Secondary Education Students' Self-Assessment: The Effects of Feedback, Subject Matter, Year Level, and Gender

Ernesto Panadero[1 2], Javier Fernández Ruiz[3] & Iván Sánchez-Iglesias[4]

[1] Facultad de Psicología y Educación, Universidad de Deusto, Bilbao, España.

[2] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain.

[3] Departamento de Psicología Evolutiva y de la Educación, Universidad Autónoma de Madrid, Spain.

[2] Departamento de Psicobiología y Metodología de las Ciencias del Comportamiento, Universidad Complutense de Madrid, Spain.

Recommended citation:
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*, (Online first). doi: 10.1080/0969594X.2020.1835823

Funding:

Research funded by: personal grant to first author under Ramón y Cajal framewok (RYC-2013-13469); Fundación BBVA call Investigadores y Creadores Culturales 2015 (project name Transición a la educación superior id. 122500); and by Spanish Ministry of Economy and Competitiveness (Ministerio de Economía y Competitividad) National I+D Call (Convocatoria Excelencia) project reference EDU2016-79714-P.

Correspondence concerning this manuscript should be addressed to: Ernesto Panadero. Despachos IRPO, Facultad de Psicología y Educación. Universidad de Deusto, Bilbao. 48007. Spain. E-mail: ernesto.research@gmail.com

Abstract

The effects of relevant factors related to self-assessment have not been systematically investigated. We explored four factors and their effects on self-assessment and self-efficacy: (1) feedback (with vs without), (2) subject matter (Spanish vs mathematics), (3) year level (K7 vs K10 vs K11), and (4) gender. The participants included 64 secondary education students who self-assessed during a set of Spanish and mathematics activities while being video-recorded. Data came from think-aloud protocols, direct observations, and self-reported instruments. The use of self-assessment strategies and criteria was more frequent and advanced without feedback and in females. There were differences in the self-assessment of Spanish and mathematics. As for year level, results showed more similarities than expected, though the use of advanced strategies and criteria varied across levels. Additionally, none of the factors had significant effects on self-efficacy. This study opens a new avenue for self-assessment research, unveiling the *black box* of self-assessment.

Keywords: self-assessment; subject matter; feedback effects; developmental effects; gender effects.

Highlights

- Research has studied self-assessment as a unified strategy, here we open the *black box* of self-assessment
- Four factors were explored: feedback occasion, subject matter, year level, and gender
- The dependent variables included self-assessment strategy use (number and type), criteria (number and type), and self-efficacy
- The four factors showed significant effects on self-assessment strategy and criteria use but just an interaction was significant for self-efficacy

Secondary Education Students' Self-Assessment: The Effects of Feedback, Subject Matter, Year Level, and Gender

Self-assessment is and has been one of the main lines of work in educational research. As pointed out in a review of the state of the field (Panadero, Brown & Strijbos, 2016), from the early studies that focused on students' self-grading – i.e. self-assessment understood as self-awarding a grade (Falchikov & Boud, 1989) – to the more recent ones on formative self-assessment (e.g. Andrade, 2018), some aspects of self-assessment have remained unexplored. The same review emphasized the need to increase our knowledge on the effects that factors such as development or subject matter could have on self-assessment. Our goal is to identify key areas that research has not yet covered and to investigate those factors effects. Most importantly, this study will try to capture the students *online* self-assessment by recording their thinking aloud protocols and actions while self-assessing, which is an empirical approach almost never employed.

Self-assessment and Variables of Interest

Panadero et al. (2016) argue that self-assessment "…generally 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" (p. 804). There have been at least three main lines of research in self-assessment. First, it was studied whether students could accurately self-assess when compared to teachers' scores. In general humans have a tendency for flawed self-assessment (Dunning, Heath, & Suls, 2004), but when it comes to scoring accurately, evidence shows that students can be reasonably accurate if a number of factors are given (e.g., training, opportunities for practice, and simple tasks) (Andrade, 2018; Brown, Andrade, & Chen, 2015; Falchikov & Boud, 1989). Secondly, self-assessment has been found to have a positive effect on students' performance, especially when implemented under formative approaches (Brown & Harris,

2018; Dochy, Segers, & Sluijsmans, 1999). For example, Brown and Harris' (2013) meta-analysis found that the median effect of self-assessment on academic performance lay between .40 and .45. And thirdly, there has been an interest in how self-assessment influences self-regulated learning, with a recent meta-analysis showing positive effects (Panadero, Jonsson & Botella, 2017). These two last lines of work, self-assessment effects on academic performance and self-regulated learning, have increased the interest on how different variables might affect student's self-assessment. Here, we will explore the effects of four factors that previous publications have pointed out as crucial, but the empirical evidence of the influence of these factors is missing or limited. The factors we will investigate include feedback, year level, gender, and subject matter.

Butler and Winne (1995) proposed that self-regulated learners go through loops of internal feedback that gets adjusted by external feedback. In the self-regulated models, there is a "self-evaluation" process that refers to self-assessment (e.g. Pintrich, 2000; Zimmerman, 2000). Self-assess is crucial for self-regulated students, and for this assessment to be accurate, feedback is needed, up to a point that some scholars have started talking of "self-feedback" (Andrade, 2018). Interestingly, studies exploring the impact of feedback on self-assessment processes are extremely scarce. We know of two. Panadero, Alonso-Tapia and Huertas (2012) found that process feedback -in interaction with the self-assessment occasion- increased students' self-efficacy more than performance feedback. However, the interaction type of feedback X self-assessment tools (rubric, script, and control) did not produce any effect on performance or self-efficacy. Raaijmakers and colleagues (2019) found that feedback on the accuracy of self-assessment did not increase the subsequent accuracy, regardless of the presentation of the correct answer. Though these two studies do not hold much promise for the effects of feedback on self-assessment, more research is needed as "…some student self-assessment (SSA) scholars have recommended that SSA requires

training in which students receive feedback about their own SSA so as to become more accurate self-assessors" (Panadero et al., 2016, p. 815). This study will be innovative, because we will evaluate the effects on the self-assessment process, not on other dependent variables (Panadero et al., 2012) or the accuracy of self-assessment (Raaijmakers et al., 2019).

A second factor of interest is subject matter. Assessment practices are shaped by the discipline and by the teachers' pedagogical content knowledge (Wiliam, 2019). In higher education changes in assessment practices are influenced by the academic discipline (Panadero et al., 2019) as different courses (e.g. medicine vs. history) are evaluated in distinct ways. While the review by Falchikov and Boud (1989) on self-assessment accuracy found that hard sciences seem to produce more accurate self-assessment, the recent state-of-the-art review pointed out that research has largely overlooked the effects of subject matter (Panadero et al., 2016). To our knowledge, there are no studies that explore two different subjects simultaneously. In this study, subject matter will be investigated via a Spanish essay task and through mathematical exercises. Spanish and mathematics tasks might require different levels of cognitive load, working memory demands, or the specific strategies for mathematics could be more difficult to verbalize. Nevertheless, profiling different subjects will amplify our knowledge about self-assessment.

Third, regarding year level, previous research shows that there is a tendency for more accurate self-assessment in individuals who have greater academic ability or performance (Barnett & Hixon, 1997) and that self-assessment is more difficult for novices (Kostons et al., 2009, 2010). Panadero et al. (2016) argued that both the developmental stage and the domain expertise had not been sufficiently explored. Our participants were drawn from the first (K7) and last (K10) years of compulsory secondary education and the first (K11) two years at the university preparatory level. Therefore, we will compare three different years and three

different levels with students at different developmental stages, while controlling for domain expertise by selecting activities in which the groups have similar levels of expertise.

Lastly, the effect of gender has been largely studied in education (e.g. Butler & Hasenfratz, 2017). The effects of gender on self-assessment have been a largely overlooked line of research. For example, Boud and Falchikov (1989) found in their review that only six out of 48 studies had explored the role of gender on self-assessment accuracy. Panadero et al. (2017) meta-analysis also found an extraordinarily low number of studies exploring these effects. More importantly, those authors found a high interaction of self-assessment interventions and gender, with girls clearly outperforming boys in self-efficacy after intervention. Therefore, we will investigate if boys and girls perform differently on self-assessment as it has been argued that girls might benefit motivationally from being more conservative on their self-evaluations (Butler & Hasenfratz, 2017).

Regarding the methodology to explore self-assessment processes, we will use a combination of data methods (i.e., thinking aloud protocols, direct observation, and self-report) to open the *black box* of self-assessment. This information will be organized around the strategies (number and type) and criteria (number and type) that students use while self-assessing. By analyzing these, we will disentangle self-assessment into more specific actions that will allow for extracting empirical conclusions. To our knowledge, there is only one previous study that explores how complex forms of self-assessment are specifically deployed by students (Yan & Brown, 2017). According to this model, self-assessment deploys around three main actions: determining performance criteria, seeking self-directed feedback, and self-reflection. One important limitation of Yan and Brown (2017) is that their data is derived from interviews, not from analyzing the students' actions while self-assessing, which we will do here.

