University students' strategies and criteria during self-assessment: Instructor's feedback, rubrics, and year level effects

Ernesto Panadero[1 2], Daniel García Pérez[3], Javier Fernández Ruiz[4], Juan Fraile[5], Iván Sánchez-Iglesias[6] & Gavin T. L. Brown[7]

[1] Facultad de Psicología y Educación, Universidad de Deusto, Bilbao, España.
[2] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain.
[3] Universidad Europea de Madrid, Spain.
[4] Departamento de Psicología Evolutiva y de la Educación, Universidad Autónoma de Madrid, Spain.
[5] Universidad Francisco de Vitoria, Spain.
[6] Departamento de Psicobiología y Metodología de las Ciencias del Comportamiento, Universidad Complutense de Madrid, Spain.
[7] Faculty of Education & Social Work, The University of Auckland, New Zealand.

Funding:
Research funded by: 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: Javier Fernández Ruiz. Aula PDIF, Facultad de Psicología. Universidad Autónoma de Madrid, Madrid. 28049. Spain. E-mail: javier@fernandezruiz.com

Recommended citation: Panadero, E., Pérez, D. G., Ruiz, J. F., Fraile, J., Sánchez-Iglesias, I., & Brown, G. T. (2022). University students' strategies and criteria during self-assessment: Instructor's feedback, rubrics, and year level effects. *European Journal of Psychology of Education*, 1-21. DOI: 10.1007/s10212-022-00639-4

This is a post-peer-review, pre-copyedit version of an article published in European Journal of Psychology of Education. The final authenticated version is available online at: https://doi.org/10.1007/s10212-022-00639-4. This manuscript may not exactly replicate the published version due to editorial changes and/or formatting and corrections during the final stage of publication. Interested readers are advised to consult the published version.

The authors declare that they have no conflict of interest.

Abstract

This study explores the effects of feedback type, feedback occasion and year level on student self-assessments in higher education. In total, 126 university students participated in this randomized experiment under three experimental conditions (i.e., rubric feedback, instructor's written feedback, and rubric feedback plus instructor's written feedback). Participants, after random assignment to feedback condition, were video-recorded performing a self-assessment on a writing task both before and after receiving feedback. The quality of self-assessment strategies decreased after feedback of all kinds, but the number of strategies increased for the combined feedback condition. The number of self-assessment criteria increased for rubric and combined conditions, while feedback helped shift criteria use from basic to advanced criteria. Student year level was not systematically related to changes in self-assessment after feedback. In general, the combination of rubric and instructor's feedback produced the best effects.

Keywords: self-assessment; feedback effects; rubric; higher education.

University students' strategies and criteria during self-assessment: instructor's feedback, rubrics, and year level effects

Self-assessment of learning is linked to greater self-regulation (Andrade, 2018; Yan, 2019) and achievement (Brown & Harris, 2013). Further, the ability to evaluate one's own work and processes is an important objective of higher education (Tai et al., 2017). However, our understanding of how students integrate feedback within their self-assessment processes is limited (Panadero et al., 2016), though we have a considerable knowledge on how feedback concerning task, process, and self-regulatory processes has been shown to improve educational outcomes (Butler & Winne, 1995; Hattie & Timperley, 2007). In one of the few studies exploring self-assessment and external feedback, Yan and Brown (2017) showed in an interview study with teacher education students that students claim to seek external feedback to form a self-assessment. Hence it is important to understand how to support the development of realistic and sophisticated self-assessment. A successful formative assessment practice has been the introduction of rubrics or scoring guides into classroom practice (Brookhart & Chen, 2015). Hence, it was expected that students would describe more complex self-assessment processes when provided feedback based on a rubric.

In a randomized experiment with university students, this study systematically extends our understanding of the role feedback plays on self-assessment by manipulating the type of feedback, its timing, and the expertise level of tertiary students. The study extends our understanding of the self-assessment "black box" by examining the strategies and criteria students used. Hence, this study provides new insights into how robust self-assessment can be supported.

Self-assessment

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). This definition indicates that self-assessment can take different shapes, from self-grading (e.g. Falchikov & Boud, 1989) to formative approaches (e.g. Andrade, 2018). However, what exactly happens when students self-assess is still largely mysterious.

Yan and Brown (2017) interviewed 17 undergraduate students from a teacher education institute using six general learning scenarios (e.g., *How good are you at learning a new physical skill*?) and five questions specific to self-assessment (e.g., *What criteria did you use to conduct self-assessment*?). From that data, the authors built a schematic cyclical self-assessment process consisting of three subprocesses: (1) determining performance criteria, (2) self-directed feedback seeking, and (3) self-reflection. Despite being an early effort to unpack the black box, the results are limited by a small sample and highly descriptive and interpretive analysis of interview data.

More recently, Panadero et al. (2020) analyzed the behaviour of 64 secondary education students when self-assessing Spanish and mathematics tasks. Multi-method data sources (i.e., think aloud protocols, direct observation and self-report via questionnaires) described self-assessment actions as either strategies or criteria. The study showed that (1) the use of self-assessment strategies and criteria was more frequent and advanced without feedback and among girls; (2) there were different self-assessment patterns by school subject; (3) patterns of strategy and criteria use differed by school year, and (4) none of the self-assessment strategies or criteria had a statistically significant effect on self-efficacy.

