Espinosa [b], Javiera Figueroa [c], Carolina Castillo [a]

[a] Pontificia Universidad Católica de Chile, Chile
[b] Universidad Diego Portales, Chile
[c] Universidad Alberto Hurtado, Chile

ARTICLE INFO

Keywords:
Large-scale assessment
Construct-irrelevant variance
Experimental design
Informative writing
Social justice

ABSTRACT

The aim of this article was twofold: first, to introduce a design for a writing test intended for application in large-scale assessments of writing, and second, to experimentally examine the effects of employing a device for activating prior knowledge of topic and genre as a means of controlling construct-irrelevant variance and enhancing validity. An authentic, situated writing task was devised, offering students a communicative purpose and a defined audience. Two devices were utilized for the cognitive activation of topic and genre knowledge: an infographic and a genre model. The participants in this study were 162 fifth-grade students from Santiago de Chile, with 78 students assigned to the experimental condition (with activation device) and 84 students assigned to the control condition (without activation device). The results demonstrate that the odds of presenting good writing ability are higher for students who were part of the experimental group, even when controlling for text transcription ability, considered a predictor of writing. These findings hold implications for the development of large-scale tests of writing guided by principles of educational and social justice.

1. Introduction

Standardized writing assessment in school education can provide relevant information about learning to various stakeholders. This information is helpful in evaluating the state of an educational system and guiding public policy (Manzi et al., 2019). However, writing is a complex skill that, far from being merely a school-based ability, is learned throughout life and it is strongly socially-mediated (Bazerman, 2013; Boscolo, 2009; Wijekumar et al., 2019). Given its complexity, which encompasses factors that are individual and social, cognitive and cultural, linguistic and contextual, evaluating writing on a large scale presents numerous challenges for test developers and educational systems.

Moreover, large-scale writing assessment presents challenges related to justice and equity, which are inherent to any testing regime. As Stein (2016) argues, there is an intrinsic relationship between testing and social justice, in which an assessment infrastructure is unjust if it benefits those already advantaged while punishing those who face structural disadvantages. Due to the inherently social, cultural, and contextual nature of writing, its standardized measurement has historically tended to reproduce patterns of injustice, consistently disadvantaging more marginalized groups (Randall, Poe, Slomp, & Oliveri, 2024; Castillo & Ávila

* Corresponding author.
E-mail address: naavila@uc.cl (N. Ávila Reyes).

https://doi.org/10.1016/j.asw.2024.100898
Received 12 December 2023; Received in revised form 22 October 2024; Accepted 7 November 2024
Available online 22 November 2024

Reyes, In Review), for instance, through a lack of attention to contextual dependency, the complexity of the construct, or the diversity of test-takers (Broad, 2000).

For this reason, issues of social justice and equity have been central to a recent research agenda on writing assessment (Beck et al., 2020; Chamorro, 2022; Inoue, 2015; Poe & Inoue, 2016; Reed et al., 2023; Sims, 2023). However, these important contributions have mostly focused on local and classroom forms of assessment. Indeed, part of this tradition even shows skepticism toward standardized writing assessment (Hammond, 2017).

In this article, we align with authors such as Zachary Stein (2016), who assert that measurement can be useful to educational systems and even advance social justice in education. However, more often than not, the misuse of these tools leads to forms of injustice. Hence, it is crucial to rethink the design of assessment instruments in terms of social justice and to critically examine with technical evidence how changes in test design can contribute to fairer evaluations.

An essential concern in large-scale assessment is validity, which refers to the degree to which evidence and theory support the interpretations that can be made from test scores. One of the ways in which validity can be affected is by measuring more than the construct of interest, known as construct-irrelevant variance, which refers to how a test score can be influenced by processes that are extraneous to the test's purpose (American Educational Research Association et al., 2014; Cushing Weigle, 2002). Since writing is an activity that involves various factors, diverse elements can result in sources of variance unrelated to the writing ability, and in particular, many of them reveal implicit racial, linguistic, or cultural biases (Randall et al., 2024).

There are some research precedents that have explored the effects of implementing technical measures in the construction of standardized tests aimed at improving either their validity, fairness, or justice. For example, Ghanbari (2019) contributes to the discussion on equity in writing assessment by demonstrating how sharing evaluation criteria with test-takers grants them greater agency and understanding of the evaluative process, which not only improves their performance but also makes the assessment more transparent, democratic, and fair. This challenges the traditional power dynamic between evaluators and students, promoting more ethical and equitable assessment practices, and can be extrapolated to the design and implementation of large-scale writing tests.

Another study investigated the effect of time allocated for writing tests, based on the premise that time limitations might offer fewer opportunities for students to demonstrate their ability. The study found no significant differences but noted that more proficient students gained a greater advantage from the extra time provided (Knoch & Elder, 2010).

In the pursuit of implementing technical measures in the design of standardized tests that increase validity and consequently fairness, this article describes the overall design of a theoretically-sound larga-scale writing test, and the development of a device aimed at activating prior knowledge related to two factors that could impact students' ability to demonstrate their writing skills: genre knowledge and topic knowledge.

Motivation for writing and prior knowledge stand out among the various elements different from the writing ability that may impact the writing performance (Graham et al., 2017). A way of accounting for motivation may be letting test-takers to choose the topic on which they would write. Indeed, Perelman (2018) points out that the choice of prompts can improve fairness by allowing test takers to decide based on their prior knowledge and interests. However, there is recent evidence that the effects of topic choice in writing quality are limited (Aitken et al., 2022). Thus, this article purposefully examines only the effect of prior knowledge genre and topic knowledge, employing a fixed prompt that emerged from writing themes that were most preferred by students in a previous survey.

In an evaluative situation where students are prompted to write a text on a specific topic and in one particular genre, they would need resources to develop the topic and understand the requested format. This is essential for composing texts that meet writing quality criteria, such as fulfilling the communicative purpose or structuring coherent and cohesive responses. However, access to this prior knowledge is mediated by students' previous experiences, potentially hindering some students' ability to demonstrate their composing skills. Thus, implementing forms of scaffolding in the design of standardized assessment might be a useful measure towards more just forms of assessment (Perelman, 2018; Randall et al., 2024; Stein, 2016).