Additionally, we will investigate the effects of self-assessment on self-efficacy. This variable is one of the main predictors of academic achievement (Richardson et al., 2012), key for students' motivation (Pajares, 2008) and self-regulated learning (Zimmerman, 2000). Self-efficacy constitutes a central aspect of self-assessment as, depending on how the student interpret the results, this will impact future performance and self-assessment. Interestingly, a recent meta-analysis has found a .73 effect size of self-assessment interventions on self-efficacy while pointing out that more systematic research is needed. Therefore, we will investigate how our four factors influence self-assessment and, through it, self-efficacy.

Finally, regarding our study population, in general self-assessment has been studied more among higher education students than at any other educational level (e.g., Boud & Falchikov, 1989; Falchikov & Boud, 1989; Panadero et al., 2017). However, the existent body of knowledge on secondary education (i.e., Brown & Harris, 2013) shows that self-assessment has an important effect on academic performance. It was then decided to explore secondary education students as their independence and autonomy is increased and self-assessment seems a crucial skill.

Research Aim and Questions

This study aims to explore the effects of four factors on self-assessment and self-efficacy: feedback (with vs without), subject matter (Spanish vs mathematics), year level (K7 vs K10 vs K11), and gender (male vs female). Self-assessment was divided into strategies (number and type) and criteria (number and type) that students elicited while thinking out loud and through observation. Self-efficacy was measured via self-report.

Research Question 1 (RQ1). What are the effects of feedback, subject matter, year level, and gender on self-assessment strategy use (number and type)?

Hypothesis (H1). Self-assessment strategy use will show higher numbers and more advanced types in the content area of Spanish for older students and girls. However, with feedback the number and type of strategies will decrease.

RQ2. What are the effects of feedback, subject matter, year level, and gender on self-assessment criteria use (number and type)?

H2. Self-assessment criteria will show higher numbers and more advanced types in the content area of Spanish for older students and girls. However, with feedback the number and type of criteria will decrease.

RQ3. What are the effects of feedback, subject matter, year level, and gender on self-efficacy?

H3. Self-efficacy will remain stable regardless of the feedback occasion; there are no hypotheses for year level; and self-efficacy will be higher for females and Spanish self-assessment task. Previous research shows that females benefit more from self-assessment and that students have lower self-efficacy for mathematics.

Method

Sample

The initial sample included 67 secondary students from a high school in Madrid, Spain. The final sample after deleting cases with incomplete data was 64 students from three different years representing three different levels (first year K7, fourth year K10, and first year of university preparation K11), with a gender distribution of 26 boys (40.6%) and 38 girls (59.4%). The distribution per year level was 28 students from K7 (43.8%, M_{Age} = 12.3 years, SD = .94), 22 from K10 (34.4%, M_{Age} = 15.6 years, SD = 1.14), and 14 from K11 (21.9%, M_{Age} = 16.8 years, SD = 1.31). These three-year levels were selected because one is the beginning and another at the end of compulsory secondary education in Spain (K7 and K10). Lastly, the fifth year (K11) is the first of two years' preparation for university entry and

has distinct features – i.e., more academically able students. Even though the statistical power always benefits from a larger sample, our data collection and analysis procedures was demanding and involved a significant workload. Nevertheless, in terms of power a sample size of 64 participants is sufficient for addressing our objectives. In a 2 (subject matter) X 2 (feedback condition) repeated measures ANOVA, given $\alpha = .050$, $N = 64$, and supposing an effect size $f = 0.20$, a statistical power $1 - \beta = .883$ would be achieved.

The sampling method at the high school level was based on convenience sampling i.e., the research group and the institution had a research agreement. At the participant level it was volunteer sampling as parental permission was needed. The participants did not receive any reward.

Research Design and Variables

This study followed a mixed methods approach that included qualitative and quantitative data using a convergent design data collection process (Creswell & Plano-Clark, 2018) using the coding of thinking aloud protocols, questionnaire data, and combining qualitative and quantitative analysis and interpretation. This is a cross-sectional design (three-year levels K7, K10, and K11), exploring the effects of subject matter (Spanish vs mathematics), feedback occasion (without vs with), and gender. The dependent variables were comprised of (a) self-assessment strategies (number and type), (b) criteria (number and type), and (c) self-efficacy.

Data Collection and Instruments

Coded Video-recorded Data on the Thinking Aloud Protocols and Direct Observation of Students' Actions. While in the experimental setting, the students were video-recorded. These videos were coded, thereby analyzing students' thinking aloud protocols and actions. Most of the categories were designed as dichotomous (absence or presence) (see Appendix A for the complete coding sheet). Next, we explain the process for

the creation of the coding system. Firstly, a deductive approach was employed to create five general coding categories of self-assessment elements: strategies, criteria, emotion, judgment, and reaction to feedback. Additionally, we created subcategories for those five general categories. Afterwards, the general and specific categories were contrasted with the data using an inductive approach. This process was performed over fifteen video-recordings reaching the final codebook. Later, it was decided to only keep strategies and criteria as these had the strongest and most reliable thinking aloud and direct observation coding. The other three categories had weaker measurements (e.g., reactions to feedback were interpretation of the researchers of the students' facial reactions). Additionally, we created two subcategories (i.e., number and type) for the two categories of strategy and criteria. The final coding sheet contains 13 types of strategies and 13 types of criteria. Importantly, during the writing of this manuscript we further organized each set of 13 categories into four levels for clarity in interpretation (see column level in Appendix A).

Three research assistants coded the data coordinated by the first author. The four met three times in three consecutive weeks to establish the procedure and ensure appropriate inter-judge agreement procedures. In the first week, the four coded six video-recordings reaching an average Krippendorff's Alpha of .81 with six categories below .70. The second week, five new video-recordings were coded reaching an average of .84 with four categories below .70. Finally, in the third week, with more than six new video-recordings, the average reached .86 and none of the categories were below .70. From that point onwards, all the videos were divided among the three assistants including the videos already coded. Five of the video-recordings were coded by all three, but the assistants did not know which ones. The average Krippendorff's Alpha was .85 with no categories falling below .70. The coding process was time consuming and took on average three hours for each video, as the coders visualized them several times.

Students' Self-efficacy Scale. This instrument, comprised of 14 items (8 for Spanish, 6 for mathematics), was slightly adapted to reference the specific contents and competences for each year and level. The items were answered in a 7-point Likert scale, ranging from *not capable at all* to *extremely capable*. The internal consistency of this instrument, computed for all three measurement moments (pre self-assessment, without feedback, and with feedback) ranged from .81 to .82 for Spanish, and from .87 to .92 for mathematics.

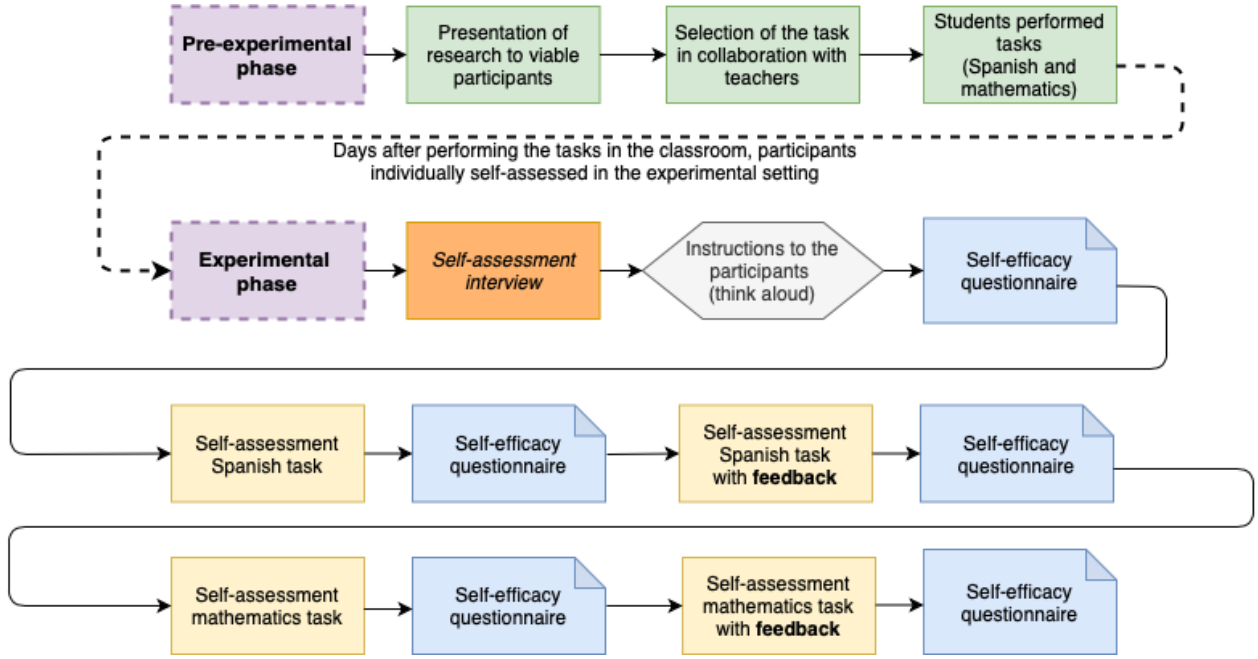
Procedure

Figure 1 shows the research procedure. The research was presented during classroom time in twelve classroom groups highlighting the need for parental permission for participation. The research team collaborated with the Spanish and mathematics teachers selecting a set of exercises, aiming for equal difficulty among the three-year levels. In this collaboration, the teachers and researchers (1) identified exercises with similar levels of difficulty and performance time, (2) teachers checked previous cohorts scores to ensure the variability in results for each year level, and (3) after the participants performed the exercises score variability was checked. Students performed the exercises during class time. Then, the teachers and the research team collaborated in grading and providing feedback for the exercises. Teachers discussed with the research team exemplars of previous years (high, medium and low performance) and provided a criteria list for correcting the exercises. Then the research team graded and provided feedback to five cases for each year level using the directions given the teacher. These fifteen cases were jointly analyzed to ensure correct grading and feedback. The latter aimed at the task and process levels as per Hattie and Timperley's model (2007). For common errors, a database with feedback samples was created and used when needed. The rest of the cases were always corrected by two research assistants: first one of them performed the correction, then the second revised it. From this

point on, for simplicity, we will refer to Spanish task and mathematics task in relationship to the sets including different exercises.