Factors influencing self-assessment

Feedback in general has been shown to improve academic performance, especially when focused on specific tasks, processes, and self-regulation (Hattie & Timperley, 2007; Wisniewski et al., 2020). Butler and Winne's (1995) feedback review showed that self-regulated learners adjust their internal feedback mechanisms in response to external feedback (e.g., scores, comments from teachers, etc.). Scholars have claimed that students need instructor's feedback about their self-assessments as well as about content knowledge (Andrade, 2018; Brown & Harris, 2014; Boud, 1995; Panadero et al., 2016). Previous studies have shown little effect of external feedback on student self-assessment (Panadero et al., 2012, 2020; Raaijmakers et al., 2019). Thus, understanding how external feedback such as instructor's or via instruments (e.g. rubrics) can influence students' self-assessment is important.

Among feedback factors that influence student outcomes (Lipnevich, Berg, & Smith, 2016), the timing of feedback is important. In general, delayed feedback is more likely to contribute to learning transfer, whereas prompt feedback is useful for difficult tasks (Shute, 2008). However, linking feedback to self-assessment is relatively rare. Panadero et al. (2020) found that secondary education students self-assessed using fewer strategies and criteria after receiving feedback. This has crucial implications for instructors as to when they should deliver their feedback, if they want students to develop calibrated self-assessments.

One potentially powerful mechanism for providing feedback is a marking, scoring, or curricular rubric, which has been shown to have stronger effects on performance than other assessment tools, such as exemplars (Lipnevich et al., 2014, in press) or scripts (Panadero et al., 2012). The use of rubrics in education and research has grown steadily in the last years (Dawson, 2017), due to its instructional value with positive effects for students, teachers and even programs (Halonen et al., 2003). Rubric

use has been associated with positive effects on self-assessment interventions and academic performance (Brookhart & Chen, 2015). Previous research has demonstrated that a rubric alone produced better results than combining rubrics with exemplars (Lipnevich et al., 2014, in press). Although there is previous research exploring the effects of rubrics when compared or combined with feedback (Panadero et al., 2012, 2020; Wollenschläger et al., 2016), we still need insights around the impact of rubrics with or without feedback on student self-assessment.

It was established in the self-assessment literature that more sophisticated and accurate self-assessments are conducted by older and more academically advanced students (Barnett & Hixon, 1997; Boud & Falchikov, 1989; Brown & Harris, 2013; Kostons et al., 2009, 2010). As Boud and Falchikov (1989) demonstrated it was subject specific competence that reduced discrepancy between self-assessments and teacher evaluations. However, recent research shows that the relationship might not be so straight forward (Panadero et al., 2020; Yan, 2018). Additionally, it is unclear at what level of higher education students need to be to have sufficient expertise to self-assess appropriately. Thus, an investigation with students in consecutive years of study in the same domain might clarify the role of year level on self-assessment capacity.

Research Aim and Questions

The current study adds to this body of research by examining the number and type of self-assessment strategies and criteria among higher education students in a randomized experiment which manipulated three feedback conditions (rubric vs. instructor's vs. combined) without a control group because the university Ethics Committee did not grant permission. Importantly, we also examined feedback occasion (before vs. after) and year level (1st, 2nd and 3rd university undergraduates). This is a

single group, multi-method study (i.e., think aloud, observation, and self-report; though only the two first ones are analyzed here).

We explored three research questions (RQ):

RQ1. What are the self-assessment strategies and criteria that higher education students implement before and after feedback?

Hypothesis 1 (H1): Self-assessment strategies and criteria will decrease when feedback is provided. In line Panadero et al. (2020).

RQ2. What are the effects of feedback type and feedback occasion on self-assessment behaviors (i.e., number and type of strategy and criteria)?

H2: Rubric feedback will provide better self-assessment practices than other feedback types. In line with Lipnevich et al. (2014, in press).

RQ3. What is the effect of student year level on the results?

H3: Students in higher years within a discipline will use more sophisticated strategies and criteria in their self-assessments. There are results in different directions from no differences in primary education but less self-assessment in more advanced secondary education students (Yan, 2018), to more similarities than expected yet some differences identified in secondary education students (Panadero et al., 2020). Nevertheless, as our participants are higher education students it is expected they will behave differently with more advanced students showing higher self-assessment skills.

Method

Sample

A convenience sampling method at one university site where the first author worked created a sample of 126 undergraduate psychology students (88.1% females) across first, second and third year of study (34.9%, 31.7%, and 33.3%, respectively).

Participants were randomly assigned to one of three feedback conditions: rubric only (n = 43), instructor's written feedback (n = 43), and rubric and instructor's written feedback combined (n = 40). Participants received credit in accordance with the faculty volunteering programme. In a 3 x 3 ANOVA, given a risk level of $\alpha = .005$, and a statistical power of $1 - \beta = .800$, the current sample size would detect a medium effect size, $f = 0.280$ (G*Power 3.1.9.2; Faul, Erdfelder, Lang, & Buchner, 2007).

Data Collection and Instruments

Data from the video-recorded think aloud protocols was inductively coded using the categories defined in a previous study (Panadero et al., 2020). In addition, two structured feedback intervention tools were used (i.e., rubric and instructor's feedback).

