Practical Assessment, Research, and Evaluation

Volume 26 *Volume 26, 2021* Article 16

June 2021

Variability In The Accuracy Of Self-Assessments Among Low, Moderate, And High Performing Students In University Education

Samuel Parra León
University of Jaén, Spain

Antonio Pantoja Vallejo
University of Jaén, Spain

James Byron Nelson
University of the Basque Country, Spain

Recommended Citation

León, Samuel Parra; Pantoja Vallejo, Antonio; and Nelson, James Byron (2021) "Variability In The Accuracy Of Self-Assessments Among Low, Moderate, And High Performing Students In University Education," *Practical Assessment, Research, and Evaluation*: Vol. 26, Article 16.
DOI: https://doi.org/10.7275/6q91-az58
Available at: https://scholarworks.umass.edu/pare/vol26/iss1/16

Variability In The Accuracy Of Self-Assessments Among Low, Moderate, And High Performing Students In University Education

Cover Page Footnote

This work was supported by the by the Spanish Ministry of Economy and Competitiveness under the Grant PSI2014-52263-C2-1-P; Junta de Andalucía under the Grant HUM642; Eusko Jaurlaritza (Government of the Basque Country) under the Grant IT1341-19 from

Practical Assessment, Research & Evaluation, Vol 26 No 16
León, Vallejo, & Nelson, Validity in Self-Assessment

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 26 Number 16, June 2021 ISSN 1531-7714

Variability In The Accuracy Of Self-Assessments Among Low, Moderate, And High Performing Students In University Education

Samuel Parra León, *University of Jaén, Spain*
Antonio Pantoja Vallejo, *University of Jaén, Spain*
James Byron Nelson, *University of the Basque Country, Spain*

The present work empirically examines the validity of Student Self-Assessment (SSA) as an educational assessment in higher education. We briefly review the principle methodological factors that could affect SSA validity, as well as the main findings identified in the literature. One empirical study is presented that compares student-self evaluations on a test with the evaluation made by the course instructor while controlling for students' experience with SSA, criteria, rubric, and scales used by the student and teacher, and that the teacher was blind. Results show a strong correlation overall between the SSA and the instructor's evaluation and show that lower-performing students tend to over-estimate their performance while higher-performing students under-estimate their performance. The results support that SSA is valid for the average student, but less so for those that deviate above and below average in the absence of measurements of potentially mediating variables. The need to consider metacognitive factors in SSA is proposed.

Introduction

Evaluation has always been an essential part of the instructional processes, as it measures and assigns value to achievement in the teaching and learning process (e.g., Ainscow, 1988; Ysseldyke & Matson, 1988). Of existing types of evaluation, Student Self-Assessment (SSA, hereafter) has aroused substantial interest in the research community. Panadero et al. (2016, p. 2) describe SSA as "…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".

SSA has been an important topic in the analysis of teaching and learning processes for decades. The first important SSA review (Boud & Falchikov, 1989) provided answers to many doubts that the SSA practice raised at that time. However, the review raised a series of unresolved questions about factors that influence SSA. Since 1989, the popularity of self-assessment has increased considerably, becoming common practice for many teachers (e.g., Berry 2011; Black & Wiliam, 1998). Panadero et al. (2014) showed that 90% of the professors in their survey had used SSA in their courses and that 90% indicated a positive experience.

The increasing use of this evaluation technique has led to heterogeneity in how it is defined and understood. Currently up to 20 different categories of SSA can be found in the literature (see Panadero et al., 2016 for review). These different categories are organized from

practices that seek the student's summative and general SSA without any established criteria (e.g., Stanton, 1978), to those that view SSA as part of the student's self-regulation process (e.g., Panadero & Alonso-Tapia, 2014). Roberts et al (2019, p. 79) describe the latter as "… the ability to set, monitor, and reflect on goals and then set new goals to monitor and reflect upon. This cycle of learning is needed to be an efficient and effective active participant in one's own learning." In that regard, self-regulation would represent the highest level of acquisition achieved by the student within his or her own learning process and, consequently, of self-evaluation. It is necessary that teachers stimulate reflection in the classroom and provide tools and strategies for self-regulated learning (Torres & Tackett, 2016), which also includes self-assessment.

Numerous studies (e.g., Dochy et al., 1999) suggest a relationship between the SSA process and other factors regarding learning, such as a) improvement in both the effectiveness and quality of learning (e.g., Brown & Harris, 2013; Ramdass & Zimmerman, 2008; Topping, 2003); b) the use of self-regulation strategies for learning (Kostons et al., 2012; Panadero et al., 2016) and c) self-efficacy (Olina & Sullivan, 2004; Ramdass & Zimmerman, 2008; c.f., Andrade et al., 2009). SSA is also deemed important in that it produces active learning by involving students in the evaluation process (Black & William, 1998; Nicol & McFarlane-Dick, 2006; Tan, 2012; Taras, 2010).

The role that SSA has taken in educational research, combined with the heterogeneity in its conception regarding its purpose, its execution, and even its interpretation, has produced several methodological approaches to its study. Consequently, different ideas are associated with the validity of SSA results (e.g., Boud & Falchikov, 1989; Falchikov & Boud, 1989; Gordon, 1991; Panadero & Alonso-Tapia, 2013; Ros, 2006; Ward et al., 2002). The range ways to implement or develop experiences with SSA requires a theoretical framework that facilitates its effective implementation.

Ross (2006) has reviewed the reliability, validity, and usefulness of SSA in education. He highlights some of the characteristics of SSA that are relevant to effectively implementing it in the classrooms, such as a) its reliability as a technique, b) its validity as evidence of the student's performance, c) its formative nature, and d) its usefulness as an evaluation technique. Similarly, Panadero et al., (2016) proposed different factors that could impact the validity of SSA. These are: a) the medium of SSA, b) the delay between SSA and instruction, c) expectations of students, and d) whether or not criteria for evaluation are provided.

When considering the variables proposed to affect the validity of SSA, it is necessary to define what is meant with regard to validity. In general terms, validity refers to the evidence provided to support or refute the meaning or explanation given to the evaluation data or results; "To validate a proposed interpretation or use of test scores is to evaluate the claims being based on the test scores. The specific mix of evidence needed for validation depends on the inferences being drawn and the assumptions being made." (Kane, 2006, p. 131). SSA Validity has been defined as the degree of agreement, rapprochement, or consistency that exists between the student's evaluations and those of the teachers (e.g., Andrade, 2019; Gordon, 1991; Ros, 2006).

