Unleashing think-aloud data double function to investigate self-assessment: quantitative and qualitative approaches

Ernesto Panadero[1], Leire Pinedo[2] & Javier Fernández-Ruiz[3]

[1] Centre for Assessment Research Policy and Practice in Education (CARPE), School of Policy and Practice, Institute of Education, St. Patrick's Campus, Dublin City University, Ireland.
[2] Facultad de Educación y Deporte, Universidad de Deusto, Bilbao, Spain.
[3] Departamento de Psicología, Sociología y Filosofía, Universidad de León, Spain.

Ernesto Panadero https://orcid.org/0000-0003-0859-3616

Leire Pinedo http://orcid.org/0000-0002-3046-5226

Javier Fernández Ruiz https://orcid.org/0000-0001-5419-7687

Recommended citation: Panadero, E., Pinedo, L., & Fernández Ruiz, J. (2025). Unleashing think-aloud data to investigate self-assessment: Quantitative and qualitative approaches. *Learning and Instruction*, 95, 102031. https://doi.org/10.1016/j.learninstruc.2024.102031

Funding: (1) Spanish National R+D call from the Ministerio de Ciencia, Innovación y Universidades (Generación del conocimiento 2020), Reference number: PID2019-108982GB-I00. (2) Basque Government Call for Grants to support the activities of research groups of the Basque University System (2022-2025) project reference IT1624-22.

Correspondence concerning this article should be addressed to Ernesto Panadero, Dublin City University St. Patrick Campus (Drumcondra) Institute of Education C112 Main building Dublin 9 D09 DY00 Ireland. E-mail: ernesto.research@gmail.com

The authors declare that they have no conflicts of interest regarding this manuscript. This research has been approved by the corresponding ethical committees.

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

During the preparation of this manuscript, the authors utilized ChatGPT for proofreading and simplifying complex sentences. Following the use of this tool, the authors thoroughly reviewed and edited the content as necessary. The authors assume full responsibility for the content of this publication.

Abstract

Background. Think-aloud, a method that allows access to students' cognitive, emotional, and motivational processes, has been highlighted as a powerful research tool.

Aim. This study aims to examine the use of think-aloud in a self-assessment research project, seeking insights into students' actions during self-assessment and illustrating its potential for both qualitative and quantitative research in educational assessment. We introduce three empirical studies from two data collections, utilizing both quantitative and qualitative methods to investigate the 'black box' of self-assessment.

Sample. A total of 67 secondary education students and 126 university students participated in two data collections.

Methods. Participants were instructed to think-aloud concurrently while self-assessing academic tasks they had previously performed. Think-aloud protocols were then analyzed and coded based on two types of data/content: (1) self-assessment strategies and criteria, and (2) processes. For the first type, a quantitative approach was used, while for the self-assessment processes, a qualitative approach was adopted.

Results. This approach resulted in a more robust categorization of the actions undertaken by students when self-assessing, yielding a general/specific model of self-assessment, and providing empirical results regarding the effects of feedback, year and educational level, and gender.

Conclusion. Our studies demonstrate how think-aloud can be effectively utilized through complementary quantitative and qualitative methods to provide comprehensive insights into the studied phenomena.

Keywords: thinking aloud protocols, think aloud, quantitative, qualitative, self-assessment, feedback.

Unleashing think-aloud data double function to investigate self-assessment: quantitative and qualitative approaches

1. Theoretical framework

For a number of years, there has been a drive to integrate process data into educational research, notably in the self-regulated learning (SRL) field. A pioneering publication that steered SRL research in this direction was by Winne and Perry (2000), who contrasted a more static vision of SRL - as an aptitude - with a dynamic one - as an event. For the latter, data about the specific process were needed, extending beyond self-reported scales or questionnaires. During this period, think-aloud emerged as a powerful SRL research method (Boekaerts & Corno, 2005). A key aspect of think-aloud is its ability to combine the students' perspective with a more "objective" approach to the investigated phenomena (Ericsson & Fox, 2011; McIntyre et al., 2021). While think-aloud has gained attention since then, it remains less prominent as a major research method. However, there is a growing interest in its use in combination with other process data methods (e.g., eye-tracking), as it offers crucial insights into students' processes during learning (Holmqvist et al., 2011). Significantly, think-aloud provides valuable opportunities for both quantitative and qualitative research, though it is rare to see projects that report both approaches. In this paper, we introduce a self-assessment research project utilizing think-aloud to unveil the 'black box' of self-assessment. Utilizing this project's results, our aim is to illustrate how think-aloud data can be effectively utilized in both qualitative and quantitative approaches, enhancing the value of think-aloud data.

Educational assessment, particularly when targeted for formative purposes, is a crucial factor in enhancing students' academic performance (Endowment Educational Foundation, 2018). Notably, research in educational assessment often relies on traditional

methods, not incorporating process data that is crucial to understanding how students process assessment information (Panadero, 2023; Winstone & Nash, 2023). Self-assessment is one area that could greatly benefit from process data, providing insights into the actions students undertake during their self-assessment (Yan & Brown, 2017). In this regard, think-aloud can provide access to students' cognitive, emotional, and motivational processes, assuming certain conditions are met. In this paper, we discuss a self-assessment research project that employed think-aloud to explore the black-box of self-assessment. By analyzing these project results, our aim is to demonstrate how think-aloud data can be applied in both qualitative and quantitative research, thereby amplifying the value of think-aloud data.

1.1. Think-aloud as process data

As just presented, process data has become mainstream in educational research especially in the last 15 years due to the irruption in popularity of measurement technologies such as eye-tracking, electrodermal activity, facial recognition or extensive videorecording (Azevedo et al., 2018; Panadero, 2023). In the context of SRL, process data has emerged as a significant disruptor in how researchers conceptualize their studies (Saint et al., 2022). Notably, SRL has historically been at the forefront of employing innovative measurement techniques compared to other psychological disciplines (Karoly et al., 2005). This trend was vividly illustrated by Boekaerts and Corno (2005), who enumerated various methods to measure self-regulated learning. Among these methods, think-aloud stands out as a method that bridges the perspective of the student with a more “objective” approach to the phenomena (Ericsson & Fox, 2011; McIntyre et al., 2021).

Think-aloud serves both as a research method and a type of data collection, wherein participants articulate the processes they are experiencing. Typically, educational

researchers encourage participants to vocalize their thoughts and emotions while engaged in an academic task. In some studies, participants are trained in think-aloud techniques, for instance, by being asked to count the windows in their home, narrating their mental journey through the house.

Think-aloud has a long history as a data collection method, but its use has been the subject of debate. Centuries ago, William James (1890) viewed introspection as a primary tool in psychological research, yet acknowledged its challenges and fallibility. Subsequent advances in cognitive psychology during the 50s and 60s validated the study of participants' performance through verbalized solutions, enabling inferences about cognitive constructs (Fox et al., 2011). Consequently, in recent decades, think-aloud has been recognized as a valuable tool for representing students' internal states (Ericsson & Simon, 1980).

A major challenge for think-aloud research has been ensuring that verbalizations accurately reflect participants' thought processes. Concerns were raised that think-aloud might influence cognitive processes, casting doubt on its validity. However, a meta-analysis involving nearly 3,500 participants found the "think-aloud" effect size to be negligible (Fox et al., 2011). Notably, when participants had to recode information for reporting, think-aloud was observed to enhance performance compared to silent conditions. This finding underscores the importance of researchers clearly delineating their instructions and procedures in think-aloud studies to accurately assess its impact on cognitive processes.**1.2. Self-assessment: the perfect arena to employ think-aloud**

Self-assessment "involves a wide variety of mechanisms and techniques through which students describe (i.e., assess) and possibly assign merit or worth to (i.e., evaluate) the qualities of their own learning processes and products" (Panadero et al., 2016 p. 804).