Coded Video-Recorded Data

Think-aloud protocols. Participants were asked to think aloud while conducting two self-assessments of their written essay. The first was an unguided self-assessment in which students were asked to evaluate the quality of their essay and the reasons for their evaluation. Participants were asked to express their thoughts and feelings and reminded that if they were silent they would be prompted to think out loud. After the feedback was provided, students were asked to talk about their thoughts and feelings concerning the feedback and to repeat the think aloud process of self-assessing their essay. If the participant remained silent for more than 30 seconds, they were reminded verbally to think out loud. There were no time restrictions to perform the self-assessment.

A closed coding process was followed, as the codes were already defined as part of a previous study (see Panadero et al., 2020) with secondary education students. In such study, a deductive approach was employed to create the two general coding categories of self-assessment elements: strategies and criteria. Additionally, we created codes for those general categories. The categories were contrasted with the data using

an inductive approach, to ensure that they were applicable to the new sample and procedure.

The video-recorded think-aloud content was coded to identify the strategies and criteria each student used. As in our previous study, we further organized each set of 13 categories into four levels for clarity in interpretation (0-3). Such levels classify the categories depending on their type and complexity. Details of the levels, categories, definitions, and exemplar comments are provided in Table 1.

Table 1

Category description and examples

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	Category	Description	Example comment
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"*

Intervention Prompts

Rubric (Appendix A). It was created for this study using experts' models of writing composition. It contains three types of criteria: (1) writing process, (2) structure and coherence, and (3) sentences, vocabulary and punctuation. There are three levels of quality: low, average and high. The rubric is analytic as three criteria should be scored independently. The rubric was provided to some of the students during the experimental

procedure, depending on the experimental condition; but it was not explicitly used by de instructor to provide feedback on the essays.

Instructor's feedback (Appendix B). The instructor provided feedback to each essay using the same categories as the rubric. For the "writing process" criterion, as that was not directly observable by the instructor, he provided feedback by suggesting whether some of those strategies had been put into places (e.g. planning). Additionally, it included a grade ranging from 0 to 10 points. All essays were evaluated by the second author. The first author evaluated a third of the essays reaching total agreement in the rubric categories.

Procedure

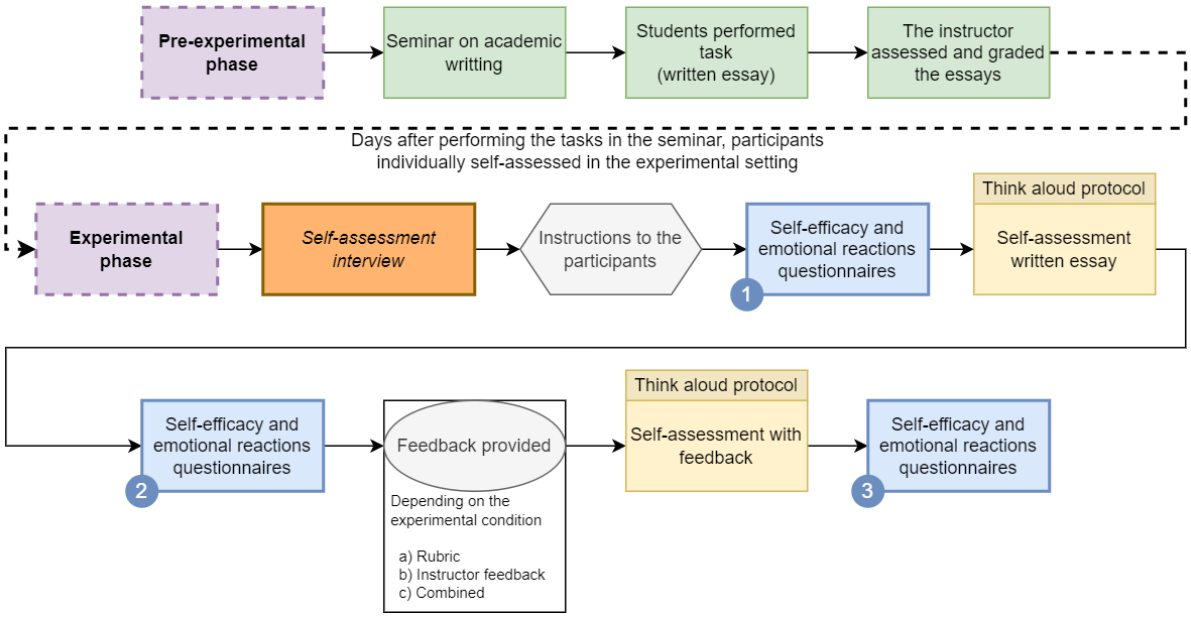
This randomized experiment is part of a larger study; this report focuses on the specific self-assessment strategies and criteria students elicited (See Figure 1), as measured via thinking aloud protocols and observations. After attending a 3 hours' group seminar on academic writing, participants wrote a short essay answering the question: "Why is the psychologist profession necessary?". This topic was directly directed to the participants' psychology programme. There was no length limitation for the essays that were written in the participants' computers, which then submitted it to the research team. This essay did not have implications outside of the research experiment but we emphasized its utility for the students' academic perspective of the programme. Some days later (approx. 1 week), participants went individually to the laboratory setting. There, they participated in the experiment face-to-face with one of the authors.

First, they received the instructions for self-assessing their essay that was handed out to them in its original form, in other words with no feedback. Students were instructed to while self-assessing think aloud their thoughts, emotions, and motivational

reactions. Then they performed the first think aloud self-assessment of the essay they had written. Right after, participants were given feedback on their essay according to the condition they had been assigned to (rubric vs. instructor vs. combined) and asked to self-assess again. The rubric group was handed out the rubric with the instruction of using it for their self-assessment. In the instructor's feedback group, the participants were said that they should use the instructor's feedback for their self-assessment. Finally, the combined group received both instructions. After reading the feedback, each participant repeated the self-assessment thinking aloud.