In the case of SSA, the evaluation made by the teacher is considered as an expert evaluation against which to compare the scores that students provide during SSA. In accord with Kane (23006), the validity of SSA should be greater the more consistent student achievement evaluations are with teacher evaluations. Data in the literature on SSA validity appear inconclusive. Authors such as Boud and Falchikov (1989) indicated that, although there was a moderate consensus between SSA and the judgments expressed by professors, most of the literature suffered from a number of errors; methodological, conceptual, and interpretative, making a general picture of the findings complex. Some of the main limitations are summarized below:

- The evaluation scales used were not specific (e.g., Boud & Tyree, 1979).
- Students and teachers used different evaluation criteria (Doleys & Renzaglia, 1963; Gaier, 1961; Keefer, 1971; Mueller, 1970; Murstein, 1965; Sumner, 1932).
- The dependent variables measure more than knowledge gained. For example, Davis and Rand (1980) asked students to report their overall course performance, without making a clear distinction between performance and effort expended. Thus, the dependent variable sometimes reflected the effort invested by the student, providing a high degree of subjectivity

and variability (e.g., Filene, 1969; Davis & Rand, 1980).

- Abuse of reporting correlations between SSA and teacher evaluation (e.g., Gaier, 1961; Doleys & Renzaglia, 1963; Morton & Macbeth, 1977). The correlation coefficient is sensitive to atypical scores, and is not the best technique when assessing a group that may not be heterogeneous with respect to variables that could affect the correlation. Interpretations that compare SSA and teacher evaluation studies assume that individuals within the group share the same SSA ability (e.g., Ward et al., 2002).

There is a consensus that a moderate correlation between SSA and teacher evaluations exists, but there are also cases of a lack of correspondence that could call into question the validity of SSA. Some studies show trends towards overvaluation or undervaluation (precision errors). The variables that conclusively predict these trends are not known (Boud & Falchikov, 1989). The most common precision error reported is overestimation. It is generally assumed, with some exceptions, that students' SSA are higher than that of the teachers. There are also results that indicate that cognitive ability relates to precision errors. Overestimation occurs in young children, for example, and that has been attributed to the absence of the ability to assess achievement based on a criterion (Butler, 1990).

Students who usually achieve better marks tend to be more precise/realistic (e.g., Cochran & Spears, 1980; Doleys & Renzaglia, 1963; Keefer, 1971; Murstein, 1965), or even underestimate (e.g., Sumner, 1932) their performance, while students who usually achieve worse marks tend to overvalue (e.g., Daines, 1978; Moreland et al., 1981). When analyzing the relationship between students' academic achievement and SSA accuracy, students tend to be grouped into two groups, high and low academic achievement. This clustering may be overshadowing part SSA/Teacher evaluation relationship. Analyzing the relationship by grouping students into finer-grained levels of academic achievement may provide greater sensitivity to changes in SSA accuracy.

A variable that consistently affects accuracy with regard to the scores given by teachers is the experience or ability that the student has with the subject to be evaluated. Results show better accuracy in SSA with an increase of experience or mastery of the subject (e.g., Ross et al., 1999; Sung et al., 2005; Longhurst & Norton, 1997; Ross, 1998).

Knowledge of the evaluation criteria is also an important variable. When students do not know the evaluation criteria, the SSA is less accurate (Panadero & Romero, 2014). This relationship implies that the students' experience with SSA and their previous knowledge about the evaluation criteria, as well as the instruments used for SSA, will help to achieve more accurate SSA relative to the teacher (expert) evaluation. Also, accuracy improves when the criteria are simple (Pakaslahti & Keltikangas-Järvinen, 2000), or when the students have participated in the development of evaluation criteria (Brown & Harris, 2013). However, the effect of this latter variable is inconsistent, and contradictory results have been found (Andrade et al., 2010; Orsmond et al., 2000).

Overall levels of education do not necessarily have much impact on SSA accuracy. Accuracy in primary school students (Brown & Harris, 2013; Finn & Metcalfe, 2014; Ross, 2006) and higher education (Boud & Falchikov, 1989; Falchikov & Boud, 1989) is very similar ($.30 < r < .50$). SSA can be affected by motivational factors, such whether the SSA affects, or not, the final grade (e.g., Boekaerts, 2011; Dunning et al., 2004; Tejeiro et al., 2012). Tejeiro et al., (2012) find that when the SSA does not influence the grade, the SSA is very similar to that of the teacher. When the SSA influences the grade, the discrepancy increases notably with overestimates becoming more likely (Boud & Falchikov, 1998). Moreover, different motivational components can influence SSA. In a meta-analysis, Sitzmann et al., (2010) concluded that affective factors (e.g., satisfaction with the evaluation outcomes) have a greater impact on SSA than those associated with cognitive learning, which they determined to be only moderately related to SSA accuracy.

Although the studies discussed so far have shown a moderate level of precision of SSA, the literature shows varied results on the validity of this test (e.g., Brown & Harris, 2014; Dunning et al., 2004; Eva & Regher, 2005; Lew et al., 2010). These works show that the agreement between SSA and other measures (test grade, expert judgments, etc...) is moderate only in the best of cases (e.g., Brown & Harris, 2013). Though correlations vary from .2 to .8, there are few studies that report correlations higher than 0.6 (e.g., Brown & Harris, 2013;

Falchikov & Boud, 1989). Nevertheless, Panadero et al., (2016) propose that, even being an imprecise measure with an unknown amount of error, its use continues to be beneficial for teaching practice (Andrade, 2010; McMillan & Hearn, 2008).

As we have discussed, SSA precision / validity can be affected by numerous methodological factors. Some of those variables described above have been considered in the analysis of SSA validity. However, we have identified no studies that explicitly controlled for the variables already identified in the literature as "threatening" as a whole, when assessing the accuracy of the SSA.

The main objective of the study we present is to assess the validity of SSA as evaluation method. The study was conducted with a natural university sample while controlling factors that threaten SSA's validity. We controlled the evaluation criteria (teacher and students were familiar with the evaluation criteria); the scale and rubric used (teacher and students used the same scale and rubric for the evaluation); all students had previous SSA experience; the SSA was conducted at the end of the course when student's knowledge should be at its highest; and the examiner was blind to the identity of the student being evaluated. Finally, we analyzed how the precision of SSA can vary according to the level of achievement acquired by the students beyond a simple pass-fail categorization (cf., Ćukušić, et al., 2014). To achieve this, we not only focus on students who pass or fail, but we also set different levels of academic achievement to ensure sensitivity in the accuracy of self-assessments made by students, with respect to the evaluation of the teacher (expert).