Figure 1.

Research procedure



Afterwards, the participants individually accompanied a research assistant to a quiet room within the high school. There, this protocol was followed: (a) participants were interviewed about the assessment and self-assessment (i.e. Do you usually self-assess your work?). Each interview lasted between 5 to 10 minutes; (b) Participants conducted a self-assessment of the Spanish task without being given any further instruction; (c) participants conducted a self-assessment of the Spanish task with feedback; (d) participants conducted a self-assessment of the mathematics task without being given any further instruction; and (e) participants conducted a self-assessment of the mathematics task with feedback. Between the different phases, the participants filled out a self-efficacy scale for a total of five times.

The participants were prompted before their first self-assessment to *think out loud* all their thoughts, emotions, and motivational reactions. If the participant remained silent for

more than 30 seconds, the researcher reminded him/her of the need to think out loud. The whole process took an average of 45 minutes and was video-recorded for later analysis. Due to the length of the transcripts, the interview data will not be presented in this study.

Data Analysis

The categorical variables are described using multiple dichotomy frequency tables, as each subject could display more than one behavior. For quantitative variables, the descriptive analyses of the items were made calculating the mean, standard deviation, median, and interquartile range. To investigate the effects of feedback and subject matter, we used a two-way repeated measures ANOVA; while for gender and year level we used Mann-Whitney and Kruskal-Wallis tests, respectively as the dependent variables were not normally distributed. All statistical analyses were performed using SPSS 20.

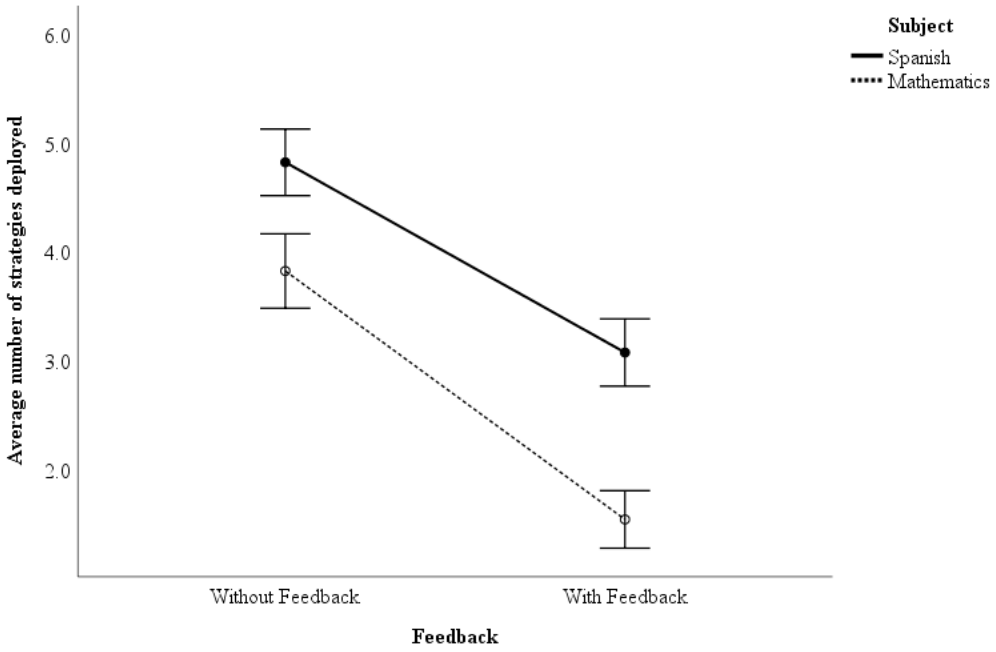
Results

RQ1. What are the effects of feedback, subject matter, year level, and gender on self-assessment strategy use (number and type)?

Number of Strategies. During the self-assessment of the Spanish task, the number of strategies was higher (without feedback $M = 4.82$, $SD = 1.15$, with feedback $M = 3.07$, $SD = 1.16$) than during the mathematics task (without feedback $M = 3.82$, $SD = 1.28$, with feedback $M = 1.54$, $SD = 0.99$). A two-way repeated measures ANOVA was conducted to assess the effect of subject matter and feedback on the number of strategies used by the participants. There was a significant effect for subject matter, with a higher number of strategies in Spanish ($M = 3.95$, $SD = 0.82$) than in mathematics ($M = 2.68$, $SD = 0.94$), $F(1, 55) = 66.47$, $p < .001$, $\eta^2 = .547$. There was also a significant effect for feedback with more strategies used without feedback ($M = 4.32$, $SD = 0.88$) than with feedback ($M = 2.30$, $SD = 0.87$), $F(1, 55) = 175.44$, $p < .001$, $\eta^2 = .761$. An interaction effect between both factors was found; the difference between Spanish and mathematics (in average number of strategies

enacted) was significant without and with feedback ($p < .001$), but it was greater with feedback, $F(1, 55) = 4.65$, $p = .035$, $\eta^2 = .078$. This interaction can be seen in Figure 2.

Figure 2.
Number of strategies deployed, by feedback and subject matter. Error bars represent +/-2 SE.



Then we analyzed the number of strategies enacted, considering feedback occasions and subject matter by year level (Table 1). Interestingly, there were significant differences only in the before feedback occasions for both subjects. In Spanish, K7 outperformed K10 and in Mathematics, K10 outperformed K7. Therefore, year level effects seem to be balanced when it came to the number of strategies.

Analyzing the number of strategies enacted, considering feedback occasion and subject matter by gender (Table 2) in three of the four self-assessments, females used a higher number of strategies. However, the difference was only significant in the Spanish with feedback.

Table 1

Number of strategies by feedback, subject matter and year level. Descriptive statistics and group comparisons

Number of strategies							
Without feedback. Spanish							
Year level	*N*	*M*	*SD*	*Mdn*	*IQR*	Year level comparison	*p***
K7	23	5.22	1.20	5.00	1.00	K7 - K10	.004
K10	22	4.41	0.80	4.00	1.00	K7 - K11	.300
K11	11	4.82	1.40	5.00	2.00	K10 - K11	.271
	*p**	.018					
With feedback. Spanish							
Year level	*N*	*M*	*SD*	*Mdn*	*IQR*	Year level comparison	*p***
K7	23	2.91	1.20	3.00	1.00	K7 - K10	-
K10	22	3.36	1.14	3.00	1.00	K7 - K11	-
K11	11	2.82	1.08	2.00	2.00	K10 - K11	-
	*p**	.124					
Without feedback. Mathematics							
Year level	*N*	*M*	*SD*	*Mdn*	*IQR*	Year level comparison	*p***
K7	23	3.30	1.26	3.00	2.00	K7 - K10	.005
K10	22	4.32	1.32	4.00	1.50	K7 - K11	.061
K11	11	3.91	0.83	4.00	2.00	K10 - K11	.489
	*p**	.014					
With feedback. Mathematics							
Year level	*N*	*M*	*SD*	*Mdn*	*IQR*	Year level comparison	*p***
K7	23	1.17	0.39	1.00	0.00	K7 - K10	-
K10	22	1.91	1.27	1.00	2.00	K7 - K11	-
K11	11	1.55	1.04	1.00	1.00	K10 - K11	-
	*p**	.064					

* *p* - values computed in a Kruskal-Wallis test.
***p* - values computed in a *post hoc* Mann-Whitney test. $\alpha_{corrected} = .017$.

Table 2

Number of strategies by feedback, subject matter and gender. Descriptive statistics and group comparisons

Number of strategies						
	Gender	*M*	*SD*	*Mdn*	*IQR*	*p**
Without feedback Spanish	Male	5.14	1.15	5.00	2.00	.071
	Female	4.63	1.11	4.00	1.00	
With feedback Spanish	Male	2.76	1.14	2.00	1.00	.030
	Female	3.26	1.15	3.00	1.00	
Without feedback Mathematics	Male	3.67	0.80	3.00	1.00	.267
	Female	3.91	1.50	4.00	2.00	
With feedback Mathematics	Male	1.24	0.44	1.00	0.50	.150
	Female	1.71	1.18	1.00	2.00	

Note: Male, $N = 21$; Female, $N = 35$.
* *p*- values computed in a Mann-Whitney test.

Type of Strategies. Importantly, all the tables in this section describe multiple dichotomous variables (as participants could enact more than one strategy). As such, the observations were not independent and statistical inference was not used. Therefore, only descriptive data is presented.