Figure 1

Experimental procedure



Data Analysis

The coding of the think aloud utterances for strategies and criteria was evaluated in three rounds of inter-judge agreement. In round one, agreement between two judges on 15 videos reached an average Krippendorff's $\alpha = .78$, with three categories below .70. After discussion and consensus building around the low agreement categories, a second set of 15 videos was coded with an average Krippendorff's $\alpha = .83$. A third round, using 15 new videos, produced Krippendorff's $\alpha = .87$. This indicates the final

coding values are dependable. The direct observation was performed in situ during data collection but more intensively during the coding of the video-data. The observation data was used to inform and confirm the thinking aloud categories via defining the participants' behavior, so as supplementary data to further establish the categories.

The categorical variables were described using multiple dichotomous frequency tables, as each participant could display more than one behavior. To study the effect of the factors (feedback occasion, condition, and year level) on self-assessment strategies and criteria frequencies we conducted ANOVAs and square test to compare differences among the levels.

Results

RQ1. What are the self-assessment strategies and criteria that higher education students implement before and after feedback?

Type of strategies. Table 2 shows the multiple self-assessment strategies enacted by the participants. The most used before feedback were *Read the essay, Think of different responses,* and *Read the instructions.* After the feedback, the most used were *Read the feedback or rubric received* and *Compare essay to feedback or rubric*. These strategies are low level according to our code except for *Think of different responses* which show a deeper level of self-assessment elaboration. Three main results can be extracted. First, the strategies used before and after feedback are similar in nature, with five categories occurring at both moments. However, second, once the students received the feedback there was a general decrease in the number of frequency of strategies with three out of the five strategies showing significant decreases. This is logical as most of the strategies were basic and participants did not need to enact them again (e.g. read the essay, which they had done just minutes before). Also, there was the appearance of two new strategies that were not present before the feedback as they are specific to the

reception of feedback (i.e. *Read the feedback or rubric received* and *Compare essay to feedback or rubric*). Third, after the feedback there was also a new category that the participants did not activate it before: *Compares question and response*.

Table 2

Type of Strategies Deployed by Feedback Condition and Time

	Condition			
Strategy	Rubric *N*	Instructor's feedback *N*	Combined *N*	Grand total
Before feedback (n = 121)				
Remember the instructions	17	18	12	47
Remember the seminar	11	14	10	35
Read the essay	41	38	39	118
Compare instructions and essay	13	17	11	41
Perform the essay again	2	5	0	7
Think of different responses	18	22	9	49
After feedback (n = 124)				
Remember the instructions	4**	2***	1***	7
Remember the seminar	5^{ns}	5^{ns}	3^{ns}	13
Read the essay	5***	9***	5***	19
Read the feedback or rubric received	43***	41***	40***	124
Compares question and response	3*	3**	2*	8
Compares essay to feedback or rubric	42***	41***	40***	123
Perform the essay again	0^{ns}	0^{ns}	0^{ns}	0
Think of different responses	3*	4**	5^{ns}	12

Note. binomial χ^2 comparison between times *=p<.05; **=p<.01; ***=p<.001; ^{ns}=p>.05

Type of criteria. As the students could choose more than one criterion, we described multiple dichotomous variables. In general, the most used criteria before the feedback were: *Sentences and punctuation marks*, *Negative intuition*, *Positive intuition*, and *Paragraph structure* (Table 3). The most used after the feedback were: *Feedback received*, *Sentences and punctuation marks*, *Paragraph structure*, and *Writing process*. When it comes to the trajectories, most of the criteria frequencies decreased significantly after receiving the feedback. However, there were three criteria that increased after feedback (significantly *Writing process* and *Paragraph structure*, non-

significantly *Sentences and punctuation marks*) all being advanced strategies and all increasing in the rubric and combined condition but decreasing in the instructor's condition. Additionally, a new criterion was used *Feedback received*, which, for obvious reasons, only occurred after feedback.

Table 3

Type of Criteria Deployed, by Feedback and Condition

		Condition			
	Criteria	Rubric *N* (%)	Instructor's feedback *N* (%)	Combined *N* (%)	Grand total
Before feedback (*n* = 122)	Negative intuition	32	24	17	73
	Positive intuition	21	18	23	62
	Negative hindsight	12	9	14	35
	Positive hindsight	2	0	1	3
	Writing process	12	8	15	35
	Paragraph structure	23	26	23	72
	Sentences and punctuation marks	34	30	31	95
	Instructions	10	15	11	36
	Spelling	5	2	8	15
	Without clear criteria	0	4	2	6
After feedback (*n* = 124)	Negative intuition	13**	10*	6*	29
	Positive intuition	9 ^{ns}	3**	4***	16
	Negative hindsight	2*	4 ^{ns}	3*	9
	Positive hindsight	0 ^{ns}	0 ^{ns}	1 ^{ns}	1
	Feedback received	43***	40***	40***	123
	Writing process	42***	3 ^{ns}	38**	83
	Paragraph structure	41*	26 ^{ns}	37 ^{ns}	104
	Sentences and punctuation marks	41 ^{ns}	27 ^{ns}	40 ^{ns}	108
	Instructions	1*	5*	4 ^{ns}	10
	Spelling	1 ^{ns}	6 ^{ns}	7 ^{ns}	14

Note. binomial χ^2 comparison between times *=p<.05; **=p<.01; ***=p<.001; ^{ns}=p>.05

RQ2. What are the effects of feedback type and feedback occasion on number and type of strategy and criteria in self-assessment behaviours?