Method

Participants

Sixty-four students from the third year of the Degree in Social Education of the University of Jaén (Spain) participated in the experiment. The sample constitutes a natural group. The age of the students was between 20.54 and 39.62 years (Median 22.48 years). The sample was 82.54% female and 17.46% male. These percentages are proportional to the distribution of males and females in the total population of students in Spain (Spanish National Institute of Statistics, 2015). One male student chose not to participate, so the final number of participants was 63.

Instruments

The test that was used as a basis to evaluate the validity of the SSA was the final exam of the theoretical component one of the subjects of the third year of the degree in Social Education. This test consisted of six open-ended questions, four of them with a short answer (e.g., requiring 1 to 2 paragraphs to answer correctly) and the other two with a long answer (requires 1 to 2 sheets to respond correctly). The questions were drawn from the most prevalent theoretical contents of the subject. An example of a short question is "*What are the principles of guidance? Explain briefly what they are.*" An example of a

Table 1. Example of evaluation criteria by rubric.

Degree of achievement	Description of the achievement
Level 1(Excellent) 100%	The student includes the required content and demonstrates mastery of it. The information is relevant, accurate and written in a coherent manner.
Level 2 (Good) 50-75%	The student includes three parts of the required content and demonstrates mastery of it. The information is relevant, accurate and written in a coherent manner.
Level 3 (Medium) 25-50%	The student includes half of the required content and demonstrates partial mastery of the content. The information is partially relevant, accurate.
Level 4 (Poor) 0-25%	The student includes summary information of the required content and does not demonstrate mastery of it. The information is not relevant, accurate and is not written in a coherent manner.

long question is "*Explain the phases that should be followed in the development of an Orientation Program (as we saw in topic 3).*" The rubric for the evaluation of the questions is presented in Table 1. To evaluate the validity of the rubric, 5 expert teachers evaluated 5 exams with the rubric designed. The intraclass correlation coefficient was .962.

At the beginning of the written test students were given instructions to complete the test and provided the point distribution for the various question types (e.g., up to one point for each short question answered correctly and up to 3 points for each long question answered). The maximum score that could be obtained in the test was ten. Students were told that once the test was completed, they should calculate the grade that they expect to obtain on the test, based on their performance, and write it in a box titled "Self-Assessment" that appeared in the upper left of the sheet. In the study, the teacher's score was assumed to be the accurate one and both evaluations (expert-teacher and student) followed the same criteria.

Procedure

At the beginning of the course, students were informed about the evaluation criteria and were told in the course's practice sessions that SSA would be used to evaluate the work they gave to the professor. Thus, and in accordance with the recommendations of the literature (Doleys & Renzaglia, 1963, Gaier, 1961, Keefer, 1971, Mueller, 1970, Murstein, 1965, Sumner, 1932), we ensured that students were aware of the evaluation criteria and practiced in their use. In order to control the bias caused by lack of knowledge about the use of SSA, students received instruction on how to evaluate works using criteria established by a rubric, in the same way as the teacher would do when evaluating the students' exams. Thus, during SSA students were practiced at evaluating works using rubric-based criteria and used the same criteria as the teacher, reducing the influence of factors not related to the knowledge assessed in the exam (e.g., Filene, 1969; Davis & Rand, 1980). The classes on how to perform SSA were part of the content of the subject regarding how achievement can be quantified.

On the day of the written test, the students were seated, the exams were distributed, and the previously mentioned instructions were read. Participants were informed of the purpose of the investigation and that participation in it was entirely voluntary. Those who did not want to participate simply left the self-assessment box blank. Students were informed that any student who had participated in the study and wanted to receive information about the findings could email the teacher to receive the information. The degree of participation was 98.4%.

Before beginning the evaluation of the exams by the teacher, the first page of each exam containing the student's identifying data, the exam questions, and the SSA response was removed. Thus, the teacher was blind to the identity of the student and his/her self-assessment.

Statistical Analysis

The scores resulting from the evaluation by the teacher (hereinafter called Exam) and the SSA responses issued by the students (hereinafter called SSA) were recorded for each student. The results were evaluated initially using the Pearson correlation coefficient (r) and linear regression was used to characterize the relationship between the variables. Participants were further grouped into Fail (<5), Pass (≥ 5 and <7), Very Good (≥ 7 and <9) and Outstanding (≤9) categories based on the Exam. These groupings were not arbitrary but followed common evaluation practices in university education (e.g., as defined by Real Decreto 1125/2003, 5 de September). Student's t-tests were used to compare each groups Exam score to its' SSA. Percent agreement between the SSA and Exam groupings as well as Cohen's Kappa were calculated.

Effect sizes were computed using Hedges g, as it is useful with small sample sizes (Hedges, 1981). Bootstrapping was used to determine the confidence intervals around the effect sizes using methods described by Efron (1992). Five-thousand random samples (with replacement) were drawn from each Exam grouping. The bootstrapping, and the associated statistical estimates, were made with the *dabestr* R package (Ho et al., 2018).

Results

Table 2 shows the descriptive statistics for the Exam and Self-assessment (SSA) variables. As can be seen in these data, the means of both variables are very close (6.62 and 6.75). Overall, there was no difference between SSA and Exam scores, $F < 1$.

The inter-rater agreement between the SSA and the teacher's evaluation with respect to the grouping (Fail, Pass, Very Good, Outstanding) was good (63.5%, K =

.61). Figure 1 shows the overall correlation between SSA (X) and Exam performance (Y) on the left along with the regression line of best fit predicting Exam grade with SSA. The points at right show the distribution of the scores, where it is evident that there were no outliers or extreme scores in either the SSA or the Exam. Within Figure 1, at left, the different symbols represent the scores grouped by exam achievement. Overall, exam scores were predictable given SA scores, $r = .66$, $r^2 = .44$, $p < .0001$. The SSA scores of failing students were shifted to the right of the regression line, indicating that their SSA overestimated their exam score (1.99 points overestimation, $t(7) = 5.39$, $p = .001$). Those who passed were more accurate in their SA, with scores spanning the regression line (.292 overestimation, $t(24) = 1.21$, $p = .24$). Those doing "very good" under-estimated their final grade, with scores being shifted to the left of the regression line (-.40, $t(23) = 5.21$, $p < .0001$), as did those doing "Outstanding", who underestimated their grade, on average, by -.81, $t(5) = 2.74$, $p = .04$.

