

This is an accepted manuscript of an article that is published by SAGE in ACTIVE LEARNING IN HIGHER EDUCATION on 2019.

Cite: González-Betancor, S.M., Bolívar-Cruz, A., & Verano-Tacoronte, D, (2019). Self-assessment accuracy in higher education: the influence of gender and performance of university students, *Active Learning in Higher Education*, 20 (2), 101-114.

© 2019 This manuscript version is made available under the CC-BY-NC-ND 4.0 license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

This document is the Accepted Manuscript version of a Published Work that appeared in final form in *Active Learning in Higher Education*. To access the final edited and published work see <https://doi.org/10.1177/1469787417735604>

Self-assessment accuracy in higher education: the influence of gender and performance of university students

Sara M. González-Betancor, Alicia Bolívar-Cruz and Domingo Verano-Tacoronte

Contact/biographical details of authors

Sara M. González-Betancor, Economía de la Salud y Políticas Públicas, Departamento de Métodos Cuantitativos en Economía y Gestión, University of Las Palmas de Gran Canaria, ULPGC, Las Palmas de Gran Canaria, Spain. Tel: +34 928 458 162. Email: sara.gonzalez@ulpgc.es

Alicia Bolívar-Cruz, Departamento de Economía y Dirección de Empresas, University of Las Palmas de Gran Canaria, ULPGC, Las Palmas de Gran Canaria, Spain. Email: alicia.bolivar@ulpgc.es

Domingo Verano-Tacoronte, Departamento de Economía y Dirección de Empresas, University of Las Palmas de Gran Canaria, ULPGC, Las Palmas de Gran Canaria, Spain. Email: domingo.verano@ulpgc.es

Sara M. González-Betancor is an Associate Professor of Statistics and Econometrics at the University of Las Palmas de Gran Canaria. Her research interests are in the field of economics of education, bibliometric research and economics of poverty.

Alicia Bolívar-Cruz is an Associate Professor of Business Administration at University of Las Palmas de Gran Canaria. Her research interests are focused in competence assessment and entrepreneurship.

Domingo Verano-Tacoronte is an Associate Professor of Human Resource Management at University of Las Palmas de Gran Canaria. His research interests are focused in performance evaluation and competence assessment.

Abstract

Self-assessment activities are increasingly popular in the classrooms. But regarding self-assessment accuracy, mixed results are reported without clear reasons for this. This paper addresses the impact of gender and performance on self-assessment accuracy. To evaluate self-assessment accuracy, peer and professor assessments are used as reference. A research project, relating to the assessment of oral presentations using scoring rubrics, is conducted in two university degrees. Data from 155 self-assessments and more than 12,000 peer and professor assessments are gathered. The analysis differentiates by degree, gender and level of performance in the assessed competency. The results show that self-assessment accuracy is low and related to the student's gender, that men rate themselves higher than women do, and that even using a scoring rubric students receiving higher marks from the professors are more accurate than students with lower marks.

Keywords

Self-assessment, higher education, oral skills, scoring rubrics

Self-assessment accuracy

Higher education has shown a growing tendency towards student-centered teaching-learning processes. In any profession that incorporates the principles of autonomy and self-regulation in its day-to-day functioning, people have to be aware of the need to self-assess and orient their ongoing learning throughout their careers. One of educators' responsibilities is to teach students to behave like professionals; that is, they should be capable of giving and receiving feedback and evaluating their own work and that of others, as a way to increase their professional competencies (Boud and Falchikov, 1989). One of the areas where students demonstrate their responsibility is in self-assessment, which refers to students' commitment to judging their own learning, especially regarding their achievements and results.

Students' participation in their own assessment presents a series of advantages (Falchikov, 2005; Gessa Perera, 2011; Ibarra Sáiz and Rodríguez Gómez, 2007; Marín-García, 2009; Topping, 2003; Gil Flores and Padilla Carmona, 2009; Dochy et al., 1999; Moreno-Murcia et al., 2014), which can be summarized as follows. It helps to develop competencies that are highly valued in the job market, such as a critical attitude toward their work. It increases students' involvement in their learning. It gives the professor time to spend on tasks with greater educational value. Specifically regarding the development of the oral presentation competency, self-assessment leads to improvements in grades, perceived learning, confidence in making better presentations, and the development of assessment skills (De Grez et al., 2012). The use of self-assessment is defended (Boud, 1989; Boud and Falchikov, 1989; Taras, 2010), even though its accuracy is less than optimal.