There is strong evidence that self-assessment contributes to students' academic performance, self-regulated learning, and a number of motivational variables (Sitzmann et al., 2010; Yan et al., 2023). Self-assessment is a powerful learning and instructional strategy, as it enables students to recognize the quality of their work through evaluation and reflection. Consequently, self-assessment has emerged as an extensively researched area, particularly since the latter decades of the previous century (e.g., Falchikov & Boud, 1989).

Crucially, self-assessment primarily occurs as an internal process, and our educational efforts should aim to externalize this implicit process (Panadero et al., 2019). Given that self-assessment largely transpires within the student's mind, employing activities that yield 'tangible' elements for reflection, such as written self-assessments, is vital. This internal nature complicates research in this domain, a challenge often described as 'opening the black box of self-assessment' (Yan & Brown, 2017). This complexity arises because self-assessment is frequently viewed as a standalone strategy, without adequately dissecting the specific actions comprising it (Panadero et al., 2016).

Moreover, in considering methods to explore this internal process, think-aloud emerges as a particularly effective approach. This effectiveness is twofold: (1) think-aloud seeks to reveal participants' inner thoughts, motivations, and emotions (Fox et al., 2011), and (2) concurrent think-aloud facilitates real-time measurement. Despite its efficacy, think-aloud remains underutilized in self-assessment research, although its use is increasing (e.g., Rickey et al., 2023). However, our group has consistently employed this method from the past decade to the present, establishing it as a cornerstone in our multimodal data approach. Through three of our studies, we will demonstrate its potential."

1.3. Quantitative vs qualitative research: approaches in think-aloud

Think-aloud is inherently a qualitative data collection method due to its rich and evolving nature, aiming to elucidate participants' inner processes (Shapiro, 2014). Yet, its analysis can be approached from both quantitative and qualitative perspectives, especially in mixed-methods research (Creswell & Plano-Clark, 2018). As Charters (2003) explains, both approaches are valuable but yield different insights: the quantitative approach extracts numerical patterns from the codes to generalize participants' experiences, while the qualitative approach provides a naturalistic perspective, detailing individual experiences. In our research, we have also used the quantitative approach to extract those numerical patterns (Panadero et al., 2020, 2022) and the qualitative approach to create general profiles of student self-assessment methods (Panadero et al., under review), as will be discussed later.

Notably, studies in self-regulated learning and educational assessment predominantly employ a quantitative approach when analyzing think-aloud data. The typical procedure involves coding the processes utilized by students during a task and then quantifying them through frequencies, means, or other statistical measures. This trend is evident in SRL research, where quantitative analyses such as chi-square tests and logistic regression models are commonly used (Azevedo et al., 2004; Azevedo et al., 2008; Greene & Azevedo, 2009). Similarly, in educational assessment, studies often focus on the frequencies of comments during self-assessment or use ANOVAs to compare mental operations during feedback processing (Kostons et al., 2009; Mañez et al., 2019). According to Panadero's (2023) categorization of multimodal research, these studies align with the second (unimodal research + performance) and third (multimodal data without triangulation) categories, respectively.

However, qualitative approaches, though less common in educational assessment, also play a significant role. For instance, Lui and Andrade (2022) combined quantitative survey data analysis with narrative descriptions of think-aloud responses, integrating both data sources to interpret survey results through participants' perspectives. Tian et al. (2022) adopted a similar strategy, initially presenting general survey findings before delving into case studies that illustrate feedback revision strategies. Such qualitative approaches involve selecting specific cases and presenting their think-aloud content for a more in-depth understanding of participants' processes, thereby augmenting the survey data. Recently, Panadero (2023) has emphasized that this method of data utilization represents an advanced form of multimodal data analysis, involving triangulation of different data sources.

In conclusion, while the majority of SRL and educational assessment studies tend to favor either a quantitative or qualitative approach for analyzing think-aloud data, there are examples where think-aloud is used to contextualize quantitative data from self-reports or surveys. Nonetheless, it is relatively rare for the same think-aloud dataset to be analyzed through both quantitative and qualitative lenses.

1.4. Aim, research questions, and think-aloud approach

Our aim is to present and discuss an innovative way of utilizing think-aloud data integrating quantitative and qualitative methods in the context of self-assessment research. Our findings are derived from three prior studies where think-aloud served as the pivotal method to examine self-assessment processes (Panadero et al., 2020, 2022, under review). In the initial study, Panadero and colleagues (2020) applied quantitative methods to analyze think-aloud data from a secondary education sample. The subsequent study (Panadero et al., 2022) also utilized quantitative approaches in think-aloud but with university students. The third study, currently under review, combines data from both collections, employing

qualitative methods to analyze think-aloud and using quantitative techniques to validate these findings. The common goal of these studies was to demystify the 'black box' of self-assessment. In this paper, we address two research questions (RQs):

RQ1. What insights emerged from the quantitative coding and analysis of think-aloud data?

RQ2. What insights emerged from the qualitative coding and analysis of think-aloud data?

Importantly, all three studies involved concurrent think-aloud while students engaged in self-assessment tasks. The innovative aspects of our research include: (1) employing dual analytic approaches (quantitative and qualitative) to gain a multifaceted understanding of the same phenomena, (2) utilizing think-aloud across a diverse age range, from 12 to over 20 years, covering both secondary and higher education, and (3) leveraging think-aloud to pioneer research into decoding the 'black box' of self-assessment. Through think-aloud data, we accessed students' self-assessment actions, identifying their primary strategies and criteria quantitatively, while qualitatively grouping students into distinct profiles. We contend that this approach leads to more robust and precise conclusions, thereby enhancing the effectiveness of educational interventions.

2. Methods

2.1. Participants

We conducted two separate data collections, one focused on secondary education and the other on higher education. In the secondary education data collection 67 students participated while the higher education data collection involved 126 university students. Secondary education participants included first-year K7 students, fourth year K10 students, and first year of university preparation K11 students. For higher education, participants were drawn from the 1st, 2nd and 3rd years of Psychology degree program. Previous research indicates that the youngest participants in our study are capable to elicit think-aloud

(Greene & Azevedo, 2009). Additionally, there is substantial evidence that higher education students can effectively think-aloud while self-assessing (Kostons et al., 2009).

Crucially, in two of our studies, the final sample size was reduced due to missing data. In Panadero and colleagues (2020) the sample comprised 64 secondary education participants. Meanwhile, in Panadero and colleagues (in review), the sample included 67 secondary education and 111 higher education participants. Table 1 summarizes the main research characteristics of these three studies.

Table 1

Summary of studies using think-aloud data.

	Panadero and colleagues (2020) Data collection 1	Panadero and colleagues (2022) Data collection 2	Panadero and colleagues (in review) Data collection 1 & 2
Final sample	64 secondary education participants from: K7: 28 K10: 22 K11: 14	126 higher education participants from: 1st year: 44 2nd year: 40 3rd year: 42	193 participants that included: - 67 secondary education participants from K7 (28), K10 (25), and K11 (14). - 111 higher education participants from 1st year (39), 2nd year (36), and 3rd year (37).
Data collection methods	- Video-recorded concurrent think-aloud - Direct observation - Questionnaires	- Video-recorded concurrent think- aloud - Direct observation - Questionnaire collected but not analyzed in this study	- Video-recorded concurrent think-aloud - Direct observation - Questionnaire collected but not analyzed in this study
Data nature	Quantitative and qualitative	Qualitative	Qualitative
Data coding	Quantitative	Quantitative	Qualitative
Data analysis	Quantitative	Quantitative	Quantitative
Results	Self-assessment strategies and criteria	Self-assessment strategies and criteria	Self-assessment processes and profiles

2.2. Research design

We employed a cross-sectional design encompassing three year levels each in secondary and higher education. For secondary education, we carried out an intra-subject study to explore the effects of subject matter (Spanish and Mathematics), feedback occasion (without and with feedback) and gender. In higher education, we conducted an experimental study to investigate the impact of three types of feedback (rubric vs.

instructor's vs. combined). The studies utilized a mixed methods approach, incorporating both qualitative and quantitative data.