Four results were identified when exploring feedback occasion and subject (Table 3). First, participants enacted more types of strategies in Spanish. Second, there are four strategies that were shared between Spanish and mathematics: *read the question*, *read the response*, *perform the exercise again*, and *read/process the feedback received*. The similarities in frequency among the four types is striking, however there is one difference and that is the strategy: *read the question*. The use of this strategy decreased much less in mathematics as participants reviewed more frequently their understanding of the task once they had the feedback. Third, there were a large number of strategies specific to each subject, for example *Compare text – response* (Spanish) or *Replace the X* (mathematics) (Appendix A for more information). This indicates that the enactment of self-assessment strategies is influenced by domain knowledge. And, fourth, the majority of types of strategies decreased in frequency with feedback, some of them not being used at all; while two new strategies

were enacted with feedback: *Read/Process the feedback received* and, only in Spanish, *Compare own evaluation to feedback.*

Table 3

Strategies for the self-assessment, by feedback and subject matter

		Without Feedback	With Feedback
	Strategy	*N* (%)	*N* (%)
Spanish *N* = 64	Read the question	60 (93.8)	10 (15.6)
	Read the response	62 (96.9)	19 (29.7)
	Read the text	48 (75.0)	2 (3.1)
	Read/Process the feedback received	0 (0.0)	63 (98.4)
	Compare text-response	28 (43.8)	0 (0.0)
	Compare question-response	23 (35.9)	0 (0.0)
	Access their memory to compare	30 (46.9)	0 (0.0)
	Compare own evaluation to feedback	0 (0.0)	32 (50.0)
	Perform the exercise again	22 (34.4)	0 (0.0)
	Think of different responses	3 (4.7)	3 (4.7)
Mathematics *N* = 63	Read the question	55 (91.7)	23 (41.8)
	Read the response	23 (38.3)	11 (17.5)
	Read/Process the feedback received	0 (0)	62 (98.4)
	Review the signs	9 (15.0)	0 (0.0)
	Replace the X	4 (6.7)	1 (1.6)
	Evaluate the procedure followed	39 (65.0)	10 (15.9)
	Perform the exercise again	29 (48.3)	1 (1.6)

When we compared the year levels some interesting observations were extracted (Table 4). *Performing the exercise again* is mostly used among younger participants. We have considered this an advanced strategy, because it provides them with more accuracy on their self-assessment. However, the low numbers of K10 and K11 indicate that the participants might not need to perform the exercise again, especially in Spanish, to be able to self-assess. So, it might be a differential effect only occurring with younger students, in which they show different types of advanced strategies than the students in K10 and K11.

Other interesting differences were that K10 and K11 performed more: *Compare text-response*, *Review the signs*, and *Evaluate the procedure followed*; that *Read the response* in mathematics without feedback was most used by K10; or that *Compare question-response* and *Think of different responses* in Spanish without feedback was used more by K11. The results are then mixed and seem to indicate that self-assessment processes vary among year level being difficult to determine, at least with the current data analysis, which year level is more advanced.

Table 4

Type of strategies deployed, by feedback, subject matter, and year level

		Year level		
	Strategy	K7 *N* (%)	K10 *N* (%)	K11 *N* (%)
Without feedback Spanish *N* = 64	Read the question	27 (96.4)	22 (100)	11 (78.6)
	Read the response	28 (100)	22 (100)	12 (85.7)
	Read the text	22 (78.6)	18 (81.8)	8 (57.1)
	Compare text-response	6 (21.4)	15 (68.2)	7 (50.0)
	Compare question-response	8 (28.6)	6 (27.3)	9 (64.3)
	Access their memory to compare	15 (53.6)	8 (36.4)	7 (50.0)
	Perform the exercise again	19 (67.9)	3 (13.6)	0 (0.0)
	Think of different responses	0 (0.0)	0 (0.0)	3 (21.4)
With feedback Spanish *N* = 64	Read the question	6 (21.4)	3 (13.6)	1 (7.1)
	Read the response	5 (17.6)	10 (45.5)	4 (28.6)
	Read the text	2 (7.1)	0 (0.0)	0 (0.0)
	Read/Process the feedback received	27 (96.4)	22 (100)	14 (100)
	Compare own evaluation to feedback	12 (42.9)	15 (68.2)	5 (35.7)
	Think of different responses	1 (3.6)	1 (4.5)	1 (7.1)
Without feedback Mathematics *N* = 60	Read the question	23 (82.1)	21 (95.5)	11 (78.6)
	Read the response	3 (10.7)	18 (81.8)	2 (14.3)
	Review the signs	0 (0.0)	6 (27.3)	3 (21.4)
	Replace the X	1 (3.6)	2 (9.1)	1 (7.1)
	Evaluate the procedure followed	10 (35.7)	18 (81.8)	11 (78.6)
	Perform the exercise again	19 (67.9)	6 (27.3)	4 (28.6)
With feedback Mathematics	Read the response	5 (17.6)	3 (13.6)	3 (21.4)
	Read/Process the feedback received	26 (92.9)	22 (100)	14 (100)

N = 63	Replace the X	0 (0.0)	1 (4.5)	0 (0.0)
	Evaluate the procedure followed	0 (0.0)	7 (31.8)	3 (21.4)
	Perform the exercise again	0 (0.0)	1 (4.5)	0 (0.0)

The descriptive analyses of the types of strategies by gender (Table 5) shows that females employed three categories more frequently: *Compare own evaluation to feedback* in Spanish; *Read the response* in mathematics without feedback; and *Evaluate the procedure followed* with feedback. This last category was only performed by females. The first and the last strategies mentioned are advanced strategies used by females to a larger extent. Therefore, females seem to enact a more strategic self-assessment.

Table 5

Type of strategies deployed, by feedback, subject matter, and gender

		Gender	
	Strategy	Male *N* (%)	Female *N* (%)
Without feedback Spanish *N* = 64	Read the question	24 (92.3)	36 (94.7)
	Read the response	24 (92.3)	38 (100)
	Read the text	19 (73.0)	29 (76.3)
	Compare text-response	9 (34.6)	19 (50.0)
	Compare question-response	8 (30.7)	15 (39.4)
	Access their memory to compare	14 (53.8)	16 (42.1)
	Perform the exercise again	11 (42.3)	11 (28.9)
	Think of different responses	2 (7.6)	1 (2.6)
With feedback Spanish *N* = 64	Read the question	3 (11.5)	7 (18.4)
	Read the response	6 (23.1)	13 (34.2)
	Read the text	1 (3.8)	1 (2.6)
	Read/Process the feedback received	25 (96.2)	38 (100)
	Compare own evaluation to feedback	7 (26.9)	25 (65.8)
	Think of different responses	1 (3.8)	2 (5.2)
Without feedback Mathematics *N* = 60	Read the question	25 (96.2)	30 (78.9)
	Read the response	5 (19.2)	18 (47.3)
	Review the signs	3 (11.5)	6 (15.8)
	Replace the X	1 (3.8)	3 (7.9)
	Evaluate the procedure followed	14 (53.8)	25 (65.8)
	Perform the exercise again	14 (53.8)	15 (39.4)
With feedback Mathematics *N* = 63	Read the response	5 (19.2)	6 (15.8)
	Read/Process the feedback received	24 (92.3)	38 (100)
	Replace the X	1 (3.8)	0 (0.0)
	Evaluate the procedure followed	0 (0.0)	10 (26.3)
	Perform the exercise again	0 (0.0)	1 (2.6)

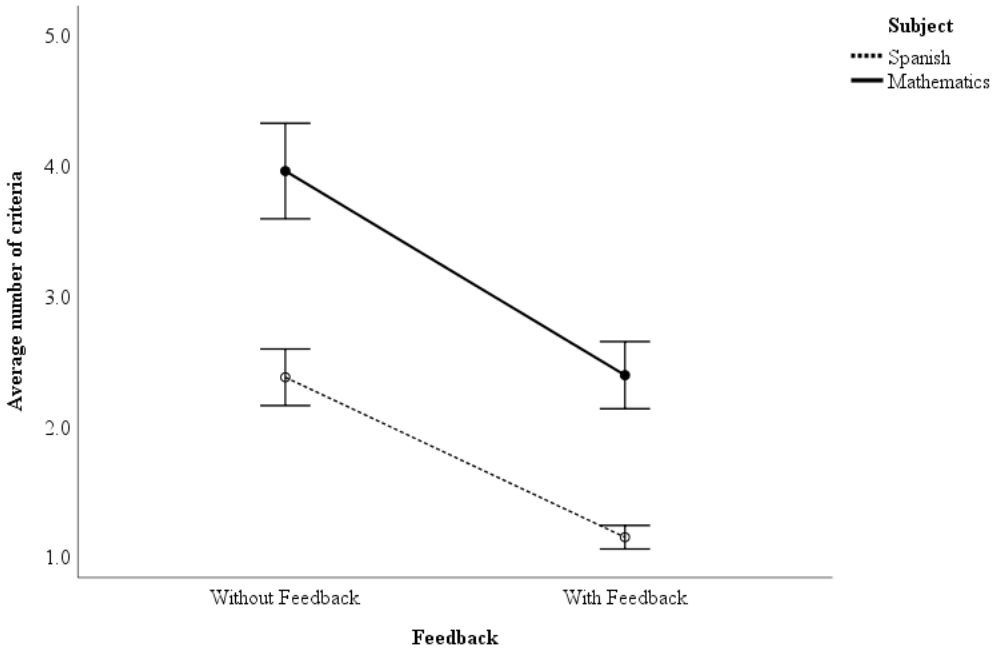
In conclusion, regarding our hypothesis (H1) self-assessment strategy data was supposed to be higher in number and more advanced types for Spanish (*partially maintained for number of strategies*), older students (*mixed results*), girls (*maintained*), and that with feedback the number and type of criteria would decrease (*maintained*).