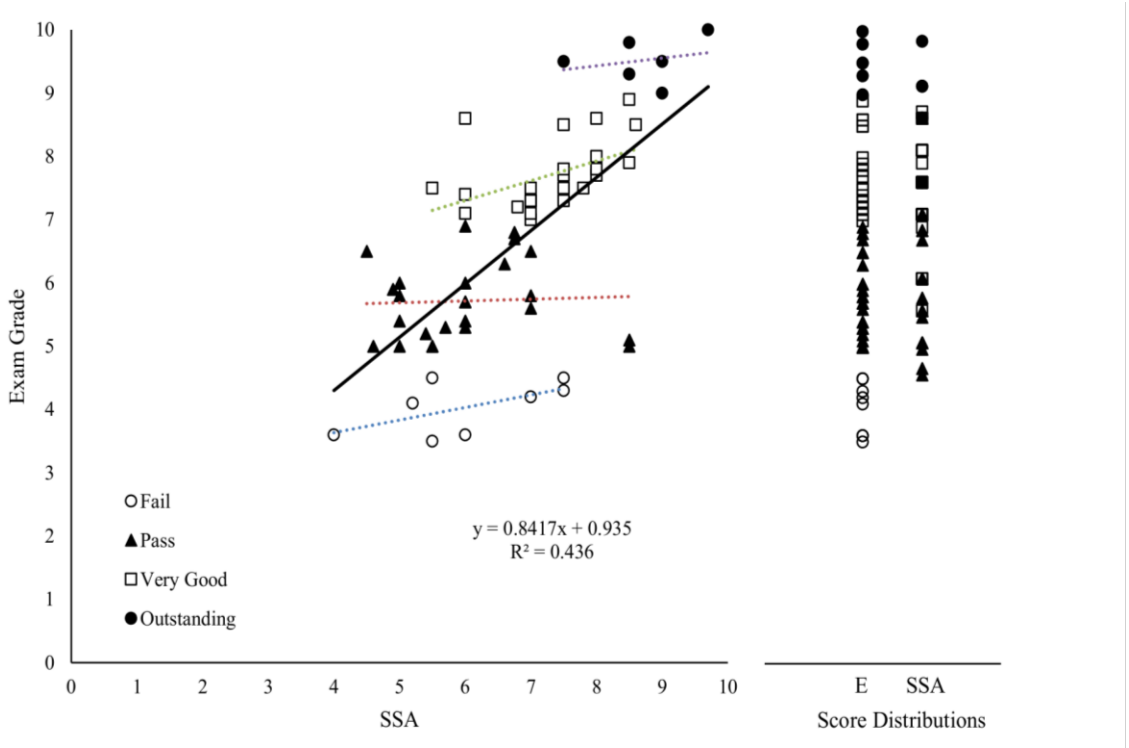
The regression lines within each grouping of exam performance (dotted lines in Figure 1) suggest little relationship between SSA and Exam performance, beyond the achievement (Fail, Pass, Very Good, and Outstanding) groupings. Examining the relationship between Exam Grade and SSA within each achievement grouping would yield weak conclusions due to the small sample sizes within each group. To remove the influence of the exam grouping, we subtracted each grouping's mean exam score from each exam score so.

Table 2. Descriptive statistics for the Exam and Self-assessment variables by level of achievement[1]

		Range	$\bar{x}$	S	S^2	Skew (Tip. Error.)	Kurtosis (Tip. Error)
Exam	Total	3.50-10	6.62	1.65	2.74	0.06 (0.3)	-0.75 (0.59)
	Fail	3.50-4.5	4.04	0.41	0.17	-0.27 (0.75)	-1.93 (1.48)
	Pass	5.00-6.9	5.72	0.62	0.38	0.52 (0.46)	-0.89 (0.90)
	Very Good	7.00-8.9	7.71	0.55	0.30	0.84 (0.47)	-0.37 (0.92)
	Outstanding	9.00-10	9.52	0.35	0.13	-0.08 (0.85)	-0.29 (1.74)
SSA	Total	4.00-9.7	6.75	1.30	1.69	0.04 (0.3)	-0.74 (0.59)
	Fail	4.00-7.5	6.03	1.23	1.52	-0.18 (0.75)	-0.77 (1.48)
	Pass	4.50-8.50	6.01	1.07	1.14	0.85 (0.46)	0.53 (0.90)
	Very Good	5.50-8.60	7.30	0.84	0.70	-0.45 (0.47)	-0.29 (0.92)
	Outstanding	7.50-9.70	8.70	0.73	0.54	-0.52 (0.85)	1.17 (1.74)

[1] *Note.* Exam refers to the variable measured by the Exam score determined by the teacher. SSA refers to the student's self-assessment. Total shows the scores by grouping all the students in the class. Fail, Pass, Very Good and Outstanding are classifications according to the score given by the teacher on the test.

Figure 1. Exam Grade by Self-Assessment Grade[2]



that each group had the same mean (zero). Thus, the only source of variation remaining in the exam scores was that within each achievement group. There was no relationship within achievement categories between the Exam and the SSA, r = .19, p = .13. The lack of relationship is, perhaps, not surprising given that there is a restricted range of exam scores within each exam achievement category.