The following pages explain how a standardized test was designed using a sociocognitive approach (Bazerman et al., 2009; Corrigan & Slomp, 2021) and based on an authentic task (Duke et al., 2006; Purcell-Gates et al., 2007; Tolchinsky, 2008). The test protocol includes a device to provide access to information about the topic through an infographic and a model of the genre to be written. Using an experimental design, this task was administered to an intervention group (applying the activation device) and a control group (without applying the activation device). The text produced was later evaluated using a rubric with five dimensions that do not directly assess aspects of the topic or genre but other traits central to composition. The results and potential applications for test design in standardized writing assessments are discussed.

2. Genre knowledge and topic knowledge: theory and evidence

In their seminal work on the strategies used by novice writers, Bereiter and Scardamalia (1987) identify the crucial role of knowledge of both genre –understood as knowledge of prototypical discourse structures– and of the topic –understood as the theme to be developed in their texts. In the "knowledge telling" strategy, novice writers use memory cues derived from either the topic or the genre of the assigned task and retrieve significant information with which to compose their text (Bereiter & Scardamalia, 1987). In a comprehensive review of processes involved in managing the complexity of writing, McCutchen (2011) identifies, in addition to linguistic processes of textual production, knowledge of the genre and knowledge of the topic as other "writing-related knowledge" that intervenes in writing. The following review addresses both knowledge types, providing evidence to support their pivotal role in written composition.

2.1. Genre knowledge

According to McCutchen (2011), familiarity with the genre theoretically impacts writing by providing access to schemas that can facilitate planning, revision, and even affect working memory demands by translating ideas into text. This is further supported by the discursive-rhetorical knowledge possessed by advanced writers. The evidence reviewed by the author suggests a link between genre knowledge and writing ability. For example, one study demonstrated significant improvement in the quality of texts written by children after genre instruction (Fitzgerald & Teasley, 1986). In various studies, children showed greater ability in familiar genres, such as narratives over expository texts (Cox et al., 1991; Hidi & Hildyard, 1983; Langer, 1986, McCutchen, 1987). In another study, students derived useful macrostructures for writing based on their familiarity with the genre (McCutchen et al., 1997).

Several pieces of evidence confirm the influence of genre knowledge in recent studies evaluating participants' writing. For example, Olinghouse and Graham (2009) conducted a study tracing how discourse knowledge—including, but not limited to genre—contributes to predicting the writing performance of young children. In their hierarchical regression analysis, structural knowledge of story accounted for unique variability in writing quality above the contributions of other measures. Similarly, Wang and Troia (2023) included measures of register knowledge in a predictive model of writing quality, comprising a) identifying the literary genre of each passage, b) the textual structure of informational text, and c) organizational features of passages, alongside motivation for register, measured through motivational scales for informational, narrative, and opinion texts. In this model, the influence of knowledge and motivation related to specific registers predicted writing quality, even after considering student demographic and linguistic factors. While these variables explained only a small portion (2 %-3 %) of the variance in writing quality across different registers, register-related knowledge and motivation have a stable impact on predicting writing quality.

Studies conducted at other educational levels support this relationship. For instance, (Driscoll et al., 2020) investigated the connection between genre knowledge and writing gains in university students. Different themes emerging in student reflections and gains in writing over a term were explored using correlations. Genre awareness—whether through simplistic or nuanced views of genre—was the only factor among those coded in the reflections that correlated with writing gains.

Furthermore, the study by Olinghouse et al. (2015) is noteworthy in addressing genre and topic knowledge together. Their findings regarding discourse knowledge—including, but not limited to genre—and topic knowledge corroborate the roles they play in the architecture of Bereiter and Scardamalia's (1987) "knowledge telling" model. Discourse knowledge played a unique and statistically significant role in predicting the quality and integration of genre-specific elements in various forms of writing, such as narratives, persuasive texts, and informational pieces. This contribution remained significant beyond topic knowledge and other controlled variables. Furthermore, topic knowledge independently predicted the quality of narratives, persuasive texts, and informational writing, beyond discourse knowledge and other controlled variables. Additionally, topic knowledge was identified as a predictor for incorporating genre-specific elements, particularly in informational texts.

2.2. Topic knowledge

The role of topic knowledge was early addressed by Kellogg (1987), who found no differences in allocating processing time, but found instead that high-knowledge writers expended less effort in writing overall. Topic knowledge has also been understood as a determining factor in textual production. For Graham (2018), writing ultimately depends on having something to write about. This knowledge can be retrieved either partially or entirely from long-term memory. McCutchen (2011) points to various sources of evidence showing that writers who know more about the topic can write more coherent texts than those who do not (McCutchen, 1986). Likewise, the quality of text revision improves with a familiar topic, so writers make more changes at the discursive and meaning levels of texts when revising compared to when they are unfamiliar with the topic (McCutchen et al., 1997).

The influence of topic knowledge has also been traced in measures of participants' writing. From the premises regarding the lesser effort required by high-knowledge writers to retrieve and use relevant knowledge for their writing, as well as to produce longer texts with greater content revision, Proske and Kapp (2013) examine the effects of topic knowledge on university students' composition of academic texts. Specifically, they tested whether supporting the construction of the situation model through learning questions associated with reading a source led to the production of better texts. The results found positive and significant associations with writing process times, readability, and the length of the text measured in the number of words, indicating the advantages of stimulating the development of a situation model. A study conducted by Tabari et al. (2021), also at the university level, explored whether familiarity with the topic influenced the linguistic complexity and emotional tone of writers in L2. In their theoretical model, they define familiarity with the topic as subject-matter knowledge, domain-specific knowledge, and discipline knowledge, which may be acquired through formal instruction or informal channels such as life, work, and study (i.e., experiential knowledge). While there was previous evidence of the positive relationship between topic familiarity and writing, there was no consensus on whether this effect extends to linguistic complexity. The results of the study demonstrated that familiarity with the topic among the study participants facilitated the production of greater linguistic complexity, particularly at the phrasal and clausal levels.

Lastly, in a study that measured the complexity of factors associated with writing quality (motivation, knowledge, skills, and strategic behavior) in fifth-grade students, Graham et al. (2019) found that topic knowledge had a unique and statistically positive relationship with compositional quality. Additionally, the knowledge variable—which included topic knowledge and knowledge of discourse markers—accounted for unique and statistically significant variance in compositional length.