At time 1, before receiving feedback, the number of strategies by condition (Table 4) differed statistically and substantially, ($F_{(2, 121)}$ = 4.22, p = .017, η^2 = .65) with

a significant post hoc difference between the instructor condition (*M* = 2.78, *SD* = 0.183) and the combined condition (*M* = 2.06, *SD* = 0.185); the rubric condition did not differ from any of the two (*M* = 2.37, *SD* = 0.179). When it comes to number of criteria used the conditions were equivalent ($F_{(2, 121)} = 0.48$, $p = .62$, $\eta^2 = .008$, $1 - \beta = .127$) with no differences among the three groups: instructor (*M* = 3.32, *SD* = 0.224), rubric (*M* = 3.51, *SD* = 0.219) or combined (*M* = 3.63, *SD* = 0.227). We also analyzed if there were differences within the different levels of strategies ($\chi^2_{(6)} = 8.38$, $p = .21$), and levels of criteria, ($\chi^2_{(6)} = 6.32$, $p = .39$), but both were equivalently distributed across conditions.

Table 4

Number and Level of Strategies Deployed by Condition and Time

		Level				Number	
Condition	*N*	0	1	2	3	*M*	*SD*
				Strategies			
Time 1							
Rubric	43	41	13	38	20	2.60	1.18
Instructor	41	38	17	32	27	2.78	1.29
Combined	40	39	11	22	9	2.03	1.03
Total	124	118	41	92	56		
Time 2							
Rubric	43	48	45	9	3	2.44	0.77
Instructor	41	50	44	7	4	2.56	0.98
Combined	40	45	42	4	5	2.40	0.67
Total	124	143	131	20	12		
				Criteria			
Time 1							
Rubric	43	0	67	15	69	3.51	1.28
Instructor	41	4	51	17	64	3.32	1.52
Combined	40	2	55	19	69	3.63	1.50
Total	124	6	173	51	202		
Time 2							
Rubric	43	0	24	45	124	4.49	1.03
Instructor	41	0	17	51	66	3.02	1.33
Combined	40	0	14	51	115	4.50	0.82
Total	124	0	56	149	308		

At Time 2, after feedback, the number of strategies by condition (Table 4) did not differ ($F_{(2,121)}$=0.42, p=.66, η^2 =.007, 1 - β =.118): instructor (M = 2.56, SD = 0.976), rubric (M = 2.44, SD = 0.765) or combined (M = 2.40, SD = 0.671), showing that the effects of rubric had no meaningful impact on the number of strategies. However, the number of criteria differed substantially ($F_{(2,121)}$=25.30, p<.001, η^2 =.295) with significant post hoc differences for Rubric (M = 4.48, SD = 0.165) and combined conditions (M = 4.50, SD = 0.171) that outperformed the instructor condition (M = 3.02, SD = 0.169), both at p<.001. Similar to the number of strategies, the level of strategies was equivalently distributed across conditions, ($\chi^2(6) = 2.29$, p = .89). However, and to be expected, the level of criteria differed significantly ($\chi^2(4) = 12.00$, p = .02), which is likely to be a function of the large sum of criteria differences across conditions at Time 2 (i.e., 193, 134, 180, respectively). When viewed as differences based on percentage of responses at each level, this is statistically not significant ($\chi^2_{(4)} = 7.74$, p = .10).

When we explored the interaction condition by feedback occasion, we found no significant effect in self-assessment strategies ($F_{(2,121)}$=1.74, p=.180, η^2 =.028). However, we found a significant main effect of condition in self-assessment criteria ($F_{(2,115)} = 7.97$, p = .001, η^2 = .116). The pre-post increase in number of strategies deployed was greater (post hoc p = .002) in the rubric (M = .938, SE = .247) than in the instructor's feedback (M = -.291, SE = .253) condition. The combined condition (M = .881, SE = .256) also yielded a greater increase (post hoc p = .004) compared to the instructor's feedback.

RQ3. What is the effect of student year level on the results?

We calculated the differences in strategies and criteria by year level between pre- and post-feedback conditions in two-way ANOVAs with condition and year level

as factors. When it comes to the use of strategies neither main effects (i.e., year level, $F_{(2, 115)} = 1.04$, $p = .359$, $\eta^2 = .018$, $1 - \beta = .227$; feedback type, $F_{(2, 115)} = 1.72$, $p = .183$, $\eta^2 = .029$, $1 - \beta = .355$) nor interaction ($F_{(2, 115)} = .973$, $p = .425$, $\eta^2 = .033$, $1 - \beta = .300$) were significant, largely due to lack of power. Likewise, in the use of criteria the same result was seen (i.e., year level, $F_{(2, 115)} = 1.68$, $p = .192$, $\eta^2 = .028$, $1 - \beta = .347$; feedback type, $F_{(2, 115)}$=7,57, p<.001, η^2 =.116, $1 - \beta = .940$; and interaction, $F_{(2, 115)} = 0.25$, $p = .911$, $\eta^2 = .009$, $1 - \beta = .102$). Therefore, our hypothesis that older students would show more advanced self-assessment action is not supported.