2.3. Think-aloud method

We collected concurrent think-aloud as students engaged in self-assessment. Before beginning the experimental phase, secondary education participants were instructed as follows for the think-aloud process: "Imagine that you have just finished the task and still have some time to review it before handing it out to the teacher, this is, you have time to self-assess. We are interested in understanding what you do during that self-assessment process. Thus, I would like to ask you to say out loud everything that you are doing, thinking, and feeling while you self-assess". In higher education, the instructions for the think-aloud protocol were identical. However, since these participants were already familiar with self-assessment, we omitted the preamble of "you have just finished a task and you have time before handing it to the teacher".

Additionally, if any participant remained silent for more than 30 seconds, the researcher prompted them to verbalize their actions, by asking "What are you doing now?". The students were recorded using high-definition cameras for the corresponding coding of their think-aloud.

2.4. Procedure for data collection

Before the experimental phase began, participants were briefed on procedure and the think-aloud protocol. The verbal instructions required students to articulate their actions while conducting a self-assessment of an academic task (e.g., written essay or mathematical exercises). They were also informed that if they would be prompted to vocalize their thoughts if they remained silent. The experimental procedure was then divided in two phases: one without feedback and one with feedback. In the first phase, participants initially

filled a set of questionnaires (i.e., self-efficacy and emotions), and subsequently engaged in think-aloud while performing their self-assessment of the task. Immediately afterward, they filled the questionnaires again. In the second phase, participants conducted a second self-assessment of the same task, this time aided by feedback. In secondary education they received instructor's feedback, while in higher education, the feedback type varied depending on the experimental condition (rubric vs. instructor's vs. combined). The participants were again asked to verbalize their actions during the self-assessment. Lastly, they filled the questionnaires for the third time. The entire process was video-recorded, and there were no time constraints on the self-assessment.

Throughout the data collection process, we paid special attention to three specific situations. First, students could stop verbalizing during the experimental phase. As each student was accompanied and observed by a team member, the researcher could discern whether the silence resulted from external disturbances (e.g., corridor noise) or from a lack of engagement with the task. However, if a student remained silent for 30 seconds, the researcher would prompt them to articulate what they were doing at that moment.

Second, students could ask for more specific instructions about what they should do during the experimental phase. In such cases, students were remained of the written instructions provided at the beginning, and no further explicit guidance was provided to avoid biasing their cognitive processing. Third, there were cases where students inquired about how to use the feedback received (e.g., a rubric). In response, the researcher only provided a brief description of the materials. Additionally, students were informed that choosing not to use the provided feedback was also a valid approach.

2.5. Data analysis and category creation

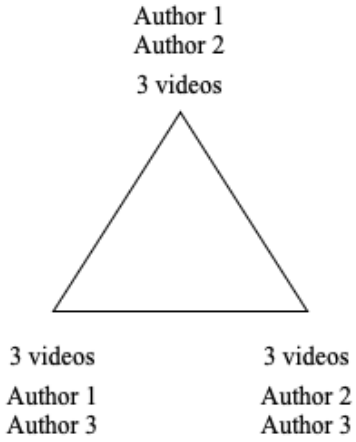
A detailed explanation of data analysis is available at this link: https://osf.io/4zpts/ As a summary, the think-aloud protocols were analyzed and coded according to two types of data/content: (1) self-assessment strategies and criteria (Panadero et al., 2020, 2022), and (2) processes (Panadero et al., under review). We will now describe each of these in detail.

First, we employed a quantitative approach to analyze the data, focusing on the presence or absence of the categories in self-assessment strategies and criteria. The first author, along with two research assistants, reviewed 15 videos to discuss the content of participants' verbalizations. We aimed to identify a set of primary categories, which were then grouped into two large blocks: self-assessment strategies and criteria. Given that participants could report multiple strategies and criteria, we utilized dummy dichotomous variables for coding. Then, based on participants' verbalizations, specific categories for strategies and criteria were defined by consensus among the researchers.

Second, in Panadero and colleagues (in review) we analyzed the data using a qualitative approach. In the first round, the three authors independently watched 6 videos, with every video being reviewed by two authors. This resulted in a total review of 9 videos as it is shown in the Figure 1. They then met to share their thoughts on the observed self-assessment behaviors, combining their notes to identify similarities and differences.

Figure 1

Example of coding distribution



After the initial round, we conducted three iterations, each involving the review of 3 videos (2 per author), and a posterior discussion. During these rounds, the authors developed a preliminary coding scheme for the self-assessment processes and created a template for subsequent coding. This included a clear definition of each process, and guidelines for qualitatively assigning profiles. The preliminary coding scheme was then applied to additional videos for testing. Specifically, the second and third authors independently coded 15 videos using the coding categories defined in the first phase. They later met again to discuss and refine certain categories (i.e., dividing some of the actions into two levels), leading to minor adjustments in the coding of the original videos. Finally, the remaining videos were distributed among the three researchers for analysis, completing the full database.

2.6. Categories and coding

As mentioned in the previous section, we coded different types of data: (1) self-assessment strategies and criteria, and (2) process data. These were further specified in different categories.

First, we identified 8 self-assessment strategies (e.g., *Read the question*) and 11 self-assessment criteria (e.g., *Based on positive intuition*). Importantly, not all of them were found in all educational levels or subjects. A dummy variable was created for each category, for a total of 19 dichotomous variables. Considering each participant engaged in multiple self-assessment occasions during the experimental procedure (four times in secondary education, twice in higher education), a different set of variables was created for each self-assessment occasion.

Second, for the process data, we created four variables representing the first, second, third and fourth process executed by the participant. We opted for four variables because the number of participants using more than four different processes was extremely low. In this case they were not dummy variables; instead, the name of the process was coded in the correspondent variable, depending on whether it was performed first, second, third or fourth by the participant). This coding method was designed to capture the sequence of the processes rather than merely their occurrence or absence. Similarly to the approach for strategies and criteria, we created a set of four variables for each specific self-assessment occasion. Based on the processes performed, we identified and coded four profiles: (1) No Self-Assessment, (2) Superficial Self-Assessment, (3) Intermediate Self-Assessment, and (4) Advanced Self-Assessment. These profiles were coded as ordinal variables, and independent profiles were coded for each of the self-assessment occasions.

3. Results

3.1. RQ1. What insights emerged from the quantitative coding and analysis of think-aloud data?

We will present the results from three areas: think-aloud coding, type of data analyses, and results for the self-assessment field.

3.1.1. Think-aloud coding

One of the main results of the two quantitative papers (Panadero et al., 2020, 2022) is the categorization of self-assessment strategies and criteria in itself. These levels and categories can be used by self-assessment scholars to explore their own data, whether it is think-aloud or a different data source, as they represent basic self-assessment actions.

First in the secondary education data, we were able to identify four levels of strategies and four levels of criteria. Even further, in most of those levels we were able to further identify more specific categories of strategies and criteria. Interestingly, when we tried to use the same coding with the higher education sample this was not only applicable but also the same four levels of strategies and criteria were found to be valid. There were differences at the specific level of categories especially for criteria as we identified some differences there. However, in general the resemblance is that of a very similar process regardless of the educational level which is surprising considering the wide range of ages we investigated. Next, in Table 2 and Table 3 we present the levels and categories identified in each study.