RQ2. What are the effects of feedback, subject matter, year level, and gender on self-assessment criteria use (number and type)?

Number of Criteria. During the self-assessment of the Spanish task, the number of criteria used was lower (without feedback $M = 2.37$ $SD = 0.85$, with feedback $M = 1.15$ $SD = 0.36$) than during the mathematics task (without feedback $M = 3.95$ $SD = 1.44$, with feedback $M = 2.39$ $SD = 1.01$). A two-way repeated measures ANOVA was carried out to assess the effect of subject matter and feedback on the number of criteria adopted by the participants (see Figure 3). A significant effect for subject matter was found, with a lower number of criteria in Spanish ($M = 1.76$, $SD = 0.41$) than in mathematics ($M = 3.17$, $SD = 1.02$), $F(1, 61) = 110.35$, $p < .001$, $\eta^2 = .644$. Also, a significant effect for feedback was found, with more criteria used without feedback ($M = 3.16$, $SD = 0.91$) than with feedback ($M = 1.77$, $SD = 0.53$), $F(1, 61) = 133.68$, $p < .001$, $\eta^2 = .687$. No interaction effect was found, $F(1, 61) = 2.86$, $p = .096$.

Figure 3.

Number of criteria deployed, by feedback and subject matter. Error bars represent +/-2 SE.



As can be seen in Tables 6 and 7, no significant differences were found between year levels and gender when it came to the number of criteria used.

Table 6

Number of criteria by feedback, subject matter and year level. Descriptive statistics and group comparisons

	Number of criteria						
	Group	*N*	*M*	*SD*	*Mdn*	*IQR*	*p**
Without feedback, Spanish	K7	23	2.38	1.02	2.00	1.25	.992
	K10	22	2.36	0.73	2.00	1.00	
	K11	11	2.36	0.74	2.50	1.00	
With feedback, Spanish	K7	23	1.15	0.37	1.00	0.00	.588
	K10	22	1.09	0.29	1.00	0.00	
	K11	11	1.21	0.43	1.00	0.25	
Without feedback, Mathematics	K7	23	4.00	1.23	4.00	2.00	.159
	K10	22	3.45	1.14	3.00	1.25	
	K11	11	4.64	1.95	4.00	3.25	
With feedback, Mathematics	K7	23	2.42	0.99	2.50	1.00	.399
	K10	22	2.18	0.85	2.00	1.25	
	K11	11	2.64	1.28	3.00	2.25	

* *p* - values computed in a Kruskal-Wallis test.

Table 7

Number of criteria by feedback, subject matter and gender. Descriptive statistics and group comparisons

Number of criteria						
	Gender	*M*	*SD*	*Mdn*	*IQR*	*p**
Without feedback Spanish	Male	2.33	1.01	2.00	1.75	.720
	Female	2.39	0.75	2.00	1.00	
With feedback Spanish	Male	1.08	0.28	1.00	0.00	.229
	Female	1.18	0.39	1.00	0.00	
Without feedback Mathematics	Male	3.92	1.25	4.00	1.00	.684
	Female	3.97	1.57	4.00	2.00	
With feedback Mathematics	Male	2.33	1.05	2.00	1.75	.889
	Female	2.42	1.00	2.50	1.00	

Note: Male, *N* = 24; Female, *N* = 38.
* *p*- values computed in a Mann-Whitney test.

Types of Criteria. Performing statistical inference was not appropriate because the observations in this category were not independent. Therefore, we will only present data at a descriptive level.

Exploring feedback occasion and subject (Table 8) resulted in four findings. First, participants enacted more types of criteria in mathematics. Second, there were five types shared among Spanish and mathematics: *without clear criteria, based on intuition, based on hindsight, based on rules*, and *given by the teacher*. In comparison to strategy types, the frequencies in these five criteria actually have more differences: *intuition* was higher in Spanish, while *based on rules* was largely used in mathematics, along with *without clear criteria*. Third, there were a large number of strategies specific to each subject: two for Spanish and six for mathematics. The strategies used in mathematics demonstrated a high frequency of use. And fourth, the majority of types decreased in frequency with feedback, some of them not being used at all; while one new criterion was enacted with feedback: *given by the teacher,* which implied incorporating the teachers' feedback as a criterion.

Table 8

Type of criteria used, by feedback and subject matter

	Type of criteria		
	Criteria	Without Feedback N (%)	With Feedback N (%)
Spanish N = 63	Without clear criteria	2 (3.2)	1 (1.6)
	Based on intuition	40 (63.5)	1 (1.6)
	Based on hindsight	21 (33.3)	0 (0)
	Based on rules	26 (41.3)	9 (14.1)
	Based on spelling	6 (9.5)	0 (0)
	Given by the teacher	3 (4.8)	62 (96.9)
	Comparative criterion	37 (58.7)	0 (0)
Mathema-tics N = 64	Without clear criteria	6 (9.4)	0 (0)
	Based on intuition	21 (32.8)	0 (0)
	Based on hindsight	20 (31.3)	2 (3.2)
	Based on experience/self-efficacy	17 (26.6)	7 (11.1)
	Based on rules	45 (70.3)	1 (1.6)
	Based on mistakes identified	29 (45.3)	34 (54)
	Given by the teacher	1 (1.6)	60 (95.2)
	Based on adequacy to the question	15 (23.4)	6 (9.5)
	Based on the coherence of the result	34 (53.1)	12 (19)
	Based on steps followed	35 (54.7)	15 (23.8)
	Based on formula application	27 (42.2)	15 (23.8)

Table 9 shows the interesting differences between year levels. K7 participants used more *based on rules*; K10 participants used more *based on intuition*, *comparative criterion*, and *based on followed steps*, but less *based on experience/self-efficacy*; and K11 participants used more *based on hindsight* and *based on formula application,* but less *given by the teacher*. Nevertheless, the general picture seems to be that the year levels are balanced when it comes to the types of criteria used. Of course, frequencies might vary as there is a high number of criteria, but they are more similar than different.

Table 9

Type of criteria adopted, by feedback, subject matter, and year level

		Group		
	Criteria	K7 N (%)	K10 N (%)	K11 N (%)
Without feedback Spanish N = 63	Without clear criteria	0 (0.0)	1 (4.5)	1 (7.1)
	Based on intuition	14 (50.0)	20 (90.9)	6 (42.9)
	Based on hindsight	10 (35.7)	6 (27.3)	5 (35.7)
	Based on rules	24 (85.7)	1 (4.5)	1 (7.1)
	Based on spelling	2 (7.1)	1 (4.5)	3 (21.4)
	Given by the teacher	0 (0.0)	2 (9.1)	1 (7.1)
	Comparative criterion	10 (35.7)	19 (86.4)	8 (57.1)
With feedback Spanish N = 64	Without clear criteria	0 (0.0)	0 (0.0)	1 (7.1)
	Based on intuition	0 (0.0)	0 (0.0)	1 (7.1)
	Based on rules	4 (14.3)	2 (9.1)	3 (21.4)
	Given by the teacher	28 (100)	22 (100)	12 (85.7)
Without feedback Mathema-tics N = 64	Without clear criteria	4 (14.3)	1 (4.5)	1 (7.1)
	Based on intuition	7 (25.0)	11 (50.0)	3 (21.4)
	Based on hindsight	5 (17.9)	4 (18.2)	11 (78.6)
	Based on experience/self-efficacy	10 (35.7)	2 (9.1)	5 (35.7)
	Based on rules	21 (75.0)	14 (63.6)	10 (71.4)
	Based on mistakes identified	14 (50.0)	8 (36.4)	7 (50.0)
	Given by the teacher	0 (0.0)	1 (4.5)	0 (0.0)
	Based on the coherence of the result	18 (64.3)	10 (45.5)	6 (42.9)
	Based on steps followed	10 (35.7)	18 (81.8)	7 (50.0)
	Based on formula application	13 (46.4)	4 (18.2)	10 (71.4)
	Based on adequacy to the question	7 (25.0)	3 (13.6)	5 (35.7)
With feedback Mathema-tics N = 63	Based on hindsight	0 (0.0)	0 (0.0)	2 (14.3)
	Based on experience/self-efficacy	4 (14.3)	1 (4.5)	2 (14.3)
	Based on rules	1 (3.6)	0 (0.0)	0 (0.0)
	Based on mistakes identified	14 (50.0)	15 (68.2)	5 (35.7)
	Given by the teacher	26 (92.9)	22 (100)	12 (85.7)
	Based on adequacy to the question	4 (14.3)	1 (4.5)	1 (7.1)
	Based on the coherence of the result	8 (28.6)	3 (13.6)	1 (7.1)
	Based on steps followed	6 (21.4)	4 (18.2)	5 (35.7)
	Based on formula application	4 (14.3)	2 (9.1)	9 (64.3)