However, the practical application of self-assessment in the educational system is limited (Boud and Falchikov, 1989), especially if it is used to calculate students' final marks (Stefani, 1994), because teachers think students' assessments are not accurate enough (Brown and Harris, 2013; Panadero et al., 2014). If self-assessment leads people to think of themselves as good or bad in some area of knowledge, basing their personal decisions on misguided interpretations, a problem arises at both the educational and personal levels (Boud, 1989). Thus, if the accuracy of self-assessments can be guaranteed by comparing them with assessments from other sources (Boud, 1989; Brown and Harris, 2014; Brown and Harris, 2013), they can contribute to developing students' critical attitude toward their work, stimulating their ongoing learning in both academic and professional areas. Sources used to calibrate self-assessment include professors and peers (AlFallay 2004; De Grez et al., 2012; Langan et al., 2005, 2008; Ritchie, 2016).

Comparisons between peers and professors have shown better results in terms of accuracy than between self-assessment and professors (Langan et al., 2008; Patri, 2002; Ritchie, 2016), mostly when global, rather than multidimensional, criteria are used (Falchikov and Goldfinch, 2000). However, peer-assessment is not without problems. Thus, one problem linked to peer-assessment is that the range of peer marks is narrower than that of the professors (Langan et al., 2005). Furthermore, peers are inclined to be more generous than professors. Also, popular students could be awarded with high marks (Langan et al., 2008). However, since the students value positively peers' and professors' opinions (De Grez et al. 2012; Mulder et al., 2014), both sources must be considered to evaluate self-assessment accuracy.

However, the literature examining self-assessment accuracy does not offer robust results. Thus, in studies supporting its accuracy, the accuracy levels, generally established using correlation indexes, are usually positive, but without obtaining Pearson r values above 0.6 (Brown and Harris, 2013). For example, empirical results showing a high level of accuracy are reported in Dochy et al. (1999) and AlFallay (2004), while the studies by De Grez et al. (2012), Langan et al. (2008), Patri (2002) and Ward et al. (2002), among others, show low accuracy levels.

The literature points out a series of methodological and psychological reasons for this disagreement about self-assessment accuracy.

1. Regarding the activities being assessed, they have been quite different (for example, compositions, numerical exercises, cooperative work or oral presentations), so that it is not surprising to find divergent results (Marín-García, 2009). Specifically, the limited number of specific studies on oral presentation assessment makes it important to examine this process more closely to try to improve it (De Grez et al., 2012; Marín-García, 2009; Lew et al., 2010).
2. In addition, self-assessment accuracy has been verified through the analysis of correlations between students' marks and assessments by some external source, such as teachers or peers, leading to the consideration of three aspects (Ward et al., 2002). First, in the majority of the studies, the teachers' assessments are the standards used to compare the students' self-assessments. However, there are doubts about whether the marks are consistent (Topping, 2009; Falchikov and Boud, 1989), especially in assessing the oral presentation competency, where it is more difficult to find valid comparison patterns among expert raters (De Grez et al., 2012). Second, self-assessment accuracy has been analysed using the premise that raters behave as a coherent group, which is not likely because it would mean that all students measure the same aspects and use the assessment scale in the same way (Ward et al., 2002). Third, it should be kept in mind that not all students act in the same way when presented with the possibility of assessing their own performance. For example, students with higher levels on the assessed skill might demand more from themselves than those with lower skill levels (Brown and Harris, 2014; Kun, 2015).
3. Finally, it must be addressed the impact of differences between raters on self-assessment. Along these lines, a topic that sparks a lot of interest is the influence of the rater's gender on assessment accuracy (Falchikov and Magin, 1997; Archer, 1992). Although differences in self-assessment due to gender exist (Beyer, 1990), and they seem to be mainly due to lower perceptions of self-efficacy by women and less confidence in their own performance (Pallier, 2003), studies carried out in the educational setting are not conclusive (Boud and Falchikov, 1989). The same conclusion is reached when specifically analysing the assessment of the oral presentation competency. For example, Langan et al. (2005; 2008) and (Bolívar-Cruz et al., 2015) find a significant effect of the rater's gender, while Sellnow and Treinen (2004) do not. In addition, a significant influence of the student's gender has not been detected when comparing self-assessment accuracy to teachers' assessments, but gender differences have been found between peer-assessment and self-assessment (De Grez et al., 2012).