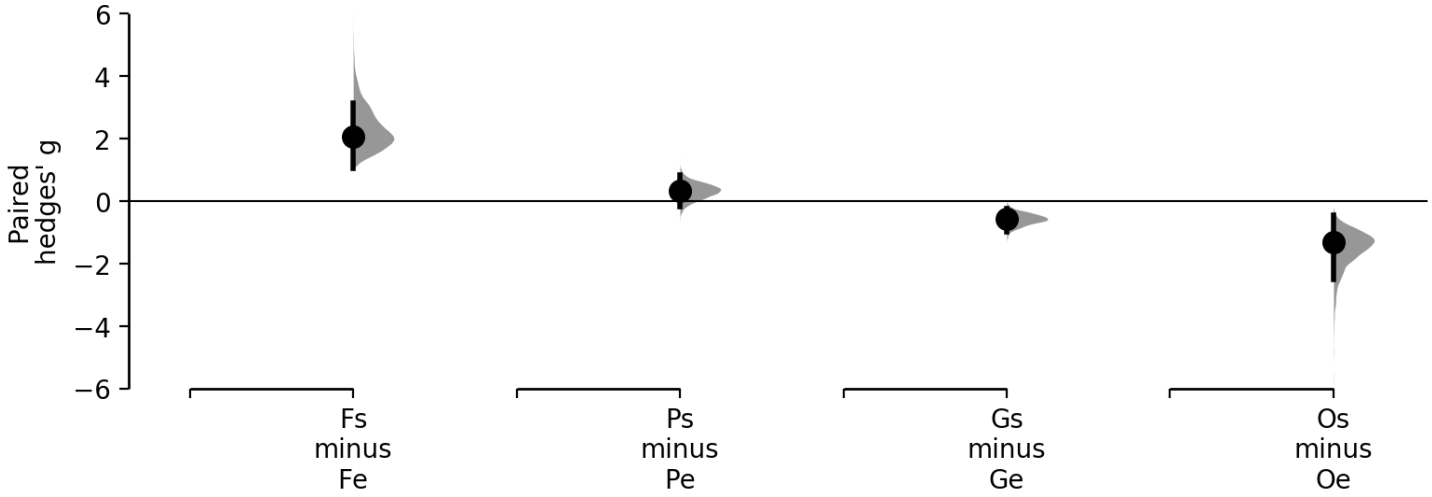
Given that the sizes of the achievement groupings were different, a bootstrap analysis (5000 samples) was performed to estimate the population effect size of each comparison. Figure 2 shows the results of the bootstrapping (F: Fail, P: Pass, V: Very Good, and O: Outstanding) for difference between the SSA assessment (subscript s) and the exam score (subscript e). Positive effect sizes show an overestimation of performance by SSA and negative effect sizes show an under-estimation. Effects near zero indicate similar SSA and Exam scores. The following effect sizes were obtained: Fail g = 2.04, [$CI_{95\%}$ 0.74, 3.18], Pass, g = .33 [$CI_{95\%}$ -0.197, 0.88], Very Good g = -.57, [$CI_{95\%}$ -1.12, -0.05] and Outstanding, g = -1.31, [$CI_{95\%}$ -2.45, -0.10]. As can be seen in the figure, the Fail group shows a robust overestimation, the Pass and Very Good groups show an accurate and slight underestimation, respectively, and the Outstanding group shows a robust underestimation.

Conclusions

The main objective of this work was to analyze the relationship between academic achievement and SSA accuracy in university education while ensuring that teachers and students used the same evaluation criteria (cf., Admiraal et al., 2015), used the same rubric and

[2] The vertical axis represents the scores obtained by the students in the exam (Exam Grade), on the horizontal axis are Self-Assessment (SSA) scores. The left part of the figure shows the linear regression between SSA and Exam Grade. The regression equation is shown at bottom right and the overall line of best fit is shown in black. Scores are shown by group (Fail = open circles, Pass = black triangles, Very Good = open squares, and Outstanding = black circles) with the lines of best fit within groups shown in dashed lines (Blue for Fail, red for Pass, green for Very Good and purple for Outstanding). The distributions of the scores for the variable Exam (E) and for the variable Self-assessment (SSA) are shown to the right of the figure.

Figure 2. Paired Hedges' *g* for the 4 groups (F: Fail, P: Pass, V: Very Goods, and O: Outstanding)[3]



scales (cf., Alameddine et al., 2018), that students were experienced in SSA (cf., Bolívar-Cruz, & Verano-Tacoronte, 2018), and that the teacher was blind to the authors of the works being evaluated (cf., Aryadoust, 2015). The results showed that students' estimates of their achievements did not differ, overall, from the evaluations made by the teacher. That correspondence demonstrates a degree of SSA validity since, overall, the students evaluated their performance similarly as did the teacher. Further, we analyzed SSA precision with respect to overall academic achievement based on standard classifications of Fail, Pass, Very Good, and Outstanding. Those standard classifications revealed biases in SSA, where lower achievers tended to over-estimate their performance while higher achievers underestimated their performance. Defining validity as the agreement between the evaluations made by the students and the teacher, our results show that SSA is a valid assessment of achievement. In Andrade's (2019, p.5) terms, there is consistency between the SSA and teacher's evaluations as the data show, "the degree of alignment between students' and expert raters' evaluations, avoiding the purer, more rigorous term accuracy unless it is fitting".

In the research presented here, all factors classified by the SSA-accuracy literature as threatening have been considered; Knowledge that students have about the evaluation criteria, the experience that students have with SSA, and familiarity with the scale, and standardization of the scale used (e.g., Panadero et al., 2013; Panadero et al., 2016). Following these recommendations, our participants had prior experience and training in evaluating work using evaluation criteria similar to that used on the SSA.

In the present work the evaluation criteria were established previously and were provided to the students before SSA. Additionally, we explained to the students what SSA is and spent several sessions on its use to familiarize them with it. To help ensure maximum objectivity regarding the teacher's evaluation, a rubric was developed prior to the evaluation and validated by five experts not involved in the evaluation (e.g., Brown & Harris, 2013; Panadero et al., 2012). In addition, the teacher was blind during the evaluation process to avoid any bias regarding the score assigned to each participant. The same, familiar, measurement scale was used both for teacher evaluation and for SSA.

A common feature in SSA studies is to analyse the results considering the class as a homogeneous group (e.g., Bolívar-Cruz, & Verano-Tacoronte, 2018).

[3] Comparisons are shown in the Cumming estimation plot. The subscript e indicates Exam score; subscript s indicates SSA score. Each paired mean difference is plotted as a bootstrap sampling distribution. Mean differences are depicted as dots; 95% confidence intervals are indicated by the ends of the vertical error bars.

Numerous factors and skills are unevenly distributed among students in a class (e.g., capacity, performance, motivation, etc...), and these factors could affect the ability to assess their own achievements. We provided additional analysis of the results by grouping students depending on their test results. The results have shown that this factor (the level of student achievement within the course) is related to the ability of students to self-assess.