Regarding gender (Table 10) both seemed to enact similar criteria. There are only three categories that show larger differences: two in favor of females (i.e., without feedback

in mathematics *based on intuition* and *based on followed steps*) and one for males (i.e., without feedback in Spanish *based on rules*). Nevertheless, in regard to year levels, males and females are more similar than different

Table 10

Type of criteria adopted, by feedback, subject matter, and gender

		Gender	
	Criteria	Male *N* (%)	Female *N* (%)
Without feedback Spanish N = 63	Without clear criteria	1 (3.8)	1 (2.6)
	Based on intuition	16 (61.5)	24 (63.2)
	Based on hindsight	7 (26.9)	14 (36.8)
	Based on rules	14 (53.8)	12 (31.6)
	Based on spelling	2 (7.7)	4 (10.5)
	Given by the teacher	1 (3.8)	2 (5.3)
	Comparative criterion	13 (50.0)	24 (63.2)
With feedback Spanish N = 64	Without clear criteria	1 (3.8)	0 (0.0)
	Based on intuition	1 (3.8)	0 (0.0)
	Based on rules	2 (7.7)	7 (18.4)
	Given by the teacher	24 (92.3)	38 (100)
Without feedback Mathematics N = 64	Without clear criteria	4 (15.4)	2 (5.3)
	Based on intuition	5 (19.2)	16 (42.1)
	Based on hindsight	8 (30.8)	12 (31.6)
	Based on experience/self-efficacy	7 (26.9)	10 (26.3)
	Based on rules	17 (65.4)	28 (73.7)
	Based on mistakes identified	13 (50.0)	16 (42.1)
	Given by the teacher	1 (3.8)	0 (0.0)
	Based on adequacy to the question	7 (26.9)	8 (21.1)
	Based on the coherence of the result	14 (53.8)	20 (52.6)
	Based on followed steps	11 (42.3)	24 (63.2)
	Based on formula application	12 (46.2)	15 (39.5)
With feedback Mathematics N = 63	Based on hindsight	0 (0.0)	2 (5.3)
	Based on experience/self-efficacy	2 (7.7)	5 (13.2)
	Based on rules	1 (3.8)	0 (0.0)
	Based on mistakes identified	12 (46.2)	22 (57.9)
	Given by the teacher	24 (92.3)	36 (94.7)
	Based on adequacy to the question	3 (11.5)	3 (7.9)
	Based on the coherence of the result	4 (15.4)	8 (21.1)
	Based on followed steps	7 (26.9)	8 (21.1)
	Based on formula application	7 (26.9)	8 (21.1)

In conclusion, regarding our hypothesis (H2) self-assessment criteria was supposed to be higher in number and more advanced types for Spanish (*rejected*), older students (*rejected*), girls (*rejected*), and that with feedback the number and type of criteria would decrease (*maintained*).

RQ3. What are the effects of feedback, subject matter, year level, and gender on self-efficacy?

First, we analyzed the effects of feedback on occasion and subject (Table 11). At a descriptive level, self-efficacy increased over time in Spanish, while it decreased in mathematics. A two-ways repeated measures ANOVA was conducted to assess the effects of subject matter and feedback on self-efficacy scores. No significant effect was found for either subject matter, $F(1, 63) = 1.99$, $p = .162$; or feedback, $F(2, 62) = 1.13$, $p = .330$. An interaction effect between both factors on self-efficacy scores was found, $F(2, 62) = 6.69$, $p = .002$, $\eta^2 = .178$. This interaction can be seen in Figure 4. Before the self-assessment, there was a significant difference in self-efficacy scores between Spanish and mathematics ($p = .001$). This difference disappeared once the students self-assessed on both occasions, without feedback ($p = .092$) and with feedback ($p = .692$). Furthermore, the pre self-assessment simple effect (i.e., Spanish and mathematics means difference) was different from the pre self-assessment without feedback ($p = .002$) and with feedback ($p = .012$) simple effects. However, there was no difference between without feedback and with feedback simple effects ($p = .135$). That is, the self-efficacy scores were higher in mathematics than in Spanish before the task. This difference disappeared with the self-assessment and did not change with feedback. Finally, as can be seen in Tables 12 and 13, there were no significant differences based on year level or gender.

Table 11

Average Self-efficacy scores, by feedback and subject matter

	M	*SD*
Pre self-assessment. Spanish	4.65	0.75
Without feedback. Spanish	4.78	0.83
With feedback. Spanish	5.15	2.04
Pre self-assessment. Mathematics	5.20	1.08
Without feedback. Mathematics	5.04	1.12
With feedback. Mathematics	5.03	1.21

Note: $N = 64$

Table 12

Self-efficacy scores by feedback, subject matter and year level. Descriptive statistics and group comparisons

	Year level	*M*	*SD*	*Mdn*	*IQR*	*p**
Pre self-assessment Spanish	K7	4.67	0.76	4.71	1.25	.664
	K10	4.53	0.69	4.44	1.03	
	K11	4.79	0.85	4.60	0.75	
Without feedback Spanish	K7	4.88	0.85	4.86	0.96	.515
	K10	4.61	0.75	4.63	0.75	
	K11	4.81	0.94	4.50	0.85	
With feedback Spanish	K7	5.13	0.93	5.14	1.44	.211
	K10	4.69	0.68	4.81	0.91	
	K11	5.89	4.08	4.70	1.45	
Pre self-assessment Mathematics	K7	5.25	0.96	5.30	1.30	.062
	K10	5.55	0.78	5.67	1.17	
	K11	4.56	1.44	4.67	1.88	
Without feedback Mathematics	K7	5.20	0.96	5.20	1.60	.453
	K10	5.14	0.92	5.17	1.33	
	K11	4.56	1.57	4.75	1.96	
With feedback Mathematics	K7	5.26	1.01	5.40	1.55	.280
	K10	5.10	1.15	5.00	1.46	
	K11	4.48	1.56	4.75	1.63	

Note: K7, $N = 28$; K10, $N = 22$; K11, $N = 14$.
* *p*- values computed in a Kruskal-Wallis test.

Table 13

Self-efficacy scores by feedback, subject matter and gender. Descriptive statistics and group comparisons

Self-efficacy scores						
	Gender	*M*	*SD*	*Mdn*	*IQR*	*p**
Pre self-assessment Spanish	Male	4.72	0.70	4.55	0.73	.613
	Female	4.60	0.79	4.59	1.21	
Without feedback Spanish	Male	4.81	0.75	4.71	0.65	.816
	Female	4.75	0.89	4.67	1.10	
With feedback Spanish	Male	5.46	3.01	5.00	1.15	.827
	Female	4.93	0.90	4.94	1.25	
Pre self-assessment Mathematics	Male	5.19	1.19	5.30	1.39	.913
	Female	5.22	1.01	5.27	1.17	
Without feedback Mathematics	Male	5.07	1.24	5.20	0.93	.598
	Female	5.02	1.05	5.08	1.42	
With feedback Mathematics	Male	5.10	1.30	5.17	1.53	.552
	Female	4.98	1.17	5.00	1.71	

Note: Male, $N = 26$; Female, $N = 38$.
* *p*- values computed in a Mann-Whitney test.

Our hypothesis for the third research question, therefore, needs to be fully rejected. There were no differences in feedback occasion, subject, year level, or gender. The only significant effect was the interaction between subject and feedback occasion, where we found a higher self-efficacy mean score in the pre self-assessment measurement for mathematics.

Finally, for better interpretability of results, Table 14 summarizes the three-research question outputs.

Table 14

Summary of Results

Dependent variable		**Feedback**	**Subject matter**	**Year level**	**Gender**
Strategies	Number	− With feedback	+ Spanish	+ K7 Spanish without feedback + K10 Mathematics without feedback	+ Females (Only in Spanish with feedback)
	Type	− With feedback but a pair of new strategies are used	Different strategies for Spanish and mathematics	Mixed results	+ Females
Criteria	Number	− With feedback	+ Mathematics	=	=
	Type	− With feedback The use of subject-specific criteria maintained in mathematics	+ Mathematics Different criteria for Spanish and mathematics	More similarities than differences	=
Self-efficacy		=		=	=
		No significant differences, except in the interaction at the "pre-self-assessment" occasions were mathematics had higher values			

Note: "–" indicates lower number or less advanced type. "+" indicates higher number or more advanced type. "=" indicates no significant differences.

Discussion

Our aim was to explore the effects of feedback, subject matter, year level, and gender on self-assessment and self-efficacy. For this, we used an innovative method disentangling self-assessment into more specific actions to better understand how students actually self-assess. Though, our hypotheses are only partially maintained, we believed this study is relevant as it helps to elucidate what happens in the "black box" of self-assessment.

The Effects of the Four Factors on Self-assessment

These effects were explored in the first and second research questions. For clarity we discuss the results of the two research questions combined and organized around the four factors.