Table 2

Category description and examples secondary education students (Panadero et al., 2020)

Level	Category	Description	Example comment
Self-assessment Strategies			
Level 0 Basic information processing	**Read the question**	The student read the question	*"First I will read what they ask"*
	Read the response	The student read his/her response	*"In this sentence I have written: You are like the sun"*
	Read the text	The student read the text to be analyzed (only in Spanish)	*"Well… I would read the text first"*
	Read /process the feedback received	The student read and process the feedback received	*"He (the teacher) has been kind. I would have graded it lower"*
Level 1 Comparing information strategies	**Compare text-response**	The student compares his/her response against the text (only in Spanish)	*"I have written the opinion of the author in the text, but anyway I am going to read it again to see if it is correct"*
	Compare question-response	The student compares his/her response against the question asked (only in Spanish**)	*"Then I would read question by question… and I would check if that is what is asked in the question"*
	Access their memory to compare	The student accesses his/her memory to compare his/her response against other data (only in Spanish**)	*"For the acronyms… I have to remember what they mean. Especially if they are in Latin"*
Level 2 Specific procedural strategies	**Review the signs**	The student reviews the signs in his/her response (only in mathematics)	*"The thing is that here I wrote 1 instead of -1, and here it should have been -1, and also here…"*
	Replace the X	The student replaces the X in his/her response (only in mathematics)	*"And after doing it again I get the same thing, that X equals -3"*
	Evaluate the procedure followed	The student evaluates the procedure followed in his/her response (only in mathematics*)	*"Then I made the tangent line here, and I now see that I forgot to draw it"*
Level 3 Advanced self-assessment strategies	**Compare own evaluation to feedback**	The student compares his/her previous self-assessment against the feedback received (only in Spanish**)	*"Now I just saw that question three I assessed it as good, but I still do not believe that that's right"*
	Perform the exercise again	The student changes the whole exercise or some parts of it	*"The verb archive… I will change it for school material. I think that I have got it wrong"*
	Think of different responses	The student thinks of different responses to the question (only in Spanish**)	*"This exercise… now I know how to do it. I would not have left it blank"*
Self-assessment Criteria			
Level 1 No criteria	**Without clear criteria**	The student doesn't use any clear criteria during his/her self-assessment	*"This I barely understand, but I think that some of it is correct"*
Level 2 Criteria based in personal reactions	**Based on intuition**	Based on his/her intuition at the moment of self-assessing	*"I think that I have got this wrong… Yes, I think this is not correct"*
	Based on hindsight	Based on hindsight at the moment of performing the exam	*"When I was doing it (the exam) I was not convinced by that answer. Now that I read it again… it does not convince me either"*
	Based on experience/self-efficacy	Based on the students' self-reported experience or self-efficacy (only in mathematics*)	*"Ii is just that, just like this exercise we have done a lot in class, so I know how it works"*
Level 3	**Based on rules**	Based on specific rules from the subject	*"The first one I think is an objective description, because the author is speaking"*

Criteria based on simple rules	**Based on spelling**	Based on the spelling of his/her response (only in Spanish)	*"I have skipped letters, which I have not noticed. Here, for example… I have skipped a letter"*
	Based on mistakes identified	Based on the mistakes identified by the student in his/her response (only in mathematics*)	*"Here, instead of writing f(1) I already wrote f(x)"*
	Given by the teacher	Based on instructions given by the teacher	*"As he (the teacher) says that the summaries have to be concise, I did not extend much"*
Level 4 Criteria based on complex rules	**Comparative criterion**	Based on the comparison made between his/her response and the text/question/both. This variable ranges from 0 (no comparison) to 2 (based on the comparison of the response and both question and text) (only in Spanish**)	*"Then I would read my answer, to see if it is what is asked… and I see that it is well answered"* *"As in this paragraph (of the text) the author points out several examples… I simply wrote that it points out examples. Maybe I should have explained it a little more"*
	Based on adequacy to the question	Based on the adequacy of the response in relation to the question asked (only in mathematics*)	*"You see that it asks you to optimize and maximize… and that is what I have done"*
	Based on the coherence of the result	Based on the mathematical coherence of the results obtained (only in mathematics)	*"I do it and I look at the result, and if it is consistent, then it is fine"*
	Based on steps followed	Based on the adequacy of the steps followed by the student (only in mathematics*)	*"I see that it is quite good, because I take out the common factor, I remove the two X's and I have… H=-1"*
	Based on formula application	Based on the adequacy of the formulas applied by the student (only in mathematics)	*"I remember the formula… and I look at it two or three times to see if it is okay. In this case I think it is okay"*

Table 3

Category description and examples for university students (Panadero et al., 2022)

Level	Category	Description	Example comment
Self-assessment Strategies			
Level 0 Basic information processing	Read the essay	The student read his essay.	*"Ok, so… Why is the psychologist profession necessary?"*
	Read the feedback or rubric received	The student read the feedback or rubric received.	*"He (the instructor) has been kind. I would have graded it lower"*
Level 1 Comparing information strategies	Compare instructions and essay	The student compares his essay with the instructions received.	*"Well I think that my text pretty much answers the question"*
	Compare essay to feedback or rubric	The student compares his essay with the feedback or rubric received.	*"Now I just saw that question three I have it well, but I still do not believe that that's right"*
Level 2 Remembering strategies	Remember the instructions	The student remembers the instructions of the task.	*"First I will read what they ask"*
	Remember the seminar	The student remembers the seminar on academic writing.	*"Yeah I remember your partner saying that we should be careful with the length of the sentences"*
Level 3 Advanced self-assessment strategies	Perform the essay again	The student changes the whole essay or some parts of it.	*"I should have made this paragraph shorter. Can I change it now?"*
	Think of different responses	The student thinks of different responses to the instructions.	*"I would have explained it differently if I had more time"*
Self-assessment Criteria			
Level 0 No criteria	Without clear criteria	The student doesn't use any clear criteria during his/her self-assessment	*"This I barely understand, but I think that some of it is correct"*
Level 1 Criteria based in personal reactions	Negative intuition	Based on a negative intuition at the moment of self-assessing.	*"I think that I have this wrong… Yes, I think this is not correct"*
	Positive intuition	Based on a positive intuition at the moment of self-assessing.	*"I am happy with my essay. It is not perfect, but I like it"*
	Negative hindsight	Based on a negative hindsight at the moment of writing the essay.	*"When I was doing it (the essay) I was not convinced by that answer. Now that I read it again… it does not convince me either"*
	Positive hindsight	Based on a positive hindsight at the moment of writing the essay.	*"I was inspired when I wrote it"*
Level 2 Criteria based on simple rules	Instructions	Based on the specific instructions of the task.	*"Well I think that my text pretty much answers the question"*
	Spelling	Based on the essay's spelling	*"I don't see spelling mistakes in my text"*
	Feedback received	Based on the feedback or rubric received by the student.	*"The instructor says that the ideas of my essay are quite confusing… and I agree"*
Level 3 Criteria based on complex rules	Writing process	Based on the process of writing the essay.	*"I should have taken a moment to think before started writing, but I was concerned about the lack of time"*
	Paragraph structure	Based on the essay's paragraph structure.	*"I think that the paragraphs are not too long. I have no paragraph longer than ten lines"*
	Sentences and punctuation marks	Based on the structure of the sentences and the adequacy of the punctuation marks used.	*"I have the problem of never knowing where I must use the semicolon. I use it randomly"*

As can be observed, both categorizations are very similar, despite the broad age range of the participants. An unintended difference arose when in secondary education, we labeled the first level of self-assessment criteria as 'Level 1,' whereas for university students, we later proposed 'Level 0' for the same level. We believe that using 'Level 0' is more appropriate as it better represents the absence of criteria.