3. A knowledge-activation device for standardized writing assessment

Following McCutchen (2011), topic and genre knowledge can be seen as factors related to writing but distinct from the attribute of writing itself. The reviewed evidence demonstrates how these crucial types of knowledge predict traits such as quality, complexity, or length, likely because it mediates the development of more complex writing processes, either through constructing situational models or facilitating access to more sophisticated writing resources. Therefore, this evidence raises the question of to what extent performance in an assessment situation, which typically utilizes a direct measure of writing stimulated by a prompt in a limited time frame (Cushing Weigle, 2002), depends on whether the student has had exposure to the specific genre of writing required and the topic they must elaborate upon. The combination of formal education and informal experiences within students' households influences their understanding of genres (Collins et al., 2021), potentially leading to unwanted sources of variance in large-scale writing assessments of this sort.

Collins et al. (2021) explain that writing prompts can either require students to rely solely on their personal experiences and prior knowledge for their writing, referred to as nonsource-based writing, or to draw evidence and examples from textual sources, known as source-based writing. The first type of task, based solely on prior knowledge, frees test-takers from the additional demands of retrieving and orchestrating information from sources. Moreover, some groups of students might struggle to comprehend the source texts, making them less effective at writing in a source-based manner. However, the authors suggest that source-based writing may scaffold student writing, offering them background knowledge for unfamiliar topics and models of how to express their ideas in words (Collins et al., 2021).

This article discusses the creation and testing process of a tool designed to provide cognitive support in a standardized testing situation by giving students access to two knowledge sources: one facilitating access to the topic—without overburdening students with the demand to read a written text—and another enabling access to the genre—providing a textual model on a different topic. These procedures are expected to act as mediators of writing-related knowledge and contribute to reducing construct-irrelevant variance, thereby enhancing students' chances of demonstrating their writing ability. Hence, our main research question is: What is the effect of a genre and topic knowledge activation device on writing performance?

4. Methods

4.1. Participants and experimental design

The participants of this study were 162 students of fifth grade (10 years of age) from 9 schools of Santiago de Chile Urban area. They were randomly assigned to each of the conditions within each school, by asking them to leave their classroom, assigning them to each condition and then applying the test to each condition in a designated classroom. 78 students were assigned to the experimental

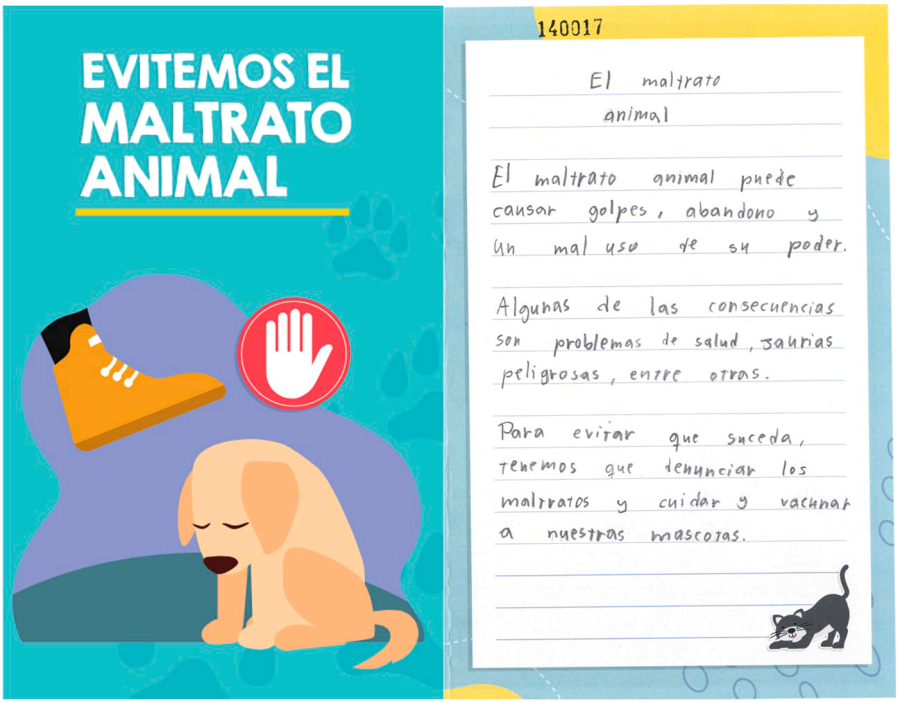


Fig. 1. Design of the response sheet: cover and inside of the brochure.

condition (with activation device) and 84 students were assigned to the control condition (without activation device). All participants had informed consents signed by their parents or tutors and additionally signed an informed assent.

4.2. The writing test

A standardized writing test, designed to be applied to fifth-grade Chilean students. The task consisted of writing an informative text. The prompt adhered to the principles of an authentic task (Duke et al., 2006; Purcell-Gates et al., 2007), which means it was embedded in a rhetorical situation with a communicative purpose, a defined genre, and an audience. To account for motivation, we previously conducted a survey to fifth graders in 9 schools (n=283), from which we selected the most chosen task (70 %) that was indistinctly preferred by girls and boys across schools with different characteristics, using the schools' previous results on national standardized language tests and the state index of poverty (known as school vulnerability index) as covariates. In the selected prompt, students were tasked with creating a brochure to be distributed in a school campaign against animal abuse. The response sheet was designed in the format of a brochure, comprised by a cover and a designated writing space within (see Fig. 1). The selection of this task aligned with a socio-cognitive approach, wherein a social context was created to support the need for writing. This also involved establishing a situation model for students to deploy writing strategies, because a communicative purpose and an audience were defined. This approach differs significantly from, for instance, responding to a decontextualized prompt that lacks a situation or purpose other than eliciting text.

The application of the test in the experimental group included the use of two instruments for prior knowledge activation. At the beginning, all the students received the exam prompt: "Lets' avoid animal abuse: This month we will make a campaign to promote animal care. Write a brochure to be distributed in our school in order to educate other children and young people. To write the brochure, explain the forms of abuse, its consequences and what to do in case of animal abuse".