Regarding the effects of feedback, though there has been a number of calls for implementing feedback in self-assessment interventions (e.g. Andrade & Valtcheva, 2009; Lipnevich & Smith, 2018; Ross, 2006) this crucial educational element is vastly absent from research. Here, we found a clear pattern: once students have the feedback their use of self-assessment strategies and criteria decreased severely. Results show that after receiving it, using the feedback became the main strategy and criteria. This is "logical" from the participants' side: following the teachers' feedback they can self-evaluate more efficiently and with higher certitude. This aligns with the model by Butler and Winne (1995) as students used the external feedback to evaluate their work even when just asked to self-assess. From the teachers' side: this indicates that it might be positive to postpone our feedback until students have had the chance to self-assess on their own. It is important to emphasize that our study might suffer from a recency bias as our participants first self-assessed without feedback and then with feedback. Therefore, as they had just self-assessed most likely they felt they did not need to be so thorough a second time. However, it could also be the case that, if we would have given them the feedback during the first self-assessment, students would mainly

pay attention to the feedback information instead of trying to self-assess by themselves. Actually, the data points in that direction as, once students had the feedback, we identified a couple of new strategies based on the feedback itself which became the most frequent ones. When comparing our results with previous research (Panadero et al., 2012; Raaijmakers et al., 2019), it is difficult to draw conclusions as none of those studies explored the effects of feedback on the self-assessment process itself. Therefore, our study has opened an interesting avenue of research on the implications of feedback on self-assessment.

Secondly, regarding the effects of subject matter, it is known that reaching an appropriate balance between domain-general vs domain-specific in assessment practices is complicated (Wiliam, 2019). It is also known that different disciplines used distinct assessment (Panadero et al., 2019), and in the same line of reasoning it has been argued that tasks with a different nature might require distinct self-assessment (Panadero et al., 2016). Our results seem to confirm such line of reasoning as our participants showed distinct profiles for each subject: more strategies for Spanish and a richer range of criteria for mathematics. The differences seem to be related to the specifics of each subject. Obviously, Spanish and mathematics activities have different nature requiring different learning and pedagogical strategies (De Corte et al., 2011; Harris et al., 2011). Therefore, the comparison is also not straightforward when it comes to self-assessment strategies and criteria. For example, in Spanish the most used categories were in relationship to reading a text, which is usually a task in Spanish; or in mathematics students needed to check the signs in the operation, a criterion not needed for Spanish. Falchikov and Boud (1989) found that more accurate self-assessment occurred in the science area, probably because of the nature of the task itself and also because the more well-defined the solution to a task (i.e. usually the case for mathematics) the more accurate the students are (Dunning et al., 2004). This clearly shows that we cannot separate self-assessment from content based knowledge and the type of task

(Panadero et al., 2016): we need to explore students' self-assessment strategies along with the specific demands for the subject matter, which is an aspect majorly overlooked in the self-assessment literature. An important limitation is that the presentation of the subjects' self-assessment was not counterbalanced (i.e. Spanish was first and mathematics second). All in all, it seems reasonable to conclude from our results, that we need to carefully consider domain knowledge in our interventions to tackle the specifics of self-assessment processes. To our knowledge, our study is the first directly comparing two different subjects' self-assessment processes, thus we cannot relate our specific results to existing studies.

Thirdly, regarding year level, historically the field has explored what type of students benefit the most from self-assessment claiming sometimes the low achievers (Ross et al., 1999) others the average students (Boud, Lawson & Thompson, 2013 & 2015), and also analyzing which students are more accurate which was explored in detail by Dunning et al. (2004). However, when it comes to the quality of the self-assessment itself, it has been argued that more mature students would have higher capacity than younger students (Panadero et al., 2016). This should not to be confused with having more experience (i.e. domain expertise) in a particular task, which is related but not necessarily the same thing. For example, a participant of K11 would have had much more experience on fractions than one from K7, which would have affected our measurement. For that reason, we selected the tasks so that the participants had similar levels of practice across the year levels to counteract the domain expertise effect. If we would have compared the three year levels in the exact same task, most likely the results have been clearly in favored of the older students as they are more experienced (e.g. Boud & Falchikov, 1989; Dochy et al., 1999). Additionally, our coding categories were created checking for their applicability in all year levels' activities. Therefore, we level the participants in the expertize to analyze is older students would have developed more advanced self-assessment skills.

Our results on the year level factor lead to two conclusions. First, in general the three year levels are showing more similarities than differences, especially in mathematics. And second, when those differences appear it is difficult to identify which students are more advanced because it seems like younger students might use different strategies and criteria that could be advanced for them while older students use others. This conflicts with the results from Yan (2018), where he found senior students reporting lower self-assessment practices in a survey. This difference in results might be caused by Yan use of a survey in which students might be biased or simply conceptualize self-assessment differently than we did here. However, our data is based in the actual self-assessment actions and a strong triangulation of data, so our results seem to be stronger. This difference between self-reported data and factual actions has been found in self-regulated learning (Winne, 2020) and it is not new to self-assessment research either (Panadero et al., 2012, 2017). Additionally, another hypothesis is, as our data shows, there are qualitative differences in what students from different year levels enact when self-assessing that it would be difficult for a survey to identify them; though, again, the self-assessments are more similar than different.

Lastly, regarding gender, our results showed that females use more number and types of strategies. The effects of gender on self-assessment research has received some attention (e.g. Bolivar-Cruz & Verano-Tacoronte, 2018; Langan et al., 2008). However, to our knowledge, the actual effects of gender on the self-assessment specific processes has only been studied in Yan (2018). Here, it was found that females reported higher levels of self-assessment practices. Again, Yan (2018) used self-reported data while we analyzed real self-assessment actions. Considering all the above-mentioned studies, our data, and an existent meta-analysis (Panadero et al., 2017), females seem to be both more advanced and obtain better benefits from self-assessment.

Effects on Self-efficacy

Our third research question explored how self-efficacy developed at three different times: before self-assessment, after self-assessment without feedback, and after self-assessment with feedback. Contrary to our hypothesis, there were no significant differences in the four factors' main effects, and only the interaction subject x feedback occasion resulted in a significant effect. Importantly, academic self-efficacy is known as a stable factor that builds up with years of academic experience (Bandura, 1997). We hypothesized that changes might occur because, right after self-assessing, the participants received a score and this might have conflicted with what they just said (e.g. stating that the task was excellently performed and then receiving a low score). However, we will need to dig deeper into our data to extract more conclusions about self-efficacy.

Comparing previous research, starting by subject, Huang's (2013) meta-analysis, also exploring gender effects, found that females showed higher self-efficacy in language and arts than males, while the latter showed higher self-efficacy in mathematics, computer, and social sciences. Here neither of these two main effects, subject and gender, were found. This might be explained by the fact that Huang (2013) found a small effect size of gender in the first place and our sample size is small in comparison with his meta-analysis. Additionally, we did not find the positive effect of self-assessment practice in females found in Panadero et al.' (2017) meta-analysis. However, this is easier to explain as we did not really intervene, but just asked our participants to self-assess. Regarding year level, the general tendency is that academic self-efficacy decreases as learners' progress through school (Pajares, 2008). This decline has been especially found through the middle school years (Pajares, 2008), which are different ages to the ones explored here and might explain the lack of differences in our study. Nevertheless, Huang's meta-analysis (2013) found the largest effect size for gender differences for students older than 23 years old. Finally, regarding the significant difference found in the interaction subject per feedback occasion before the self-assessment, it seems

that the participants were over-confident for mathematics while being under-confident for Spanish. To our knowledge, there are no previous studies to compare the results to.

Limitations and Future Lines of Research

One important limitation of our design is that we did not counterbalance our study. First, all participants self-assessed first without feedback and then with feedback. It might be the case that if feedback had been given within the first self-assessment occasion, the results might have been different (e.g., participants being less directed by feedback as the main strategy and criteria). And second, we did not counterbalance the subject matter. The Spanish self-assessment was always performed first, followed by the mathematics assessment. Because of this, we cannot reject that differences attributed to subject/feedback might be simply due to recency or practice effects (e.g., students becoming more confident at thinking aloud or fatigued).

A second limitation relates to the use of thinking aloud protocols. These are affected by cognitive load in both, the verbalization of the protocols and the performance of the task. Students could have greater difficulty verbalizing their self-assessment of mathematics because of the greater cognitive load associated with doing mathematics.

A third limitation deals with the limited time span in which we measured the changes in self-efficacy. The scale was administered five times within 45 minutes, which seems implausible that momentary changes would become long-term. However, if they did, it would be problematic because it would suggest a carry-over effect for the next time the scale was administered.

A final limitation is that the feedback students received was purely based on the task, not on the self-assessment. Though this was a research decision, it is known that self-assessment scholars recommend providing feedback on the self-assessment to enhance its quality (e.g., Andrade, 2018). Therefore, the strategies used by the participants once they had

the feedback could have been an *echo* of the feedback they received that was mostly at the task-level. However, research shows that most of the feedback delivered in the classroom is at such level (e.g. Hattie & Timperley, 2007). Thus, our study replicates the most usual classroom setting in which students are usually asked to self-assess but they receive feedback on the properties of the task, not on the self-assessment.

In terms of future lines of research, this study is opening a new area. More studies will be needed on the specific enactment of self-assessment processes, as ours has mostly focused on analyzing the strategies and criteria using a wider range of variables. For example, the emotional and motivational reactions while performing self-assessment need to be explored. Also, the use of measurement technology such as eye-tracking (Jarodzka, Holmqvist, & Gruber, 2017) would allow for a more precise understanding of the underlying mental processes. This technology, in combination with physiological reaction equipment (Azevedo, Taub, & Mudrick, 2018) would provide a complete picture in the understanding of self-assessment.