3.1.2. Type of data analyses

A very similar data analysis strategy was followed in both articles. For categorical variables, we employed multiple dichotomy frequency tables to describe them, considering that each subject could exhibit more than one behavior. In the case of quantitative variables, we conducted descriptive analyses on the items, which involved calculating statistics such as the mean, standard deviation, median, and interquartile range. To explore the impacts of the independent variables, we calculated either ANOVAS or the Mann-Whitney and Kruskal-Wallis tests, respectively. As can be seen, this strategy was quantitative.

3.1.3. Results for the self-assessment field

Regarding the two quantitative studies, in the first, Panadero and colleagues (2020) explored the effect of feedback occasion, subject matter, year level and gender on the self-assessment strategies and criteria used. In the second study (Panadero et al., 2022), we explored the effects of feedback type, feedback occasion and year level on the strategies and criteria.

A noteworthy aspect addressed in both studies was the impact of feedback on self-assessment. In the 2020 study, it was observed that feedback played a pivotal role in shaping students' self-assessment strategies and criteria. After receiving feedback, students shifted their focus towards the content of the feedback itself, making it the primary reference point for evaluating their work. Consequently, this led to a decrease in the

number and complexity of strategies employed before receiving feedback. We suggested that delaying feedback until after self-assessment may potentially be a beneficial approach.

The subject matter also emerged as a crucial factor influencing self-assessment practices. According to the 2020 study, different subjects, such as Spanish and mathematics, led to distinct self-assessment profiles. Subject-specific demands significantly influenced the strategies and criteria students employed in self-assessment.

Regarding year-level differences, the 2020 study yielded unexpected results. In contrast to some previous research, no significant differences were found among university students across various year levels. This suggests that the educational level might not necessarily be correlated with differences in self-assessment practices for certain types of tasks.

Furthermore, the gender effect was consistent across both studies. Females tended to use a more diverse range of self-assessment strategies compared to males, aligning with previous research highlighting gender-related differences in self-assessment engagement.

In the 2022 study, additional insights were gained regarding feedback. Different types of feedback conditions were compared, with a particular focus on rubric-assisted feedback. The results indicated that rubrics had a significant positive impact on self-assessment. Rubric-assisted feedback outperformed instructor feedback, leading to more sophisticated self-assessment criteria. Rubrics appeared to stimulate student reflection and promote active self-assessment more effectively than instructor-written feedback.

The timing of feedback delivery also emerged as a noteworthy consideration. Both studies suggested that feedback might be more effective when provided after students have performed self-assessment based on their own strategies and criteria. Pre-feedback performance feedback could potentially discourage constructive self-assessment strategies.

In conclusion, these two studies collectively provide a nuanced understanding of self-assessment. They highlight the pivotal role of feedback, the influence of subject-specific demands, the presence of gender-related differences, and the potential benefits of rubric-assisted feedback. Moreover, they suggest that the timing of feedback delivery should be meticulously considered in the context of self-assessment processes.

3.2. RQ2. What insights emerged from the qualitative coding and analysis of think-aloud data?

3.2.1. Think-aloud coding

Similar to the quantitative approach, we also believe that the qualitative coding created by Panadero and colleagues (under review) is a result in itself because it presents a clear and defined way to code self-assessment across six processes and four self-assessor profiles. The description of each process is presented in Table 4.

Table 4

Coding categories for self-assessment processes extracted from process data (Panadero et al., under review)

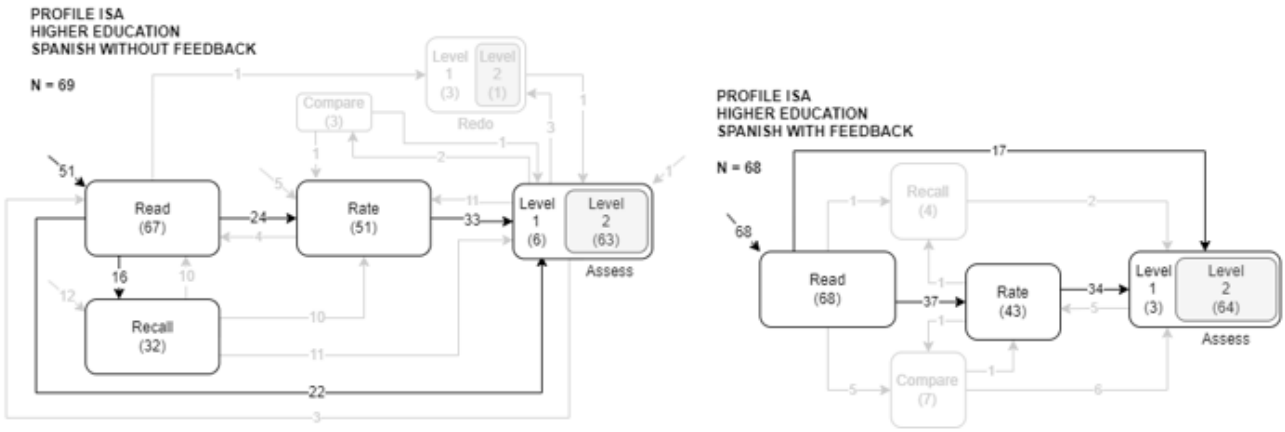
Process	Description
Read	The student reads the question, the answer, or any other part of the task
Recall	The student describes the question and their answer, without directly reading what is written and without making a judgement about the quality of the task.
Compare	The student compares different sources of information, usually between the question and their answer
Rate	The student makes an estimation (e.g., good, bad, average) about the quality of specific parts of the task, or about the whole task.
Assess	The student assesses the quality of their work using different criteria. Usually, it follows the pattern of "I think the task is _____ because _____". Assessing has two levels of complexity: In the first level, the student uses assessment criteria which are not based in rules and instructions (e.g., "I think I did a poor job because I was nervous"). In the second level, the student uses criteria based on the task, workshop, or teacher's instructions (e.g., "I think I did a poor job because some of my paragraphs are not really coherent").

Redo	The student performs again the exercise or part of it. Redoing has two levels of complexity: In the first level, the student would think on different responses to the exercise, without actively modifying their response. In the second level, the student actively modifies their response.

Additionally, we created a visualization system to facilitate comprehension and interpretation of the processes and profiles (Figure 2). The figure is a way to illustrate the different processes and depict different pathways on how students self-assess. Here, we also believe these levels and categories can be used by self-assessment scholars to explore their own data, and interestingly we have had two different colleagues (Canada and Germany) that have found our coding adjusting to their data on self-assessment via think-aloud.

Figure 2

Example figure. Intermediate self-assessment profile in higher education without and with feedback



3.2.2. Type of data analyses

In this study, we combined qualitative and quantitative data analyses. For the qualitative aspect, we used a descriptive analysis to calculate the frequencies of processes and profiles. In addition, we performed a qualitative content analysis of the think-aloud of the participants to explore differences in the profiles. Lastly, to validate our qualitative

coding, we employed quantitative analyses including the calculation of cross tables, discriminant analysis and ANOVA.

3.2.3. Results for the self-assessment field

As shown above, a powerful result for the field was to identify six key processes that students performed when they are asked to self-assess (see section *Think-aloud coding*). These six categories have been found in three independent studies currently in preparation by research teams from three countries. Thus, proving culturally valid.