Then, the test proctor activated genre knowledge by showing the students a large-size model of a brochure on a different topic (wildfires), displayed on the blackboard in front of the whole class (Fig. 2). The model is read aloud so that students can better understand the target genre through an example, and then it is removed from the blackboard. Secondly, topic knowledge was activated through an infographic given to each student. This infographic included a very limited amount of text (see Fig. 3) to avoid indirectly measuring reading comprehension skills. It was organized into three parts: forms of animal abuse, consequences, and what to do in case of animal abuse. The test proctor delivers a copy of the infographic to each participant and orchestrates a brief conversation, giving students two turns to talk about the content of the infographic. Then, the infographics are taken away from the participants. Only then do students start to write. In the control group, only the prompt and the response sheet are provided, omitting any procedure for

Fig. 2. Model of brochure for activating genre knowledge.

Fig. 3. Infographic for activating topic knowledge.

activating prior knowledge.

4.3. Rubrics

Students' texts were evaluated using a 5-indicator rubric, which assigned scores that fluctuated from 0 to 2 or 0–1 according to the scales used for each indicator. The evaluated indicators included key elements of written quality that did not directly assess text

content (topic knowledge) or textual structure (genre knowledge) as follows:

- Communicative purpose: Assessing whether the texts fulfill the purpose of informing and preventing animal abuse (scale 0–2).
- Connection: Assessing elements of text cohesion and grammatical links between propositions (scale 0–2).
- Progression: Assessing the development of ideas in the text (scale 0–1).
- Paragraphs or other devices: Assessing the discursive organization in meaning-driven parts, whether through paragraphs, titles, rhetorical questions, or multimodal devices (scale 0–1).
- Text transcription: Assessing the correct association between sound and grapheme (scale 0–1).

As can be observed in these indicators, correctness, or the amount of information (topic knowledge) are not assessed. The specific characteristics of the brochure (genre knowledge) are not evaluated either. Overall, the rubric focuses on pragmatic and discursive aspects of the produced text, except for the dimension of text transcription. Although this latter indicator refers to a skill considered a predictor of writing quality, there is concern and anecdotal evidence that the pandemic and the prolonged interruption of in-person schooling in Chile may have affected some fundamental skills for reading and writing. Thus, in the elaboration of this test, this indicator was added to the rubric in the spirit of offering actionable information to communities that would be potential users of this development and this data was considered as a covariate in the analysis of writing performance.

4.4. Raters' design and rating procedure

We recruited 12 raters, built six random pairs of raters and assigned these to each tenth written completed tasks. Random pairs assignment design helps to mitigate possible raters' noise in the measurement process (Wind, 2022). Additionally, we included a third rater in the design, assigned systematically as a diagonal to the rater design matrix, to make the raters design matrix less sparse, and connected.

Raters were recruited from undergraduate programs in language pedagogy during April and May 2022. The team advertised the position through social media, posters, and snowball techniques. Subsequently, all raters were trained using benchmarks of written samples for each indicator. The training process was conducted by two expert researchers and comprised two 90-minute sessions. Specifically, these sessions included the analysis of benchmark written samples, reading of the rubrics, and training exercises using a teaching app.

Table 1
Covariate and experimental effects in the student's quality of writing.

	M1		M2		M3		M4		M5	
Thresholds	E(SE)	P<	E(SE)	P<	E(SE)	P<	E(SE)	P<	E(SE)	P<
Communicative purpose										
τ_{y1_1}	−3.04	***	−2.43	***	−2.02	***	−1.73	***	−1.40	**
	(0.28)		(0.37)		(0.38)		(0.49)		(0.50)	
τ_{y1_2}	0.49	*	1.10	**	1.51	***	1.79	***	2.12	***
	(0.22)		(0.35)		(0.35)		(0.48)		(0.50)	
Connection										
τ_{y2_1}	−3.22	***	−2.61	***	−2.19	***	−1.91	***	−1.63	**
	(0.26)		(0.36)		(0.37)		(0.50)		(0.52)	
τ_{y2_2}	−0.15		0.46		0.87	**	1.15	*	1.47	**
	(0.19)		(0.31)		(0.31)		(0.46)		(0.49)	
Progression										
τ_{y3_1}	−0.66	**	−0.04		0.36		0.65		0.97	
	(0.21)		(0.33)		(0.33)		(0.47)		(0.50)	
Paragraphs and devices										
τ_{y4_1}	0.55	*	1.17	**	1.57	***	1.85	***	2.19	***
	(0.22)		(0.34)		(0.33)		(0.47)		(0.50)	
Fixed effects										
Text transcriptions (δ)			0.90	*	0.85	*	0.88	*	0.86	*
			(0.37)		(0.35)		(0.36)		(0.36)	
Activation (γ)					0.91	**	0.76	*	0.73	*
					(0.33)		(0.34)		(0.34)	
Variance components										
Raters' variance (σ_j^2)	0.48	**	0.48	**	0.48	**	0.48	**	0.00	
	(0.17)		(0.17)		(0.17)		(0.17)		(0.00)	
Examinee variance (σ_p^2)	3.90	***	3.72	***	3.52	***	3.35	***	3.47	***
	(0.59)		(0.56)		(0.54)		(0.52)		(0.53)	
Intra class correlation	0.89	***	0.88	***	0.88	***	0.87	***	1.00	***
	(0.04)		(0.04)		(0.04)		(0.04)		(0.00)	

Note: E = unstandardized estimates in logit scale, SE = standard error of estimates are displayed in parenthesis in the row below each estimate, P < = *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Models M4 include school fixed effects (not shown), and M5 includes schools and raters fixed effects (not shown).

Lastly, three rating sessions were conducted in person, overseen by one of the researchers. Each rater utilized a computer with a platform specifically designed for this task. Additionally, in each session, the researcher in charge reminded the raters the rubric indicators, the communicative purpose of the writing task, and the essential aspects of the rating platform used.

4.5. Data analysis

We built a multivariate response matrix where responses from each rater to each indicator were separated into columns, and all three respective raters were nested on students. We specified multilevel one parameter logistic graded response model (raters as level 1, and students as level 2) (Wang & Wang, 2020). The fitted model allows to have examinees and raters' responses in the same scale, provides specific item parameters for each indicator (thresholds), and allows to include covariate effects.