The results were analyzed with respect to different levels of student achievement to enhance sensitivity beyond the results of a binary pass/fail classification (c.f., Bolívar-Cruz, & Verano-Tacoronte, 2018). This decision is consistent with the specialized literature (e.g., Brown & Harris, 2014; Panadero et al., 2016), which argues for the importance of considering such factors in the construction and design of evaluation instruments as their absence could be detrimental. For instance, Brown and Harris (2014, p. 23) affirmed "… awarding grades or basing educational interventions or changes based on unrealistic or construct-irrelevant self-assessments is untenable. If self-assessment processes lead students to conclude wrongly that they are good or weak in some domain and they base personal decisions on such false interpretations, harm could be done, even in classroom settings (e.g., task avoidance, not enrolling in future subjects)".

Our findings are consistent with those of Sumer (1932) showing that better students tend to underestimate their performance, while those of lesser skill tend to overestimate (e.g., Daines, 1978; Moreland et al., 1981). Such an effect is obfuscated when analysing the class as a whole or using a simple pass/fail distinction. Despite the over and under estimation, accuracy was good. Overall, 66.7% of the students SSA scores were within 1 point of the teachers' score, with that portion being 12.5% among those Failing, 72% among those Passing, 79% for those classified as Very Good, and 66.7% for those in the outstanding category.

The proportion of students in each performance category (Fail: 12.60%, Pass: 39.68%, Very Good: 38.10%, Outstanding: 9.52%) is representative of the distribution of those proportions in the population of university students in Spain (e.g., Aranda, et al., 2013). Due to their different sample sizes, we used the bootstrap analysis to calculate the effect sizes in each group, which reaffirmed the previous analysis regarding the over and underestimation of scores by SSA.

Use of the student's achievement level regarding the subject matter as a moderating factor when analysing accuracy in SSA has been proposed in the literature. However, in contrast with our research, studies have tended to focus on high and low competition rankings (e.gr., Brown & Harris, 2013; Boud & Falchikov, 1989). Although results have varied, most of these studies relate higher levels of student achievement with better accuracy in SSA, and lower levels of student achievement with overestimation in SSA. The measure of the student's capacity or competence is based on a post-hoc measure (i.e., the exam), that is also the instrument being compared to the SSA. Additional measures of competency, based perhaps on pre-test measures of classroom performance, could further refine the relationship of competence and accuracy in SSA.

A possible explanation of these results could be found in the "*Dunning-Kruger effect*" (Kruger & Dunning, 1999). According to the authors, the effect is a cognitive bias in which people erroneously assess their cognitive ability as greater than it is. Kruger and Dunning (1999) propose that students lack not only knowledge of the content, they also lack metacognitive skills to recognize that they do not possess that knowledge. The cognitive bias, in "less capable students", is attributed to an internal illusion about their own cognitive abilities, leading to overestimations of performance. In the case of "more capable students" the cognitive bias may arise from an erroneous perception about the external assessment of their competences being more rigorous than it is, which leads them to underestimate their performance on the assessment. Our findings suggest that a-priori assessment of metacognitive skills (e.g., Kallio et al., 2018) would help to quantify the relationship of SSA and exam performance.

The role that SSA has in the processes of Self-Regulation-Learning (SRL) and on Self-efficacy has been the focus of many investigations (see for a review Panadero et al., 2017). However, as far as we know, there are no studies in the literature where, in addition to evaluating the formative role of the SSA, they assess the relationship between the accuracy of the students in SSA and its formative benefits. Knowing what factors positively or negatively affect the SSA accuracy

should have an impact on the formative role in the SRL process.

In summary, the results found in this study show that the validity SSA as an evaluative test in university education is somewhat relative to the skills of the student. When validity of the test was assessed across the entire group, the data show a general correspondence between the SSA and the expert evaluation. However, when students are grouped according to their achievement in the evaluation by the teacher, the measurement of validity is compromised by the student's expertise in the subject. There is an overvaluation in the less competent students, while the more competent students show undervaluation. These findings suggest that SSA may be valid test for evaluating group performance, but less so for the individual in the absence of further knowledge about that individual. Assessing metacognitive abilities related to illusion of control may be one way to better adjust individual SSA evaluations.

Strengths, Limitations, and further research

The present study provides an accurate validation of SSA with university sample while controlling factors that the literature indicates influence the accuracy of SSA. Analysis considered both the whole group and different academic achievement sub groups, beyond pass-fail. Limitations include the use of a natural class resulting in a small sample, requiring the use of a bootstrap analysis to infer a population effect. Further research should consider larger and more diverse samples (e.g. in other educational levels) and the measurement of other indicators of achievement and metacognitive factors that might mediate the accuracy of SSA.

References

Admiraal, W., Huisman, B., & Pilli, O. (2015). Assessment in massive open online courses. *Electronic Journal of E-learning, 13*(4), 207-216. Retrieved from https://academic-publishing.org/index.php/ejel/article/view/1728/1691

Ainscow, M. (1988). Beyond the eyes of the monster: an analysis of recent trends in assessment and recording. *Support for learning, 3*(3), 149-153. https://doi.org/10.1111/j.1467-9604.1988.tb00088.x

Alameddine, M. B., Englesbe, M. J., & Waits, S. A. (2018). A video-based coaching intervention to improve surgical skill in fourth-year medical students. *Journal of surgical education, 75*(6), 1475-1479. https://doi.org/10.1016/j.jsurg.2018.04.003

Andrade H. L. (2019) A Critical Review of Research on Student Self-Assessment. *Front. Educ.* 4:87. https://doi.org/10.3389/feduc.2019.00087

Andrade, H. L. (2010). Students as the definitive source of formative assessment: academic self-assessment and the self-regulation of learning. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 90–105). New York: Routledge.

Andrade, H., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education: Principles, Policy & Practice, 17(*2), 199–214. https://doi.org/10.1080/09695941003696172

Andrade, H., Wang, X. L., Du, Y., & Akawi, R. L. (2009). Rubric-referenced self-assessment and self-efficacy for writing. *Journal of Educational Research, 102*(4), 287–301. https://doi.org/10.3200/JOER.102.4.287-302

Aranda, A. F., Pastor, V. M. L., Oliva, F. J. C., & Romero, R. (2013). La evaluación formativa en docencia universitaria y el rendimiento académico del alumnado. *Aula abierta, 41*(2), 23-34. Retrieved from https://dialnet.unirioja.es/descarga/articulo/4239063.pdf

Aryadoust, V. (2015). Self-and peer assessments of oral presentations by first-year university students. *Educational Assessment, 20*(3), 199-225. https://doi.org/10.1080/10627197.2015.1061989

Berry, R. (2011). Assessment Reforms Around the World. *Assessment Reform in Education*, 89–102. https://doi.org/10.1007/978-94-007-0729-0_7

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: principles, policy & practice, 5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Boekaerts, M. (2011). Emotions, emotion regulation, and self-regulation of learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 408–425). New York: Routledge.