Another crucial result was the identification of four self-assessment profiles. These profiles were identified using a dual qualitative and quantitative approach. Regarding qualitative calculation, the four processes were categorized based on the specific processes followed by the participant. The first profile was No Self-assessment represented by a participant who did not *Rate*, *Assess* or *Redo*. The second profile was Superficial Self-assessment represented by a participant who *Rate* the quality of the task but did not properly *Assess* or *Redo* it. In general, the participants in this category judged the quality of the performance (e.g., good, bad, average) but did not explain the criteria to reach such judgement. The third one was Intermediate Self-assessment represented by a participant who *Rate* the quality of the task and *Assess* it based on some assessment criteria. Here, while the participant was able to perform a complete self-assessment of the task in base to specific assessment criteria (generally based on teacher's instructions), still did not perform any process to improve the quality or correct the previous performance. And the fourth profile was Advanced Self-assessment represented by a participant who performed all the processes, including *Redo* the task (either orally or in paper). Importantly, these participants were able to perform a complete self-assessment of the performance and improve its quality by fixing the mistakes detected during the self-assessment process.

The quantitative profiles were calculated by assigning a numerical value to each process depending on its level of complexity. After several rounds of analysis, we defined the scores shown in the Table 5 as a suitable solution.

Table 5

Processes' scores

Process	**Score**
Read	0,5
Recall	0,5
Compare	0,75
Rate	1
Assess (Level 1)	1
Assess (Level 2)	1,5
Redo (Level 1)	1,5
Redo (Level 2)	2

Using this quantification of the processes, we assigned a "self-assessment score" to each participant by adding the value of all processes performed. However, to not bias our score in favor of participants using multiple basic processes, we used a weighted score, calculated by dividing the total score by the number of processes performed. The quantitative profiles were then calculated using the weighted scores. Based on our analyses, different cut-off points were found for each profile.

- From **0 to 0,5** – Mostly No Self-Assessment
- From **0,51 to 0,92** – Mostly Superficial Self-Assessment
- From **0,93 to 1,35** – Mostly Intermediate Self-Assessment
- From **1,36 to 3** – Mostly Advanced Self-Assessment

4. Discussion

Our aim was to present and reflect on an innovative way of employing think-aloud data integrating quantitative and qualitative approaches in the context of self-assessment research. We introduced three empirical studies derived from two data collections, utilizing these methods to examine the same phenomena. We believe our research demonstrates how these approaches can complement each other and provide an integrated understanding of the phenomena.

4.1. Think-aloud coding

Our think-aloud coding generated a list of self-assessment strategies, criteria, and actions performed by students in real-time during self-assessment. Similar insights into key self-regulatory strategies and processes during learning tasks have been provided in self-regulated learning research (Azevedo & Cromley, 2004; Azevedo et al., 2004; Azevedo et al., 2008). Indeed, the coding scheme from these studies has informed subsequent research focused on process data in SRL (e.g., Greene et al., 2008; Greene & Azevedo, 2009; Torrington et al., 2023). However, in the field of self-assessment, the use of process data is less common, with few studies beyond those discussed in this article (e.g., Kostons et al., 2009). Thus, we contend that our approach offers valuable insights for researchers seeking to delve deeper into the 'black box' of self-assessment.

Additionally, our qualitative coding method and visualization system can serve as models for capturing and illustrating similar processes. This method inspired Fernández-Ruiz et al. (2022) in their investigation of how teachers design assessments. Although their focus differed, they utilized a similar approach in creating codes based on observed actions during assessment design, analyzing action patterns to identify profiles, and employing flowchart visualizations. This demonstrates the broader applicability of our qualitative approach beyond the self-assessment domain, extending to other areas like peer-assessment

and feedback processing. However, since our research was conducted within the context of self-assessment, we will next discuss our main findings, which may be valuable to other researchers in this field.

4.2. Results for the self-assessment field

The three studies reported here shared the aim of "opening the black box of self-assessment" by disentangling the actions that students enact when performing self-assessment. We believe our think-aloud work has three main implications for the self-assessment field.

4.2.1. General and specific categorization for future research

Our dual approach, incorporating both quantitative and qualitative methodologies, has enabled us to examine self-assessment at two distinct levels: a finer grain level focusing on specific strategies and criteria, and a broader level encompassing overarching processes. Crucially, we have amalgamated these two levels into a comprehensive system, as illustrated in Table 6. We argue that this integrated framework offers a nuanced perspective, capturing both the general and specific dimensions of how students engage in self-assessment.

Table 6

Integration of categories

Strategies		Processes	Criteria	
Secondary Ed.	Higher Ed.		Secondary Ed.	Higher Ed.
Read the question		**Read**		
Read the response	Read the essay			
Read the text				
Read/process the feedback received	Read the feedback/ rubric received			
	Remember the instructions	**Recall**		
	Remember the seminar			
Compare text-response		**Compare**		
Compare question-response	Compare instructions-essay			
Access their memory to compare				
Compare own evaluation to feedback	Compare essay to feedback or rubric			
Rate				
Review the signs (operators)		**Assess (Level 1)** *Subjective (guts-feeling) indicators*	Without clear criteria	Without clear criteria
			Intuition	Negative intuition
				Positive intuition
			Hindsight	Negative hindsight
				Positive hindsight
			Experience/self-efficacy	
Replace the X		**Assess (Level 2)** *Compliance with external indicators*	Rules	Instructions
			Spelling	Spelling
			Mistakes identified	
			Given by the teacher	Feedback received
			Comparative	
				Writing process
				Paragraph structure
				Sentences and punctuation marks

Evaluate the procedure followed			Adequacy to the question
			Coherence of the result
			Steps followed
			Formula application
Think of different responses	Think of different responses	**Redo (Level 1)**	
Perform the exercise again	Perform the exercise again	**Redo (Level 2)**	

Only used in one educational level due to the characteristics of the task
Only used in secondary education (mathematics)
Only used in secondary education (Spanish)

As depicted in the table, processes are positioned at the core, integrating and providing context to the specific categories of strategies and criteria. This approach stems from the understanding that various strategies and criteria may be content-dependent, varying with the task, whereas processes tend to be more transferable and indicative of broader actions. It is evident that most processes encompass several more specific strategies or criteria. For example, the process *Redo* includes more specific strategies such as *Think of different responses* or *Perform the exercise again*.

We consider this categorization as an important advance towards our aim of "opening the black box of self-assessment". We have established a system of categories that not only covers different self-assessment areas, such as strategies and criteria, but also encompasses varying levels of specificity. This achievement was made possible through our dual approach. Firstly, our quantitative analysis focused on coding data to yield the most precise and specific frequencies possible, leading to the identification of a wide array of specific and context-dependent categories (19 different categories). Secondly, our qualitative analysis, aimed at capturing a broader view of individual patterns in

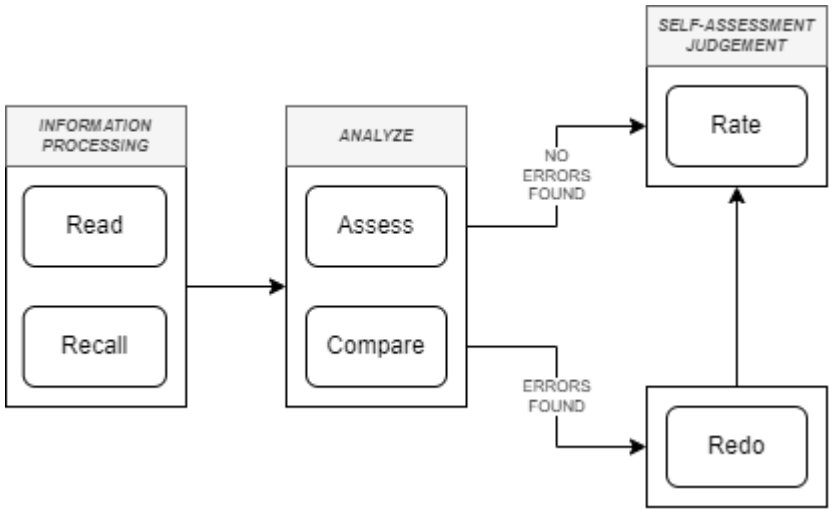
participants' self-assessment, revealed how the same phenomenon could be interpreted as a combination of 6 more general processes.