We fitted a series of five models. The first is the measurement model, where the raters nested in the examinees are modeled. The first model allows us to describe the proportion of the total variance of the raters accounted by the writing ability of the examinees compared to the proportion of variance due to the differences in correctness between raters. The second model includes the students' text transcription as a control variable (0 = more than two errors spotted by at least two of three raters; 1 = text presents only one or no transcription errors according to two out of three raters). With this second model, we were able to assess whether the presence of text transcription errors is related to the probability of showing higher writing ability and then be able to control for this effect in the following models. The third model directly answers the question of interest of the study by introducing the experimental condition as a between examinee covariate (control = 0, activation = 1). With this, we inquired if scaffolding of genre and topic knowledge impact student writing quality. The fourth and fifth models introduce fixed effects of examinee schools' membership and raters, respectively. These last two models aim to assess whether the experimental result remains robust despite controlling for these fixed effects. All models were fitted using the Mplus 8.10 software (Muthén & Muthén, 2017) using robust maximum likelihood estimator (See Annex 1 for equation).

5. Results

5.1. Fitted models

In Table 1, we include the unstandardized logit estimates of the fitted models, including a measurement model (M1), a model including only the text transcription covariate effects (M2), then a model adding the experimental effect of activation (M3), and two additional models adding school fixed effects (M4), and raters fixed effects (M5).

5.2. Variance components

The variance components of the measurement model (M1) are of.48 (SE = 0.17, p <.01) for rates within examinees and 3.90 (SE = 0.59, p <.001) for examinee written ability. Consequently, the examinees' written ability accounts for 89 % of the variance of the rates. As such, a smaller portion of the variance of the observed responses is due to differences among raters' responses.

5.3. Main effects

Text transcription is positively associated with students written ability (M2: E = 0.90, SE = 0.37, p <.05, OR=2.46; LRT (1) = 6.20 p < 0.05, $R^2_{between}$ =.097) with an odd ratio of $OR_{transcription}$ =2.46. This OR indicates that the odds of expressing a more complex writing skill increase 2.46 times in students whose texts present only one or less transcription error according to two out of three raters in contrast to the other category of this variable, i.e., more than two errors spotted by at least two of three raters. The experimental effect of activation is also positive (M3: E = 0.91, SE = 0.33, p <.01, OR= 2.48; LRT (1) = 7.74 p < 0.01, $R^2_{between}$ =.141). Thus, students in the experimental condition, that is, students exposed to the use of two instruments for prior knowledge activation on genre and topic, present 2.48 higher chances of demonstrating higher written abilities under the study rubric compared to their peers in the control condition.

The results of models fourth and fifth show that results are robust when controlled for school and raters' fixed effects. When the students' school membership is controlled through a school's fixed effect, the results show a slightly smaller effect of activation for the experimental group (M4: E = 0.76, SE = 0.34, p <.05, OR = 2.14). However, the chances of demonstrating a higher written ability are still 2.14 times higher for the group that received the knowledge activation over the control group. The same occurs when the raters' fixed effect is added to the fourth model (M5: E = 0.73, SE = 0.34, p <.05, OR=2.08) with 2.08 more chances for the activation group of present higher written ability in contrast to the control group. Finally, in terms of standardized effects, the proportion of variance of the experimental effect is of.05, compared to the unaccounted variance of the model, which can be considered a small, standardized effect (see Lorah, 2018, p5).

6. Discussion

In this article, the aim was twofold: first, to introduce a design for a writing test intended for application in large-scale assessment of writing, and second, to experimentally examine the effects of employing a device for activating prior knowledge as a means of controlling construct-irrelevant variance. The overall design indeed addresses some principles of justice-oriented standardized testing as

posited by Stein (2016), such as being evidence-based and formative, that is, "tests should enable customization and scaffolding and be learning experiences in themselves" (Stein, 2016, p.204).

Indeed, in formulating the test design, themes were identified through a survey conducted among students. Subsequently, an authentic, situated-within-a-context writing task was devised, offering students a communicative purpose and a defined audience. Moreover, two devices were utilized for the cognitive activation of topic and genre knowledge: an infographic and a genre model. The main indicators of the rubric were geared toward assessing pragmatic aspects of written composition, such as meeting the communicative purpose or ensuring text cohesion, as opposed to normative or formal elements of writing. Consequently, this instrument is posited as an example of a socio-cognitive, situated writing test that additionally endeavors to control sources of construct-irrelevant variance, such as prior knowledge relevant to writing, distinct from writing ability, and susceptible to be influenced by students' background, potentially hindering the performance of those with less access to that knowledge. This scaffolding has implications to educational justice in large-scale writing assessment, as discussed below.

The results of the main effect of this study demonstrate that the odds of presenting a good writing ability are higher for students who were part of the experimental group, even when controlling for the text transcription ability, considered a predictor of writing. By showing that activation is relevant to writing outcomes, we suggest that omitting it would cause unwanted variance. In other words, two students with the same writing ability may score differently just because one of them is more familiar with the genre or topic, a difference that may be rooted in structural imbalances. In addition to the quality of text transcription used as a covariate, the rubrics considered aspects such as communicative purpose, connection (understood as textual cohesion), progression of ideas, and the use of paragraphs or other forms of thematic delimitation. These are considered fundamental aspects of writing ability that are different from content or genre development, operationalized as knowledge needed for, but different from writing ability.

Thus, the results support the use of prior knowledge activation devices in standardized writing tests as a way to control construct-irrelevant variance and enhance the validity of the test. It is important, therefore, to allow for the use of these devices in writing tests, given that prior knowledge may be influenced by students' past experiences, potentially impacting fairness. Fairness, understood as "a full opportunity to demonstrate their standing on the construct being measured" (American Educational Research Association et al., 2014, p.52), is crucial for validity, as construct-irrelevant variance can be seen as a threat to fairness (American Educational Research Association et al., 2014; Poe & Elliot, 2019). Therefore, further studies on the specific impact of these types of activation devices on fairness are needed.

Lastly, the main results of this study also suggest the value of studying the activation of prior knowledge that is needed for but different from the skill measured in the development of standardized tests assessing constructs aside from writing. Such developments may contribute to tests designed to control unwanted variance in other disciplines.