Discussion

This study explored the effects of three factors (i.e., feedback type, feedback occasion, and year level) on self-assessment strategies and criteria. This study contributes to our understanding of what happens in the "black box" of self-assessment by disentangling the frequency and type of self-assessment actions in response to different types of feedback.

Effects on Self-assessment: Strategy and Criteria

In RQ1, we categorized self-assessment actions in a writing task in terms of strategies and criteria. Strategies were categorized on their depth or sophistication ranging from very basic activities (e.g. read the essay) to advanced ones (e.g. think of different responses). Understandably, the most common strategies were relatively low level, as they are foundational to understanding the task. However, once feedback was received most of the strategies focused on the content of the feedback received (e.g. compares essay to feedback or rubric), making the feedback as the anchor point of comparison (Nicol, 2020). In consequence, the strategies used prior to feedback were greatly reduced in number, indicating that, with feedback, self-assessment strategies were led by that information. Self-assessment criteria demonstrated similar effects. Prior

to feedback, students used a wide range of criteria ranging from very basic (e.g. negative intuition) to advanced (e.g. writing process). Upon receipt of feedback, most of the criteria responded to the feedback in a less sophisticated manner, especially in the presence of rubrics.

In terms of the three different feedback conditions (RQ2), the two conditions containing rubrics outperformed the instructor's feedback group in terms of criteria and close the initial gap in strategies. Despite of the instructor's feedback condition having a higher number of self-assessment strategies before the intervention than the combined group, that difference vanished after feedback. Both the rubric and combined conditions had a higher number and more advanced types of criteria after feedback than the instructor's feedback condition by large margins. No statistically significant differences in self-assessment strategies and criteria were found across the year levels (RQ3) regardless of feedback presence or type.

Regarding the alignment of our results to previous research, first, the feedback occasion effects on self-assessment strategies are very similar to a study with secondary education students (Panadero et al, 2020), as these strategies decreased significantly after feedback except for the ones related to the use of the feedback. In contrast, while the secondary education students decreased their number of criteria used and the type of criteria, here university students increased the number of criteria and used more advanced criteria when using rubrics, an instrument that was not implemented in Panadero et al. (2020). Wollenschläger and colleagues (2016), compared three conditions (rubric, rubric and individual performance feedback, rubric and individual performance-improvement), finding that the latest was more powerful in increasing performance than the two first conditions. An important difference of this study is that it examined the impact of rubric and feedback on self-assessment, while the

Wollenschläger et al. (2016) study examined the effects on academic performance. Hence, the impact of feedback appears to be contingent upon the kind of assessment being implemented.

Also, the secondary education students in Panadero et al. (2020) study showed differences across year levels, which was not found here with university students. This year level lack of effects aligns with Yan (2018) primary education students where he did not find differences, but is it not aligned with the same study when comparing secondary education students where he found significant differences (i.e. older students self-reporting lower levels of self-assessment). Unlike studies that have reported clearly delineated phases of self-assessment (Yan & Brown, 2017), the think aloud protocols in this study did not identify clear-cut phases, finding instead a naturally evolving process. While Panadero et al. (2012) reported that scripts were better than rubrics, this study found that the presence of rubrics led to more sophisticated criteria use; future research would need to determine if script-based feedback would have any greater impact.

Three main conclusions from this study can be reached. First, there are different effects due to the form of feedback, with rubric-assisted feedback being especially promising for self-assessment. The effect of rubrics corrected the initial difference between the instructor's feedback and the combined group so that, after receiving the feedback or/and rubric, all conditions were equal in terms of the number of self-assessment strategies. Also, and more interestingly, the rubrics conditions showed bigger effects on the use of criteria even in a situation in which the participants had already self-assessed freely before. This might indicate that rubrics as a tool are indeed very useful in stimulating student reflection on their work (Brookhart, 2018), more so than instructor's feedback which may have been perceived as external criticism rather than supportive of improvement. This effect could be caused by instructor's feedback

putting students in passive position (e.g., they are being evaluated, they are recipients of feedback), while rubrics provided them with guides to explore and reflect by themselves. This also might speak to the importance of tools, such as rubrics, to support active self-assessment, rather than of the importance of providing corrective or evaluative feedback.

This result might seem logical, as rubrics contain clear criteria and performance levels to which performance can be anchored. This may be especially pertinent to higher-education students who are used to being assessed and graded against standards (e.g. Brookhart & Chen, 2014). Therefore, one viable conclusion is that the best type of feedback among the explored ones here is using rubrics, followed by a combination of rubric and instructor's feedback.

Second, the introduction of feedback does impact self-assessment practices. Feedback decreased the number of strategies and increased the level of criteria used. A feature of this study is that students had to self-assess before they received feedback and then again upon receiving it. This process shows the impact of feedback in that it changes the strategies and criteria that students used. Therefore, for educational benefit, feedback may best be presented after students are required to implement their own self-assessment based on their own strategies and criteria. It may be that performance feedback prior to self-assessment will discourage students from the constructive strategies and criteria they exhibited in the pre-feedback stage.

And third, although self-assessment strategies did not become more advanced over years of study among our participants (i.e., our *year level* variable), this is not likely to be because there was a ceiling effect in the task itself. It is possible for students to exhibit in such a task more sophisticated strategies and criteria. It may be that, once entry to higher education is achieved, self-assessment is relatively homogeneous for this

type of task. Perhaps much more demanding tasks (e.g., research thesis) would require more sophisticated self-assessment behaviors.