4.2.2. A model of self-assessment

Derived from the empirical evidence and the two previous studies experience, in Panadero and colleagues (under review) we propose a model on how an ideal self-assessment process could take place (Figure 3). It is important to emphasize that this evidence is coming from more than 500 self-assessment observations as each of the 67 secondary education students performed four self-assessments (Spanish without feedback, Spanish with feedback, Mathematics without feedback, and Mathematics with feedback) and 126 university students performed two self-assessments (Essay without feedback, Essay with feedback).

Figure 3

Model of self-assessment



The initial step in the self-assessment process involves *information processing*. This stage usually involves students engaging with material directly presented to them (i.e., read), such as their essays, or recalling pertinent information from past performances or

instructional feedback (i.e., recall). Following this, students undertake a thorough *analysis* of their work, evaluating various factors within the assessment (i.e., assess) and enacting comparison processes (detailed descriptions of these processes can be found in Tables 6, 2, and 3). It is particularly crucial for students, especially novices, to utilize clear criteria and align with external quality benchmarks to improve the precision of those two processes (Nicol, 2021; Panadero et al., 2019).

Furthermore, the last step involves a judgment after having analyze their previous performance where students should be capable of forming a well-reasoned judgment regarding the quality of their work. This can take two routes. First, if the students do not identify errors then they *rate* the quality of their work, being this from a score to a quality judgement (e.g., "I did well in this essay"). Second, if the students identify errors, the most advanced strategy they can activate is *redo*, in other words reattempting to perform the task rectifying any errors. Unfortunately, our research shows that only a low percentage of students redo their work, and usually those students are not the ones with the poorest performances. After this, the students could rate the quality of their work. These processes are guided by specific strategies and criteria (see tables), which we have extensively documented in our 500-plus self-assessment records.

Importantly, similar to self-regulated learning models, this framework relies on the deliberate and strategic execution of these processes. Consequently, students can employ these steps even without conscious awareness, and experts may not always require all of them. Like any model, the process lacks absolute clarity, and the timing of each process, as well as the strategies and criteria employed, may vary. It is important to note that our representation is based on data from a convenient sample, which we believe closely aligns with an expert approach to self-assessment. Nevertheless, it is worth acknowledging that

there is an overlap among the three of the self-assessment profiles (Superficial, Intermediate, and Advanced), underscoring the complexity inherent in the self-assessment process.

4.2.3. Empirical conclusions on key variables and processes influencing self-assessment

Importantly, the creation and integration of these categories is not our only meaningful result for the self-assessment field, even if we consider it the most innovative. Our categories allowed us to compare our participants' self-assessment based in several variables (i.e., gender, year level, subject, feedback occasion). Our results indicate that (1) females tend to be better self-assessment performers, (2) year level does not have a significant effect on self-assessment, (3) subject matter clearly affects the strategies and criteria deployed in self-assessment and (4) the presence of external feedback could potentially discourage deep approaches to self-assessment so the timing of that feedback is key -i.e., allowing the students to first self-assess and then give them feedback.

4.3. Methodological recommendations

We offer three methodological recommendations for think-aloud research. Firstly, it's crucial to enable participants to express their internal processes freely. Such an approach is necessary to minimize potential biases in performance attributed to the think-aloud procedure. As indicated in previous research (Fox et al., 2011), bias may arise when participants are required to recode information for reporting purposes. For instance, instructing students to categorize their actions or thoughts using a predefined set of categories during think-aloud can lead to such biases. While this issue is less prevalent in concurrent think-aloud, it remains a possibility. A typical example might involve asking participants to 'verbalize their self-assessment strategies as they apply them.' This directive

could inadvertently lead to improved performance, an artifact of the instructions rather than an accurate reflection of the participants' abilities.

Secondly, we recommend having a clear understanding of the information that will be analyzed and the methodology for its analysis, which includes formulating precise and focused instructions. This approach is essential to minimize the need for interventions during the think-aloud data collection. While it is important to allow participants to freely communicate their internal processes, it is equally vital to collect data that addresses our research questions effectively. Therefore, balancing these two aspects can be challenging. On one hand, participants should have the freedom to express their thoughts and actions. On the other hand, the data collected must align with our research objectives. This tension represents a challenge that requires careful management.

Our third recommendation is to identify, after thorough reflection, the most suitable approach for analyzing the think-aloud data obtained. Given our discussion, a well-considered analysis could potentially enhance the conclusions derived from think-aloud studies. We also suggest maintaining a clear vision on how to utilize multimodal data as to ensure a thorough and nuanced analysis (Panadero, 2023).

4.4. Limitations

There are two main limitations in our empirical data. First, think-aloud is a self-reported data source that comes with a massively debated pros and cons dilemma. Importantly, concurrent think-aloud might avoid some of those cons. Second, as with any coding of data, our categories can be up to debate on their veridicality of the real phenomena. We have performed several actions to unsure this is the case. First, we use the same coding approach from Panadero and colleagues (2020) to Panadero and colleagues (2022) with a completely different population. Second, in Panadero and colleagues (under

review) we performed a quantitative validation of the qualitative coding. And third, we have elaborated a manual for future research that is being used by other research groups to analyze their self-assessment data. The three actions have supported the hypothesis that our coding is valid for other studies and the model can be maintained.

4.5. Future lines of research

First, it would be interesting to continue contrasting the validity of our coding and model in research performed by other research groups. Second, it would be beneficial for the educational assessment to study the "black box" in other areas such as peer assessment or teacher feedback. Third, triangulating think-aloud data with other data sources such as eye-tracking or physiological measures, will strengthen our understanding of think-aloud processes.

5. Conclusion

Think-aloud stands as a potent data collection technique, offering profound insights into students' internal cognitive, motivational and emotional processes. In this study, we have introduced an approach to analyzing such data, employing both quantitative and qualitative methods, to draw more robust conclusions about the mechanisms underlying self-assessment. We anticipate that fellow researchers could consider our work as a means to enhance the utilization of think-aloud data, encouraging a more comprehensive exploration of the implications within their research. The realm of think-aloud processes holds vast potential for discovery; it beckons us to delve deeper into its intricacies.

References

Azevedo, R., & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of educational psychology*, *96*(3), 523. https://doi.org/10.1037/0022-0663.96.3.523

Azevedo, R., Guthrie, J. T., & Seibert, D. (2004). The role of self-regulated learning in fostering students' conceptual understanding of complex systems with hypermedia. *Journal of Educational Computing Research, 30*(1-2), 87-111. https://doi.org/10.2190/DVWX-GM1T-6THQ-5WC

Azevedo, R., Moos, D. C., Greene, J. A., Winters, F. I., & Cromley, J. G. (2008). Why is externally-facilitated regulated learning more effective than self-regulated learning with hypermedia?. *Educational Technology Research and Development, 56*, 45-72. https://doi.org/10.1007/s11423-007-9067-0

Azevedo, R., Taub, M., & Mudrick, N. V. (2018). Understanding and reasoning about real-time cognitive, affective, and metacognitive processes to foster self-regulation with advanced learning technologies. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (pp. 254-270). Routledge. https://doi.org/10.4324/9781315697048-17

Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied psychology*, *54*(2), 199-231. https://doi.org/10.1111/j.1464-0597.2005.00205.x

Charters, E. (2003). The Use of Think-aloud Methods in Qualitative Research An Introduction to Think-aloud Methods. *Brock Education Journal, 12*(2). https://doi.org/10.26522/brocked.v12i2.38

Creswell, J. W., & Plano-Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd. ed.). Sage.