A second element worth discussing is the effect of text transcription. As mentioned earlier, in developing this test, we aimed to study this variable in written performance due to concerns at the national level about the impact of the pandemic and the interruption of in-person schooling on students, especially in the examined age group, which received part of their initial literacy instruction remotely. Consistently with previous literature (Berninger et al., 1992), a positive effect was found between text transcription and writing ability. This variable was introduced as a covariate, and the effect of knowledge activation remained positive even after controlling for text transcription. Furthermore, it is noteworthy that this variable, being of the control type, retained its significance in all the models run. Therefore, it is advisable to continue including this control variable in future studies and to explore its interaction with both the rubric indicators and other relevant sociodemographic aspects.

In addition to motivation, topic knowledge, or genre knowledge, another aspect that can introduce unwanted variance in the measured score is the rating process. As for motivation, we decided to indirectly control it by design, using data from a survey to fifth graders and employing the most chosen prompt that was selected by students regardless of their sex or the kind of schools they attended in terms of previous school's performance in tests and the school's vulnerability index. Regarding the rating process, we designed a three-rater procedure of scoring and examined the partition of variance in model 1 (M1). The proportion of variance due to differences among students' writing rates is 89 %, indicating excellent reliability, according to Cicchetti (1994). Using multiple raters for score assignment in a large-scale test is common (Engelhard, 2012; Lumley, 2002; Smith & Paige, 2019a, 2019b), but this process is not always included in the measurement model. This omission could imply a lack of awareness regarding the magnitude of this unwanted variance or the possibility of estimating the effects of the variables of interest with reduced accuracy (Guo & Wind, 2021) consequently affecting the validity of these scores. This is why including rater modeling in future research on measuring complex skills such as writing is necessary.

One of the limitations of the present study might be the raters design, relying on random raters' pairs assignment, with an additional third rater. Sparse raters' designs instead of fully-crossed raters design have an impact on examinee score estimates ordering and precision when using disconnected raters' blocks (Wind & Stager, 2019). In the present study, we systematically included a third rater to each rater pair, so each rater had shared examinees with each rater pair. The aim of this design was to diminish the disconnection between raters in the raters' design. Moreover, apart from fitting a main effect model (M3), we fitted models where we include schools and raters as fixed effects (M4 and M5), to be sure our main results were robust to schools' membership and raters' assignment. The obtained results of the present experiment were indeed robust when controlling by raters and schools fixed effects.

Overall, the results of this study reaffirm that topic and genre knowledge are forms of previous knowledge related to writing quality (McCutchen, 2011; Olinghouse & Graham, 2009; Olinghouse et al., 2015; Graham et al., 2019) and thus, they must be considered when designing a standardized writing test. Additionally, the findings demonstrate that it is possible to mitigate their influence in introducing construct-irrelevant variance to writing assessment through scaffolding, ensuring fairness and validity in the process. This hold implications for the development of large-scale tests guided by principles of educational justice.

Funding

This work was funded by grant ID21I10056, from the National Agency of Research and Development, ANID, Chile.

CRediT authorship contribution statement

Carolina Castillo: Writing – review & editing, Methodology, Investigation, Formal analysis. **María Jesús Espinosa:** Writing – review & editing, Investigation, Conceptualization. **Javiera Figueroa:** Writing – review & editing, Investigation, Conceptualization. **Diego Carrasco:** Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Rosario Escribano:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Natalia Ávila Reyes:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used DeepL and Chat GPT as translation aids. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of Competing Interest

The authors have no conflicts of interest to disclose.

Appendix. : Equation for the main effect model

$$\mathbf{logit}[\mathbf{Pr}(\boldsymbol{y} > \boldsymbol{k})] = -\boldsymbol{\tau}_{\boldsymbol{yk}} + \boldsymbol{\theta}_{\boldsymbol{p}} + \boldsymbol{\lambda}_{\boldsymbol{j}} + \boldsymbol{\delta}\boldsymbol{\omega}_{\boldsymbol{p}} + \boldsymbol{\gamma}\boldsymbol{x}_{\boldsymbol{p}}$$

Where,
$\boldsymbol{Pr}(\boldsymbol{y} > \boldsymbol{k})$= probability of examinee $\boldsymbol{p}$, reciving a rating of $\boldsymbol{k}$, on indicator $\boldsymbol{y}$, from rater $\boldsymbol{j}$
$\boldsymbol{\tau}_{\boldsymbol{yk}}$ = Threshold of indicator $\boldsymbol{y}$ for the cumulative probability change between $\boldsymbol{k}$ categories
$\boldsymbol{\theta}_{\boldsymbol{p}}$ = written ability of examinee $\boldsymbol{p}$
$\boldsymbol{\lambda}_{\boldsymbol{j}}$ = leniency of rater $\boldsymbol{j}$
$\boldsymbol{\delta}$ = overall relationship between text transcription ability $\boldsymbol{\omega}_{\boldsymbol{p}}$ and written ability
$\boldsymbol{\gamma}$ = experimental effect of topic and genre knowledge $\boldsymbol{x}_{\boldsymbol{p}}$ on examinee written ability

Data availability

Data will be made available on request.

References

Aitken, A. A., Graham, S., & McNeish, D. (2022). The effects of choice versus preference on writing and the mediating role of perceived competence. *Journal of Educational Psychology, 114*(8), 1844. https://doi.org/10.1037/edu0000765

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

Bazerman, C. (2013). Understanding the lifelong journey of writing development. *Infancia York Aprendizaje, 36*(4), 421–441. https://doi.org/10.1174/021037013808200320

Bazerman, C., Bonini, A., & Figueiredo, D. (Eds.). (2009). *Genre in a changing world*. Parlor Press.

Beck, S., Jones, K., Storm, S., & Smith, H. (2020). Scaffolding Students' Writing Processes Through Dialogic Assessment. *Journal of Adolescent Adult Literacy, 63*(6), 651–660. https://doi.org/10.1002/jaal.1039

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Lawrence Erlbaum Associates.

Berninger, V., Yates, C., Cartwright, A., Rutberg, J., Remy, E., & Abbott, R. (1992). Lower-level developmental skills in beginning writing. *Reading and Writing, 4*(3), 257–280. https://doi.org/10.1007/BF01027151

Boscolo, P. (2009). Writing in primary school. In C. Bazerman (Ed.), *Handbook of research on writing* (pp. 359–379). Lawrence Erlbaum Associates.