Bolívar-Cruz, A., & Verano-Tacoronte, D. (2018). Self-assessment of the oral presentation competence: Effects of gender and student's performance. *Studies in Educational Evaluation, 59*, 94-101. https://doi.org/10.1016/j.stueduc.2018.04.001

Boud, D. J., & Tyree, A. L. (1979). Self and peer assessment in professional education: a preliminary study in law, *Journal of the Society of Public Teachers of Law, 15*, 1, 65-74. Retrieved from https://heinonline.org/HOL/LandingPage?handle=hein.journals/sptlns15&div=1&src=home

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

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). Thousand Oaks: 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

Butler, R. (1990). The effects of mastery and competitive conditions on self-assessment at different ages. *Child Development, 61*, 201-210. https://doi.org/10.2307/1131059

Cain, K. M., Wilkowski, B. M., Barlett, C. P., Boyle, C. D., & Meier, B. P. (2018). Do we see eye to eye? Moderators of correspondence between student and faculty evaluations of day-to-day teaching. *Teaching of Psychology*, *45*(2), 107-114. https://doi.org/10.1177/0098628318762862

Cochran, S.B. & Spears, M.C. (1980). Student self-assessment and instructors' ratings: a comparison, *Journal of the American Diatetic Association, 76*, 253-257. Retrieved from http://agris.fao.org/agris-search/search.do?recordID=US8045887

Ćukušić, M., Garača, Ž., & Jadrić, M. (2014). Online self-assessment and students' success in higher education institutions. *Computers & Education, 72*, 100-109. https://doi.org/10.1016/j.compedu.2013.10.018

Daines, J.M. (1978). Self evaluation of academic performance in a continuously assessed course of study, *Research Intelligence, 4*, 1, 24-26. https://doi.org/10.1080/0141192780040106

Davis, J.K. & Rand, D.C. (1980). Self-grading versus instructor grading, *Journal of Educational Research, 73*(4), 207-211. https://doi.org/10.1080/00220671.1980.10885237

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. https://doi.org/10.1080/03075079912331379935

Doleys, E. J., & Renzaglia, G. A. (1963). Accuracy of student prediction of college grades. *Personnel & Guidance Journal, 41*(6), 528–530. https://doi.org/10.1002/j.2164-4918.1963.tb02337.x

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. https://doi.org/10.1111/j.1529-1006.2004.00018.x

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics* (pp. 569-593). Springer, New York, NY.

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

Filene, P.O. (1969). Self-grading: an experiment in learning, *Journal of Higher Education, 40*, 451-458. https://doi.org/10.2307/1979820

Finn, B., & Metcalfe, J. (2014). Overconfidence in children's multi-trial judgments of learning. *Learning & Instruction, 32*, 1–9.

https://doi.org/10.1016/j.learninstruc.2014.01.001

Gaier, E. L. (1961). Student self estimates of final course grades. *The Journal of genetic psychology, 98*(1), 63-67. https://doi.org/10.1080/00221325.1961.10534353

Gordon, M. J. (1991). A review of the validity and accuracy of self-assessments in health professions training. *Academic medicine: journal of the Association of American Medical Colleges, 66*(12), 762-769. https://doi.org/10.1097/00001888-199112000-00012

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect size and Related Estimators. *Journal of Educational Statistics, 6*(2), 107-128. https://doi.org/10.3102/10769986006002107

Ho, J., Tumkaya, T., Aryal, S., Choi, H., & Claridge-Chang, A. (2018). Moving beyond P values: Everyday data analysis with estimation plots. *Nat Methods 16*, 565–566. https://doi.org/10.1038/s41592-019-0470-3

Instituto Nacional de Estadística (2015). *Mujeres en el Profesorado Por Enseñanza que Imparten] [Females in the Teaching Body According to the Grade Level They Teach.* Available at: http://www.ine.es/ss/Satellite?L=es_ES&c=INESeccion_C&cid=1259925481851&p=1254735110672&pagename=ProductosYServicios%2FPYSLayout&param3=1259924822888

Kallio, H., Virta, K. & Kallio, M. (2018). Modelling the Components of Metacognitive Awareness. *International Journal of Educational Psychology, 7*(2), 94-122. https://doi.org/10.17583/ijep.2018.2789

Kane, M. (2006). Content-related validity evidence in test development. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 131–153). Mahwah, NJ: Lawrence Erlbaum Associates

Keefer, K. E. (1971). Characteristics of students who make accurate and inaccurate self-predictions of college achievement. *The Journal of Educational Research, 64*(9), 401-404. https://doi.org/10.1080/00220671.1971.10884203

Kostons, D., van Gog, T., & Paas, F. (2012). Training self-assessment and task-selection skills: a cognitive approach to improving self-regulated learning. *Learning and Instruction, 22*(2), 121–132. https://doi.org/10.1016/j.learninstruc.2011.08.004

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Longhurst, N., & Norton, L. S. (1997). Self-assessment in coursework essays. *Studies in Educational Evaluation, 23*(4), 319-330. https://doi.org/10.1016/s0191-491x(97)86213-x

McMillan, J. H., & Hearn, J. (2008). Student self-assessment: the key to stronger student motivation and higher achievement. *Educational Horizons, 87*, 40–49. Retrieved from https://pdfs.semanticscholar.org/bb37/9a8107153678f41b65ac68d78726ba065f26.pdf?_ga=2.39662499.1836272783.1575369325-792685113.1573028967

Moreland, R., Miller, J. & Laucka, F. (1981) Academic achievement and self-evaluation of academic performance. *Journal of Educational Psychology, 73*(3), 335-344. https://doi.org/10.1037/0022-0663.73.3.335

Morton, J.B. & Macbeth, W.A.A.G. (1977). Correlations between staff, peer, and self-assessments of fourth-years students in surgery, *Medical Education, 11*(3), 167-170. https://doi.org/10.1111/j.1365-2923.1977.tb00586.x

Mueller, R. H. (1970). Is self-grading the answer? *The Journal of Higher Education, 41*(3), 221-224. https://doi.org/10.2307/1977312