Future Research

First, our participants conducted a first self-assessment without any structure or teaching on how to effectively evaluate one's own work. Future research could introduce an intervention on self-assessment prior to the introduction of feedback to better eliminate confounds between self-assessment and feedback. Second, feedback focused on the essay writing task, not on the self-assessment process; such feedback may have had an effect on the quality of subsequent self-assessments (e.g. Andrade, 2018; Panadero et al., 2016). Third, the absence of a control group with no feedback is a limitation, although our conditions can be more realistic controls than no feedback as it is unusual to find activities without some kind of feedback in real educational settings. Additionally, internal feedback seems to be ubiquitous and automatic in any event (Butler & Winne, 1995), so even in the absence of experimenter-controlled feedback, there will be feedback. Fourth, it could be an interesting line of work to explore peer feedback and how it affects self-assessment strategies and criteria. While there has been some research in that direction (To & Panadero, 2019), it would be interesting to explore these effects using our methodology to fulfill the aim of "opening the black box of self-assessment". Fifth, it is likely that greater insights into self-assessment could be achieved by combining this self-reported approach to self-assessment with technology, such as eye-tracking (Jarodzka, Holmqvist, & Gruber, 2017) or physiological reaction equipment (Azevedo, Taub, & Mudrick, 2018). These additional tools may allow for a more precise understanding of the underlying cognitive, emotional, and motivational processes in self-assessment and in response to feedback. And sixth, future research

should also seek to determine if there are gender or content-specific effects on self-assessment and feedback (Panadero et al., 2016).

Conclusions

In general, this study shows that rubrics have the greatest potential to increase positively the quality of student self-assessment behaviors. The study also indicates that feedback has a mixed effect on self-assessment strategies and criteria use. This may explain in part why reliance on feedback from peers or markers has been shown to have a negative impact on overall academic performance (Brown, Peterson, & Yao, 2016). Students who rely more on their own evaluative and self-regulatory learning strategies are more likely to discount external feedback. The provision of rubrics is likely to enable more effective and thoughtful self-assessed judgements about learning priorities. All in all, this study helps to better understand the specific strategies and criteria higher education students enact while self-assessing, something that is key to really understanding how self-assessment works.

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.

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.

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. https://doi.org/10.1080/00220671.1997.10543773

Boud, D. (1995). “Assessment and learning: contradictory or complementary”. In P. Knight (Ed.), *Assessment for learning in higher education*, (pp. 35-48). London: Kogan.

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. https://doi.org/10.1007/BF00138746

Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education, 3*(22), 1-12. https://doi.org/10.3389/feduc.2018.00022

Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review*, *67*(3), 343-368. https://doi.org/10.1080/00131911.2014.929565

Brown, G. T. L., & Harris, L. R. (2013). Student self-assessment. In J. H. McMillan (Ed.), *The SAGE handbook of research on classroom assessment* (pp. 367-393). Sage.

Brown, G. T. L., & Harris, L. R. (2014). The future of self-assessment in classroom practice: Reframing self-assessment as a core competency. *Frontline Learning Research, 3*, 22-30. https://doi.org/10.14786/flr.v2i1.24

Brown, G. T. L., Peterson, E. R., & Yao, E. S. (2016). Student conceptions of feedback: Impact on self-regulation, self-efficacy, and academic achievement. *British Journal of Educational Psychology*, *86*(4), 606-629. https://doi.org/10.1111/bjep.12126

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

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

Dawson, P. (2017). Assessment rubrics: Towards clearer and more replicable design, research and practice. *Assessment & Evaluation In Higher Education*, 1-14. https://doi.org/10.1080/02602938.2015.1111294

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/10.3102/00346543059004395

Faul, F., Erdfelder, E., Lang, A.-G. y Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191. https://doi.org/10.3758/BF03193146

Halonen, J. S., Bosack, T., Clay, S., McCarthy, M., Dunn, D. S., Hill Iv, G. W., . . . Whitlock, K. (2003). A rubric for learning, teaching, and assessing scientific inquiry in psychology. *Teaching of Psychology, 30*(3), 196-208. https://doi.org/10.1207/s15328023top3003_01

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

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. https://doi.org/10.16910/jemr.10.1.3

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

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. https://doi.org/10.1016/j.compedu.2009.09.025

Lipnevich, A. A., McCallen, L. N., Miles, K. P., & Smith, J. K. (2014). Mind the gap! Students' use of exemplars and detailed rubrics as formative assessment. *Instructional Science, 42*(4), 539-559. https://doi.org/10.1007/s11251-013-9299-9

Lipnevich, A. A., Berg D. A., & Smith J. (2016). Toward a Model of Student Response to Feedback. In Brown, G. T. L. & Harris, L. (Eds.). *Handbook of Human and Social Conditions in Assessment*, (pp. 169–185.) London: Routledge.