In addition to the use of different measurement, it would be important to further explore the four factors investigated here. Our results show that the subject, feedback occasion, year level, and gender have important effects on self-assessment practice, yet these effects have been barely studied (Panadero et al., 2016). Finally, our research aim should be expanded to other educational levels (e.g., primary and higher education).

Educational Implications

This study has a number of implications. First, knowing the specific self-assessment strategies and criteria will help to develop more strategic interventions. These should be based on the specific aspects of the task-domain (e.g., formula application) on top of more generic strategies. Second, feedback should be given to the students once they have had the chance to self-assess without it. As observed here, once the feedback is delivered, students

radically change their self-assessment strategies and criteria, and the feedback content monopolizes the self-assessment. Third, attention needs to be given to the students' year level; younger students might need further guidance, especially directed towards the activation of specific task-domain strategies and criteria. Finally, male students need further support on self-assessment as their strategies are less advanced and used less frequently.

Conclusion

Our study has explored two new avenues of research: exploring the effects of four different factors on self-assessment and using an innovative data collection and analysis to identify specific processes involved in self-assessment. With these data, we were better able to trace what students actually do when self-assessing. By breaking self-assessment down to specific actions instead of considering it a unified strategy, we will be able to design and implement more specific interventions. We now know that (1) feedback delivery might impede self-assessment enactment; (2) that different subject matters lead to different self-assessment specific processes; (3) that students from different year levels show distinct use of self-assessment strategies and criteria according to their level of domain, so we cannot maintain that older students are best at self-assessment; and (4) females seem to be more strategic self-assessors. Our study has been innovative investigating the black box of self-assessment employing new methods and combining four factors; but as with any other study we now have new questions that need to be addressed. Our belief is that this new line of research and the use of these methods will help us answer much needed questions about self-assessment.

References

Andrade, H. (2018). Feedback in the context of self-assessment. In A. A. Lipnevich & J. K. Smith (Eds.), *The Cambridge handbook of instructional feedback* (pp. 376-408): Cambridge University Press.

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory Into Practice, 48*(1), 12-19. doi:10.1080/00405840802577544

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). New York: Routledge.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W. H. Freeman and Company.

Barnett, J. E., & Hixon, J. E. (1997). Effects of grade level and subject on student test score predictions. *The Journal of Educational Research, 90*(3), 170-174. doi:10.1080/00220671.1997.10543773

Bolívar-Cruz, A., Verano-Tacoronte, D., & Galván-Sánchez, I. (2018). Do self-efficacy, incentives and confidence in public speaking influence how students self-assess?/¿ Influyen la autoeficacia, los incentivos y la confianza para hablar en público en cómo se autoevalúan los estudiantes?. *Cultura y Educación, 30*(3), 528-555. doi:10.1080/11356405.2018.1488420

Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher-education: A critical analysis of findings. *Higher Education, 18*(5), 529-549. doi:10.1007/BF00138746

Boud, D., Lawson, R., & Thompson, D. G. (2013). Does student engagement in self-assessment calibrate their judgement over time? *Assessment & Evaluation In Higher Education, 38*(8), 941-956. doi:10.1080/02602938.2013.769198

Boud, D., Lawson, R., & Thompson, D. G. (2015). The calibration of student judgment through self-assessment: Disruptive effects of assessment patterns. *Higher Education Research & Development, 34*(1), 45-59. doi:10.1080/07294360.2014.934328

Brown, G. T., Andrade, H. L., & Chen, F. (2015). Accuracy in student self-assessment: directions and cautions for research. *Assessment in Education: Principles, Policy & Practice*, *22*(4), 444-457. doi:10.1080/0969594X.2014.996523

Brown, G., & Harris, L. R. (2013). Student self-assessment. *SAGE handbook of research on classroom assessment*.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of educational research*, *65*(3), 245-281. doi:10.3102/00346543065003245

Butler, R., & Hasenfratz, L. (2017). Gender and competence motivation. In A. J. Elliot, C. S. Dweck, & D. S. Yeager (Eds.), *Handbook of competence and motivation* (2nd. edition ed., pp. 489-511). New York: The Guilford Press.

Creswell, J. W., & Clark, V. L. P. (2018). *Designing and conducting mixed methods research*. Sage publications.

De Corte, E., Mason, L., Depaepe, F., & Verschaffel, L. (2011). Self-regulation of mathematical knowledge and skills. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 155-172). New York: Routledge.

Dochy, F. J. R. C., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher education*, *24*(3), 331-350. doi:10.1080/03075079912331379935

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological science in the public interest*, *5*(3), 69-106. doi:10.1111/j.1529-1006.2004.00018.x

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of educational research*, *59*(4), 395-430. doi:10.3102/00346543059004395

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. doi:10.3102/003465430298487

Harris, K. R., Graham, S., MacArthur, C. A., Reid, R., & Mason, L. H. (2011). Self-regulated learning processes and children's writing. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 187-202). New York: Routledge.

Harris, L. R., & Brown, G. T. (2018). *Using self-assessment to improve student learning*. Routledge.

Huang, C. J. (2013). Gender differences in academic self-efficacy: A meta-analysis. *European Journal of Psychology of Education, 28*(1), 1-35. doi:10.1007/s10212-011-0097-y

Jarodzka, H., Holmqvist, K., & Gruber, H. (2017). Eye tracking in educational science: Theoretical frameworks and research agendas. *Journal of Eye Movement Research, 10*(1), 1-18.

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. doi:10.1002/acp.1528

Kostons, D., van Gog, T., & Paas, F. (2010). Self-assessment and task selection in learner-controlled instruction: Differences between effective and ineffective learners. *Computers & Education*, *54*(4), 932-940. doi:10.1016/j.compedu.2009.09.025

Langan, A. M., Shuker, D. M., Cullen, W. R., Penney, D., Preziosi, R. F., & Wheater, C. P. (2008). Relationships between student characteristics and self-, peer and tutor evaluations of oral presentations. *Assessment & Evaluation in Higher Education*, *33*(2), 179-190. doi:10.1080/02602930701292498

Lipnevich, A. A., & Smith, J. K. (2018). *The Cambridge Handbook of Instructional Feedback*: Cambridge University Press.

Panadero, E., Alonso-Tapia, J., & Huertas, J. A. (2012). Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education. *Learning and Individual Differences*, *22*(6), 806-813. doi:10.1016/j.lindif.2012.04.007

Panadero, E., Brown, G. T. L., & Strijbos, J. W. (2016). The future of student self-assessment: A review of known unknowns and potential directions. *Educational Psychology Review*, *28*(4), 803-830. doi:10.1007/s10648-015-9350-2

Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review, 22*, 74-98. doi:https://doi.org/10.1016/j.edurev.2017.08.004

Panadero, E., Fraile, J., Fernández Ruiz, J., Castilla-Estévez, D., & Ruiz, M. A. (2019). Spanish university assessment practices: examination tradition with diversity by faculty. *Assessment & Evaluation In Higher Education*, *44*(3), 379-397. doi:10.1080/02602938.2018.1512553

Pajares, F. (2008). Motivational role of self-efficacy beliefs in self-regulated learning. In D. H. Schunk & B. J. Zimmerman (Eds.), *Motivation and self-regulated learning. Theory, research and applications* (pp. 111-168). New York: Lawrence Erlbaum Associates.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 452-502). San Diego, CA: Academic Press.

Raaijmakers, S. F., Baars, M., Paas, F., van Merriënboer, J. J., & van Gog, T. (2019). Effects of self-assessment feedback on self-assessment and task-selection accuracy. *Metacognition and Learning*, 1-22. doi:10.1007/s11409-019-09189-5

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*(2), 353-387. doi:10.1037/a0026838

Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment Research & Evaluation, 11*. Retrieved from http://pareonline.net/getvn.asp?v=11&n=10

Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (1999). Effects of self-evaluation training on narrative writing. *Assessing Writing, 6*(1), 107-132. doi:10.1016/S1075-2935(99)00003-3

Wiliam, D. (2019). Why Formative Assessment is Always Both Domain-General and Domain-Specific and What Matters is the Balance Between the Two. In H. Andrade, R. E. Bennett, & G. J. Cizek (Eds.), *Handbook of formative assessment in the disciplines*. New York: Routledge.

Winne, P. (2020). A proposed remedy for grievances about self-report methodologies. *Frontline Learning Research, 8*(3), 164 - 173. doi:https://doi.org/10.14786/flr.v8i3.625

Yan, Z. (2018). The Self-assessment Practice Scale (SaPS) for Students: Development and Psychometric Studies. *The Asia-Pacific Education Researcher*, *27*(2), 123-135. doi:10.1007/s40299-018-0371-8

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. doi:10.1080/02602938.2016.1260091

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13-40). San Diego, California: Academic Press.

Appendix A

Category description and examples

Type	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 Criteria based on simple rules	**Based on rules**	Based on specific rules from the subject	*"The first one I think is an objective description, because the author is speaking"*
		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"*

***Note*. * Indicates that category was only used during the mathematics task but it could also be used for Spanish. ** Indicates that category was only used during the Spanish task but it could also be used for mathematics.**