Murstein, B. I. (1965). The relationship of grade expectations and grades believed to be deserved to actual grades received. *The Journal of Experimental Education, 33*(4), 357-362. https://doi.org/10.1080/00220973.1965.11010894

Nicol, D., & McFarlane-Dick, D. (2006). Formative assessment and self-regulated learning, a model

and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Noonan, B., & Duncan, C. R. (2005). Peer and self-assessment in high schools. *Practical assessment, research and evaluation, 10*(17), 1-8. Retrieved from https://pareonline.net/pdf/v10n17.pdf

Olina, Z., & Sullivan, H. J. (2004). Student self-evaluation, teacher evaluation, and learner performance. *Educational Technology Research and Development, 52*(3), 5–22. https://doi.org/10.1007/BF02504672

Orsmond, P., Merry, S., & Reiling, K. (2000). The use of student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education, 25*(1), 23–38. https://doi.org/10.1080/02602930050025006

Pakaslahti, L., & Keltikangas-Järvinen, L. (2000). Comparison of peer, teacher and self-assessments on adolescent direct and indirect aggression. *Educational Psychology, 20*(2), 177-190. https://doi.org/10.1080/713663710

Panadero, E., & Alonso-Tapia, J. (2013). Self-Assessment: Theoretical and Practical Connotations. When It Happens, How Is It Acquired and What to Do to Develop It in Our Students. *Electronic Journal of Research in Educational Psychology, 11*(2), 551-576. https://doi.org/10.1016/j.stueduc.2013.04.001

Panadero, E., & Alonso-Tapia, J. (2014). How do students self-regulate? Review of Zimmerman's cyclical model of self-regulated learning. *Anales De Psicologia, 30*(2), 450–462. https://doi.org/10.6018/analesps.30.2.167221

Panadero, E., & Romero, M. (2014). To rubric or not to rubric? The effects of self-assessment on self-regulation, performance and self-efficacy. *Assessment in Education: Principles, Policy & Practice, 21*(2), 133–148. https://doi.org/10.1080/0969594X.2013.877872

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. 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., Brown, G., & Courtney, M. (2014). Teachers' reasons for using self-assessment: A survey self-report of Spanish teachers. *Assessment in Education: Principles, Policy & Practice, 21*(4), 365-383. https://doi.org/10.1080/0969594x.2014.919247

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. https://doi.org/10.1016/j.edurev.2017.08.004

Ramdass, D., & Zimmerman, B. J. (2008). Effects of self-correction strategy training on middle school students' self-efficacy, self-evaluation, and mathematics division learning. *Journal of Advanced Academics, 20*(1), 18– 41. https://doi.org/10.4219/jaa-2008-869

Real Decreto 1125/2003, de 5 de septiembre, por el que se establece el sistema europeo de créditos y el sistema de calificaciones en las titulaciones universitarias de carácter oficial y validez en todo el territorio nacional. *Boletín Oficial del Estado*, 224, de 18 de septiembre de 2003, 34,355-34,356. Retrieved from https://www.boe.es/eli/es/rd/2003/09/05/1125

Ross, J. A. (2006). The Reliability, Validity, and Utility of Self-Assessment. *Practical Assessment Research & Evaluation, 11*(10) 10. https://doi.org/10.7275/9wph-vv65

Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (1998). Skills training versus action research in-service: Impact on student attitudes to self-evaluation. *Teaching and Teacher Education, 14*(5), 463-477. https://doi.org/10.1016/s0742-051x(97)00054-1

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15*(1), 1-20. https://doi.org/10.1177/026553229801500101

Sitzmann, T., Ely, K., Brown, K. G., & Bauer, K. N. (2010). Self-assessment of knowledge: A cognitive learning or affective measure? *Academy of Management Learning & Education, 9*(2), 169-191. https://doi.org/10.5465/amle.9.2.zqr169

Stanton, H.E. (1978) Self-grading as an assessment method. *Improving College and University Teaching, 26*(4), 236-238. https://doi.org/10.1080/00193089.1978.9927589

Sumner, F.C. (1932). Marks as estimated by students, *Education, 32*, 429.

Sung, Y. T., Chang, K. E., Chiou, S. K., & Hou, H. T. (2005). The design and application of a web-based self-and peer-assessment system. *Computers & Education, 45*(2), 187-202. https://doi.org/10.1016/j.compedu.2004.07.002

Tan, K. H. K. (2012). *Student self-assessment. Assessment, learning and empowerment*. Singapore: Research Publishing.

Taras, M. (2010). Student self-assessment: processes and consequences. *Teaching in Higher Education, 15*(2), 199–209. https://doi.org/10.1080/13562511003620027

Tejeiro, R. A., Gómez-Vallecillo, J. L., Romero, A. F., Pelegrina, M., Wallace, A., & Emberley, E. (2012). Summative self-assessment in higher education: Implications of its counting towards the final mark. *Electronic Journal of Research in Educational Psychology, 10* (2), 789-812. Retrieved from http://investigacion-psicopedagogica.org/revista/articulos/27/english/Art_27_707.pdf

Topping, K. J. (2003). Self and peer assessment in school and university: reliability, validity and utility. In M. Segers, F. Dochy, & E. Cascallar (Eds.), *Optimising new modes of assessment: in search of qualities and standards* Vol. 1 (pp. 55–87). Netherlands: Springer.

Torres, K., & Tackett, S. (2016). Pre-service teachers' beliefs about teaching ESOL students in mainstream classrooms. *International Online Journal of Education and Teaching, 3*(3), 186-200. Retrieved from http://iojet.org/index.php/IOJET/article/view/130/135

Ward, M., Gruppen, L., & Regehr, G. (2002). Measuring self-assessment: current state of the art. *Advances in Health Sciences Education, 7*(1), 63-80. https://doi.org/10.1023/a:1014585522084

Ysseldyke, J.E. & Matson, D. (1988). Issues in the psychological evaluation of children. En V. Van Hasselt, P.S. Satrain, & M. Hersen (Eds.) *Handbook of developmental and physucal disabilities*. Nueva York: Pergamon

Citation:

León, S. P., Vallejo, A. P., & Nelson. J. B.. (2021). Variability in the accuracy of self-assessments among low, moderate, and high performing students in university education. *Practical Assessment, Research & Evaluation*, 26(16). Available online: https://scholarworks.umass.edu/pare/vol26/iss1/16/

Corresponding Author

Samuel Parra León
University of Jaén, Spain

email: sparra [at] ujaen.es