Broad, B. (2000). Pulling your hair out: Crises of standardization in communal writing assessment. *Research in the Teaching of English, 35*(2), 213–260.

Castillo, C. & Ávila Reyes, N. (In Review). Students' sociodemographic characteristics and writing performance: A systematic literature review.

Chamorro, M. (2022). Cognitive validity evidence of computer- and paper-based writing tests and differences in the impact on EFL test-takers in classroom assessment. *Assessing Writing, 51*. https://doi.org/10.1016/j.asw.2021.100594

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284. https://doi.org/10.1037/1040-3590.6.4.284

Collins, P., Tate, T. P., Won Lee, J., Krishnan, J. A., & Warschauer, M. (2021). A multi-dimensional examination of adolescent writing: Considering the writer, genre and task demands. *Reading and Writing, 34*(8), 2151–2173. https://doi.org/10.1007/s11145-021-10140-x

Corrigan, J. A., & Slomp, D. H. (2021). Articulating a Sociocognitive Construct of Writing Expertise for the Digital Age. *The Journal of Writing Analytics, 5*(1), 142–195. https://doi.org/10.37514/JWA-J.2021.5.1.05

Cox, B. E., Shanahan, T., & Tinzmann, M. B. (1991). Children's knowledge of organization, cohesion, and voice in written exposition. *Research in the Teaching of English*, 179–218.

Cushing Weigle, S. (2002). *Assessing writing.* Cambridge University Press. https://doi.org/10.1017/CBO9780511732997
Driscoll, D. L., Paszek, J., Gorzelsky, G., Hayes, C. L., & Jones, E. (2020). Genre Knowledge and Writing Development: Results From the Writing Transfer Project. *Written Communication, 37*(1), 69–103. https://doi.org/10.1177/0741088319882313
Duke, N. K., Purcell-Gates, V., Hall, L. A., & Tower, C. (2006). Authentic Literacy Activities for Developing Comprehension and Writing. *The Reading Teacher, 60*(4), 344–355. https://doi.org/10.1598/RT.60.4.4
Engelhard, G. (2012). Monitoring raters in performance assessments. In G. Tindal, & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students* (pp. 224–249). Routledge.
Fitzgerald, J., & Teasley, A. B. (1986). Effects of instruction in narrative structure on children's writing. *Journal of Educational Psychology, 78*(6), 424.
Ghanbari, N. (2019). Promoting Fairness in EFL Writing Assessment: Are There Any Effects of the Writers' Awareness of the Rating Criteria? *Journal of Asia TEFL, 16*(2), 735–742. https://doi.org/10.18823/asiatefl.2019.16.2.22.735
Graham, S. (2018). A writer(s)-within-community model of writing. In C. Bazerman, A. N. Applebee, V. W. Berninger, D. Brandt, S. Graham, J. Jeffery, P. K. Matsuda, S. Murphy, D. W. Rowe, M. Schleppegrell, & K. C. Wilcox (Eds.), *The Lifespan Development of Writing* (pp. 272–325). National Council of Teachers of English.
Graham, S., Harris, K., Kihuara, S., & Fishman, E. (2017). The roles of writing knowledge, motivation, strategic behaviors, and skills in predicting elementary students' persuasive writing from source material. *The Elementary School Journal, 118*(1), 82–104. https://doi.org/10.1086/693009
Graham, S., Wijekumar, K., Harris, K. R., Lei, P.-W., Fishman, E., Ray, A. B., & Houston, J. (2019). Writing Skills, Knowledge, Motivation, and Strategic Behavior Predict Students' Persuasive Writing Performance in the Context of Robust Writing Instruction. *The Elementary School Journal, 119*(3), 487–510. https://doi.org/10.1086/701720
Guo, W., & Wind, S. A. (2021). Examining the impacts of ignoring rater effects in mixed-format tests. *Journal of Educational Measurement, 58*(3), 364–387. https://doi.org/10.1111/jedm.12292
Hammond, J. (2017). Social Justice and Educational Measurement: John Rawls, the History of Testing, and the Future of Education. *Assessing Writing, 33*, 68–70. https://doi.org/10.1016/j.asw.2017.06.002
Hidi, S. E., & Hildyard, A. (1983). The comparison of oral and written productions in two discourse types. *Discourse Processes, 6*(2), 91–105. https://doi.org/10.1080/01638538309544557
Inoue, A. B. (2015). *Antiracist Writing Assessment Ecologies: Teaching and Assessing Writing for a Socially Just Future.* The WAC Clearinghouse & Parlor Press.
Kellogg, R. T. (1987). Effects of topic knowledge on the allocation of processing time and cognitive effort to writing processes. *Memory Cognition, 15*(3), 256–266. https://doi.org/10.3758/BF03197724
Knoch, U., & Elder, C. (2010). Validity and fairness implications of varying time conditions on a diagnostic test of academic English writing proficiency. *SYSTEM, 38*(1), 63–74. https://doi.org/10.1016/j.system.2009.12.006
Langer, J. A. (1986). *Children reading and writing: Structures and strategies.* Norwood, NJ: Ablex.
Lorah, J. (2018). Effect size measures for multilevel models: definition, interpretation, and TIMSS example. *Large-Scale. Assessments in Education, 6*(1). https://doi.org/10.1186/s40536-018-0061-2
Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing, 19*(3), 246–276. https://doi.org/10.1191/0265532202lt230oa
Manzi, J., García, M. R., & Taut, S. (2019). *Validez de evaluaciones educacionales en Chile y Latinoamérica.* Ediciones UC.
McCutchen, D. (1986). Domain knowledge and linguistic knowledge in the development of writing ability. *Journal of Memory and Language, 25*, 431–444.
McCutchen, D. (1987). Children's discourse skill: Form and modality requirements of schooled writing. *Discourse Processes, 10*(3), 267–286. https://doi.org/10.1080/01638538709544676
McCutchen, D. (2011). From novice to expert: Implications of language skills and writing-relevant knowledge for memory during the development of writing skill. *Journal of Research on Writing, 3*(1), 51–68. https://doi.org/10.17239/jowr-2011.03.01.3
McCutchen, D., Francis, M., & Kerr, S. (1997). Revising for meaning: Effects of knowledge and strategy. *Journal of Educational Psychology, 89*(4), 667.
Muthén, L.K., & Muthén, B.O. (2017). *Mplus User's Guide* (8th ed.). Muthén & Muthén.
Olinghouse, N. G., & Graham, S. (2009). The relationship between the discourse knowledge and the writing performance of elementary-grade students. *Journal of Educational Psychology, 101*(1), 37–50. https://doi.org/10.1037/a0013462
Olinghouse, N. G., Graham, S., & Gillespie, A. (2015). The relationship of discourse and topic knowledge to fifth graders' writing performance. *Journal of Educational Psychology, 107*(2), 391–406. https://doi.org/10.1037/a0037549
Perelman, L. (2018). *Towards a new NAPLAN: Testing to the teaching.* NSW Teachers Federation. ⟨https://www.nswtf.org.au/files/appendices_naplan.pdf⟩.
Poe, M., & Elliot, N. (2019). Evidence of fairness: Twenty-five years of research in Assessing Writing. *Assessing Writing, 42*, 1–21. https://doi.org/10.1016/j.asw.2019.100418
Poe, M., & Inoue, A. B. (2016). Toward writing assessment as social justice: An idea whose time has come. *College English, 79*(2), 119–126.
Proske, A., & Kapp, F. (2013). Fostering topic knowledge: Essential for academic writing. *Reading and Writing, 26*(8), 1337–1352. https://doi.org/10.1007/s11145-012-9421-4
Purcell-Gates, V., Duke, N. K., & Martineau, J. A. (2007). Learning to read and write genre- specific text: Roles of authentic experience and explicit teaching. *Reading Research Quarterly International Reading Association, 42*(1), 8–45. https://doi.org/10.1598/RRQ.42.1.1
Randall, J., Poe, M., Slomp, D., & Oliveri, M. E. (2024). Our validity looks like justice. Does yours? *Language Testing, 41*(1), 203–219. https://doi.org/10.1177/02655322231202947
Reed, D., Binning, K., Jemison, E., & DeSalle, N. (2023). High-Quality Formative Writing Assessment for Middle School Students in Tier 2 Literacy Interventions. *Learning Disabilities Research Practice, 38*(1), 70–79. https://doi.org/10.1111/ldrp.12300
Sims, M. (2023). Shifting perceptions of socially just writing assessment: Labor-based contract grading and multilingual writing instruction. *Assessing Writing, 57*. https://doi.org/10.1016/j.asw.2023.100731
Smith, Grant S., & Paige, David D. (2019s). A Study of Reliability Across Multiple Raters When Using the NAEP and MDFS Rubrics to Measure Oral Reading Fluency. *Reading Psychology, 40*(1), 34–69. https://doi.org/10.1080/02702711.2018.1555361
Smith, G. S., & Paige, D. D. (2019b). A study of reliability across multiple raters when using the NAEP and MDFS rubrics to measure oral reading fluency. *Reading Psychology, 40*(1), 34–69. https://doi.org/10.1080/02702711.2018.1555361
Stein, Z. (2016). *Social justice and educational measurement: John Rawls, the history of testing, and the future of education.* Routledge.
Tabari, M. A., Bui, G., & Wang, Y. (2021). The effects of topic familiarity on emotionality and linguistic complexity in EAP writing. *Language Teaching Research, 0*(0), 1–19. https://doi.org/10.1177/13621688211033565
Tolchinsky, L. (2008). Usar la lengua escrita en la escuela. *Revista Iberoamericana Délelőtt EducaciónN, 46*, 37–54. ⟨http://hdl.handle.net/11162/23440⟩.
Wang, H., & Troia, G. A. (2023). Writing Quality Predictive Modeling: Integrating Register-Related Factors. *Written Communication, 40*(4), 1070–1112. https://doi.org/10.1177/07410883231185287
Wang, J., & Wang, X. (2020). *Structural Equation Modeling: Applications Using Mplus* (2nd ed.). John Wiley & Sons, Inc.
Wijekumar, K., Graham, S., Harris, K. R., Lei, P.-W., Barkel, A., Aitken, A., Ray, A., & Houston, J. (2019). The roles of writing knowledge, motivation, strategic behaviors, and skills in predicting elementary students' persuasive writing from source material. *Reading and Writing, 32*(6), 1431–1457. https://doi.org/10.1007/s11145-018-9836-7
Wind, S. A. (2022). Rater Connections and the Detection of Bias in Performance Assessment. *Measurement, 20*(2), 91–106. https://doi.org/10.1080/15366367.2021.1942672
Wind, S. A., & Stager, C. G. (2019). The impacts of characteristics of disconnected subsets on group anchoring in incomplete rater-mediated assessment networks. *Psychological Test and Assessment Modeling, 61*(1), 13–36.