Educational Endowment Foundation. (2018). Teaching & Learning Toolkit. https://educationendowmentfoundation.org.uk/public/files/Toolkit/complete/EEF-Teaching-Learning-Toolkit-October-2018.pdf

Ericsson, K. A., & Fox, M. C. (2011). Thinking aloud is not a form of introspection but a qualitatively different methodology: Reply to Schooler (2011). *Psychological Bulletin, 137*(2), 351–354. https://doi.org/10/dnf5zn

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological review*, *87*(3), 215. https://doi.org/10.1037/0033-295X.87.3.215

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of educational research*, *59*(4), 395-430. https://doi.org/10.2307/1170205

Fernández Ruiz, J., Panadero, E., García-Pérez, D., & Pinedo, L. (2022). Assessment design decisions in practice: Profile identification in approaches to assessment design. *Assessment & Evaluation in Higher Education*, *47*(4), 606-621. https://doi.org/10.1080/02602938.2021.1937512

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological bulletin*, *137*(2), 316. https://doi.org/10.1037/a0021663

Greene, J. A., & Azevedo, R. (2009). A macro-level analysis of SRL processes and their relations to the acquisition of a sophisticated mental model of a complex system. *Contemporary educational psychology*, *34*(1), 18-29. https://doi.org/10.1016/j.cedpsych.2008.05.006

Greene, J. A., Moos, D. C., Azevedo, R., & Winters, F. I. (2008). Exploring differences between gifted and grade-level students' use of self-regulatory learning processes with hypermedia. *Computers & Education*, *50*(3), 1069-1083. https://doi.org/10.1016/j.compedu.2006.10.004

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

James, W. (1890). *The principles of psychology*. Holt.

Karoly, P., Boekaerts, M., & Maes, S. (2005). Toward consensus in the psychology of self-regulation: How far have we come? How far do we have yet to travel? *Applied Psychology*, *54*(2), 300-311. https://doi.org/10.1111/j.1464-0597.2005.00211.x

Kostons, D., Van Gog, T., & Paas, F. (2009). How do i do? Investigating effects of expertise and performance-process records on self-assessment. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 23*(9), 1256-1265. https://doi.org/10.1002/acp.1528

Lui, A. M., & Andrade, H. (2022). Inside the Next Black Box: Examining Students' Responses to Teacher Feedback in a Formative Assessment Context. *Frontiers in Education*, 7, 751549. https://doi.org/10.3389/feduc.2022.751549

Máñez, I., Vidal-Abarca, E., Kendeou, P., & Martínez, T. (2019). How do students process complex formative feedback in question-answering tasks? A think-aloud study. *Metacognition and Learning, 14*, 65-87. https://doi.org/10.1007/s11409-019-09192-w

McIntyre, N. A., Draycott, B., & Wolff, C. E. (2021). Keeping track of expert teachers: Comparing the affordances of think-aloud elicited by two different video perspectives. *Learning and Instruction*, *80*, 101563. https://doi.org/10.1016/j.learninstruc.2021.101563

Panadero (2023). Toward a paradigm shift in feedback research: Five further steps influenced by self-regulated learning theory. *Educational Psychologist, 58*:3, 193-204, DOI: 10.1080/00461520.2023.2223642.

Panadero, E., Brown, G. T., & Strijbos, J. W. (2016). The future of student self-assessment: A review of known unknowns and potential directions. *Educational psychology review*, *28*, 803-830. https://doi.org/10.1080/00461520.2023.2223642

Panadero, E., Fernández-Ruiz, J., & Sánchez-Iglesias, I. (2020). Secondary education students' self-assessment: The effects of feedback, subject matter, year level, and gender. *Assessment in Education: Principles, Policy & Practice, 27*(6), 607-634. doi:10.1080/0969594X.2020.1835823

Panadero, E., García-Pérez, D., Fernández-Ruiz, J., Fraile, J., Sánchez-Iglesias, I., & Brown, G. T. L. (2022). University students' strategies and criteria during self-assessment: Instructor's feedback, rubrics, and year level effects. *European Journal of Psychology of Education, 38*(3), 1031-1051. doi:10.1007/s10212-022-00639-4

Panadero, E., Lipnevich, A. A., & Broadbent, J. (2019). Turning self-assessment into self-feedback. In D. Boud, M. D. Henderson, R. Ajjawi, & E. Molloy (Eds.), *The Impact of Feedback in Higher Education: Improving Assessment Outcomes for Learners* (pp. 147–163): Springer.

Panadero, E., Pinedo, L., & Fernández-Ruiz, J. (under review). Identifying self-assessment profiles: what secondary and higher education students do when generating self-feedback. In *Assessment in Education: Principles, Policies & Practices.*

Rickey, N., DeLuca, C., & Beach, P. (2023). Towards a new theory of student self-assessment: Tracing learners' cognitive and affective processes. *Metacognition and Learning*, 1-37. https://doi.org/10.1007/s11409-023-09359-6

Saint, J., Fan, Y., Gašević, D., & Pardo, A. (2022). Temporally-focused analytics of self-regulated learning: A systematic review of literature. *Computers and education: Artificial intelligence*, *3*, 100060. https://doi.org/10.1016/j.caeai.2022.100060

Shapiro, M. A. (2014). Think-aloud and thought-list procedures in investigating mental processes. In In A. Lang (Eds.) *Measuring psychological responses to media messages* (pp. 1-14). Routledge.

Sitzmann, T., Ely, K., Brown, K. G., & Bauer, K. N. (2010). Self-assessment of knowledge: A cognitive learning or affective measure? *Academy of Management Learning & Education*, *9*(2), 169-191. https://doi.org/10.5465/AMLE.2010.51428542

Tian, L., Liu, Q., & Zhang, X. (2022). Self-regulated writing strategy use when revising upon automated, peer, and teacher feedback in an online English as a foreign language writing course. *Frontiers in Psychology, 13*, 873170. https://doi.org/10.3389/fpsyg.2022.873170

Torrington, J., Bower, M., & Burns, E. C. (2023). What self-regulation strategies do elementary students utilize while learning online? *Education and Information Technologies*, *28*(2), 1735-1762. https://doi.org/10.1007/s10639-022-11244-9

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In *Handbook of self-regulation* (pp. 531-566). Academic Press. https://doi.org/10.1016/B978-012109890-2/50045-7

Winstone, N. E., & Nash, R. A. (2023). Toward a cohesive psychological science of effective feedback. *Educational Psychologist*, *58*(3), 111-129. https://doi.org/10.1080/00461520.2023.2224444

Yan, Z., & Brown, G. T. (2017). A cyclical self-assessment process: Towards a model of how students engage in self-assessment. *Assessment & Evaluation in Higher Education, 42*(8), 1247-1262. https://doi.org/10.1080/02602938.2016.1260091

Yan, Z., Lao, H., Panadero, E., Fernández-Castilla, B., Yang, L., & Yang, M. (2023). Effects of self-assessment and peer-assessment interventions on academic performance: A pairwise and network meta-analysis. *Educational Research Review*, 100484. https://doi.org/10.1016/j.edurev.2022.100484