Lipnevich, A. A., Panadero, E., & Calistro, T. (in press). Unraveling the Effects of Rubrics and Exemplars on Student Writing Performance. *The Journal of Experimental Psychology: Applied*. https://doi.org/10.1037/xap0000434

Panadero, E., Tapia, J. A., & 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. https://doi.org/10.1016/j.lindif.2012.04.007

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*(4), 803-830. https://doi.org/10.1007/s10648-015-9350-2

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*, 1-28. https://doi.org/10.1080/0969594X.2020.1835823

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. https://doi.org/10.1007/s11409-019-09189-5

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189. https://doi.org/10.3102/0034654307313795

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2017). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education*. https://doi.org/10.1007/s10734-017-0220-3

To, J., & Panadero, E. (2019). Peer assessment effects on the self-assessment process of first-year undergraduates. *Assessment & Evaluation In Higher Education, 44*(6), 920-932. doi:10.1080/02602938.2018.1548559

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*(3087). https://doi.org/10.3389/fpsyg.2019.03087

Wollenschläger, M., Hattie, J., Machts, N., Möller, J., & Harms, U. (2016). What makes rubrics effective in teacher-feedback? Transparency of learning goals is not enough. *Contemporary Educational Psychology*. http://dx.doi.org/10.1016/j.cedpsych.2015.11.003

Yan, Z. (2018). Student self-assessment practices: The role of gender, school level and goal orientation. *Assessment in Education: Principles, Policy & Practice, 25*(2), 183-199. https://doi.org/10.1080/0969594X.2016.1218324

Yan, Z. (2019). Self-assessment in the process of self-regulated learning and its relationship with academic achievement. *Assessment & Evaluation In Higher Education*, 1-15. https://doi.org/10.1080/02602938.2019.1629390

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

Appendix A. Rubric

CATEGORY	LOW QUALITY	AVERAGE QUALITY	HIGH QUALITY
Writing process	I started writing the text without planning what I wanted to write. I have hardly reread what I was writing and, when I finished, I have not reviewed the text or I have only looked for misspellings.	2 options: a) Before writing, I have planned what I wanted to communicate. At the end, I have hardly reviewed the text or I have only looked for misspellings. b) I started writing without thinking much about what I wanted to tell. However, I reviewed the text several times, looking for all or some of these factors: Text structure, coherence and connection between paragraphs, clarity of the message, style, and spelling.	Before writing, I thoroughly planned what I wanted to tell and how I was going to do it. I reviewed while I was writing and, at the end, I also reviewed the full text at least once. While reviewing, I looked for all or some of these factors: Text structure, coherence and connection between paragraphs, clarity of the message, style, and spelling.
Text components: Structure & coherence /connection between paragraphs	There is no clear structure, with an introduction, a crux, and a closing. Lack of incorrect use of text connectors and/or discourse markers. Regarding paragraphs, one of these two happens: a) The text has only one or two paragraphs, without clear internal and external coherence. b) The text has many very short paragraphs, which makes it difficult to follow the argument line.	A structure is somehow present (introduction, crux and closure) but could be more clearly delimited. Connectors are most of the times used appropriately. However, there may be one or more of these flaws: Same paragraph includes different unorganized ideas. Same idea in two paragraphs when it could be in one. The paragraph where the argument is developed is too long; it could be divided. Connector/text markers are misused.	There is a very clear structure in the text: including opening, argument crux and closing. Ideas are connected and presented in well-organized paragraphs. Connectors and/or discourse markers are effectively used.
Text components: Sentences, vocabulary & punctuation.	Sentences are too long (over 40 words) or too short. Excessive use of text insertions within sentences. Punctuation is incorrect (e.g. lack of commas, the break the sentence). Too many colloquial expressions. Abuse of passive or impersonal tenses.	Most sentences are of adequate length, with a few too long or short or incomplete. Punctuation is correct, although there may be a few mistakes. The vocabulary is adequate, but different terms are used to refer to the central concept of the text. Some colloquial expression may appear.	The sentences are well constructed, usually following a simple structure, in an active language and a coherent use of the verbs. Punctuation is correct. The vocabulary is adequate, and the main terms are used with precision.

Appendix B. Instructor feedback (three samples)

GRADE: 3,5

The text structure has important flaws. It does not follow a coherent argument; on the contrary, ideas change abruptly in each paragraph. For instance, any of the first three paragraphs could actually be the introduction paragraph because each of them present different ideas as it was the introduction. Later, in the argument crux there are several ideas without connection. Finally, the previous to the last paragraph seems to be closing the text but, nonetheless, there is an additional paragraph after it. Furthermore, that previous to the last paragraph includes a new idea (about the methodology), which has not been mentioned before and it could be used as an argument in favour of Psychology.

To sum, even though a central message can be perceived (the multiple areas of application of Psychology), it is not developed nor transmitted effectively. Regarding grammar, highlighted in the text there are mistakes and comments in the footnotes.

GRADE: 6,5

The text has a quite clear structure, with a paragraph of introduction, three for crux and a closing paragraph. However, there are two arguments in the introduction, and one of them is not developed in order to refute it (the skepticism of certain people). In addition, the last paragraph includes a new idea that has not been discussed before and it does not recap and finish with the main message to be transmitted. In general, there is a correct use of connectors and discourse markers.

Regarding the style and grammar, in general, the construction of the sentences is correct, and the vocabulary is appropriate. Nevertheless, there are some mistakes in the sentence construction and some limitations in the vocabulary selection, which are highlighted in the text and commented in footnotes.

GRADE: 9

The text has an adequate argumentative structure, with and introductory paragraph, four for the argument crux and a closing paragraph. Connectors and discourse marks are properly used.

Regarding the text style, it is correct considering the vocabulary, the use of punctuation marks and the sentence construction. There are some minor mistakes highlighted in the text and commented in footnotes.