Natalia Ávila Reyes: She is an associate professor at Pontificia Universidad Católica, Chile. Her research centers on writing across the lifespan and social justice in teaching, learning, and the assessment of writing. She holds a MA in Linguistics from UC Chile and a PhD in Education from the University of California, Santa Barbara.

Diego Carrasco: He is an assistant research professor at Centro de Medición MIDE UC at Pontificia Universidad Católica de Chile. His research focuses on methodological challenges in the study of nested observations, including measurement and inferential problems in large scale assessment. He holds a PhD in Psychology from the University of Sussex.

Rosario Escribano: She is an assistant professor at Pontificia Universidad Católica de Chile. Her research has focused on large-scale learning assessment and the improvement of learning linked to information obtained from evaluations. Her interests also include the analysis of teaching practices and classroom interactions, with educational justice as a cross-cutting aspect.

María Jesús Espinosa: PhD in Education from Diego Portales and Alberto Hurtado Universities in Chile. She specializes in language teaching, literacy, and writing, with a focus on teacher education. She currently serves as an assistant professor at the Faculty of Education and Language Coordinator at Diego Portales University.

Javiera Figueroa: She is an assistant professor at the Faculty of Education at Alberto Hurtado University in Chile. She holds a Doctorate in Education from the Pontificia Universidad Católica de Chile. Her main research areas include writing in the school context, language development, and teacher education.

Carolina Castillo: She is a PhD candidate in Education at Pontificia Universidad Católica de Chile. She is an Anthropologist and holds a master's degree in Sociology. Her research interests are linked to the assessment of writing at different stages of schooling, validity, and fairness in educational measurements, and standardized assessment.