These problems indicate the need for a series of precautions to improve the results of self-assessments. Thus, first it is necessary to have more valid and reliable standards with which to compare the self-assessments, for example, by using more than one teacher or incorporating peer assessments (Ward et al., 2002). Various studies show that peer-assessment is more accurate than self-assessment (De Grez et al., 2012; Langan et al., 2008). It should be highlighted that these studies support students' capacity to identify the good or poor performance of others, but they also show that they are unable or reluctant to apply the same standards to their own performance. Second, the development of easy-to-use and reliable assessment formats with high content validity improves the standards used to compare self-assessments. One possibility is the use of rubrics, which are assessment tools that make it possible to rate the quality of students' contributions in different areas. Rubrics can also be used to rate students' performance levels. They describe, before doing the activity, the variables that will be assessed and the performance or completion levels for each (Andrade and Du, 2005; García-Ros, 2011; Jonsson and Svingby, 2007; Panadero and Romero, 2014). Rubrics reduce assessment subjectivity and, therefore, lead to a higher level of agreement on the marks given (Ramos-Alvarez and Luque, 2010; Panadero and Jonsson, 2013; Panadero and Romero, 2014). Third, providing teachers and students with training and experience in the use of assessment formats, in this case rubrics, should also contribute to improving assessment and self-assessment (Marín-García, 2009). Finally, attention should be paid to differences between raters. A segmented analysis of the group makes it possible to observe the phenomena of self-indulgence in the worst students and excessive self-demands in the best ones. This tendency, in turn, produces a correction of the 'real' marks offered by teachers or peers.

Based on these considerations, there is a need:

1. To analyse the accuracy of university students' self-assessment, compared to other assessment sources (peers and professors).
2. To find out whether self-assessment accuracy is related to the student's gender.
3. To discover whether, when classifying students according to the teachers' marks (that is, distinguishing between students with higher and lower marks on the oral presentation), there are differences in the self-assessment accuracy.

Method

A research project was carried out during 2012 in two courses from different university degree programmes housed at a small, public university located in the South of Spain: Labour Organization in Enterprises (required 6-credit course taught in the Labour Relations and Human Resources degree programme –LRHR–) and Industrial Organization and Business Administration (core course with 13.5 credits taught in the Industrial Engineering degree programme. In the former, 92 students participated (27 men and 65 women), while 63 students (47 men and 16 women) participated in the latter.

The experience consisted of giving an oral presentation in pairs. This presentation was assessed with a rubric containing ten previously validated criteria (Verano-Tacoronte et al., 2016). Each presentation was assessed by the speakers, their peers and two professors, reaching a total of 12,588 assessments, as shown in detail in Table 1. Students self-assessment represented 10% of the final course grade. Professors' mark also contributed to the final course grade (15%). Peers' score had no influence in the final course grade.

Insert Table 1 here

The rubric's reliability for these raters was tested through interrater consistency using Cronbach's alpha (García-Ros, 2011; Cortina, 1993). Good internal consistency was found in both degree programmes for the three groups of raters (Table 2).

Insert Table 2 here

The mark for each presentation was obtained by aggregating the scores on each criterion on the rubric. The final mark could range between a minimum of 10 and a maximum of 30 points. In the case of the peers' and professors' scores, the average was calculated among all the raters, and the quality of this central measure was tested using Pearson's variation coefficient (between 0.14 and 0.47 for peers; between 0 and 0.71 for professors), which indicated that it can be used without losing relevant information.

To rate the assessment accuracy (first specific objective), a graphic analysis was performed of the level of coincidence in the marks for each presentation awarded by the three rating sources. Two-sample t-tests of equality of means were also carried out.

To check the possible relationship between self-assessment and gender (second specific objective), the data were broken down by students' gender, and basic descriptive statistics were analysed for both programmes; tests of equality of means were performed. Finally, the linear correlation between the marks was analysed with the simple linear correlation coefficient, and to check for any other type of monotonic relation, Spearman's rank correlation coefficient was also calculated.

Graphic analysis was carried out to detect differences in the self-assessment accuracy of the subsamples of students who received higher and lower marks from the professors (third specific objective) in each programme. The students' placement in one group or the other was determined by constructing confidence intervals for the set of individuals from the relevant programme and gender. The students who fell outside the interval constructed as the mean score plus/minus a standard deviation for their reference group were considered the best/worst rated.

Finally, a multiple linear regression model was estimated to explore the differences in marks between the professor and the student assessed.

$$Y_i = \beta_0 + \beta_1 \cdot Degree_i + \beta_2 \cdot Gender_i + \beta_3 \cdot CompPres_i + u_i$$

Where: Y was the difference between the score given by the professor and the one given by the student assessed (if it had a negative value, the student self-assessed with a higher mark than the professor).

Degree took a value of 0 for Labour Relations and Human Resources and 1 for Engineering.

Gender took a value of 0 when the student was a man and 1 for a woman.

CompPres was a categorical variable that had 3 values: 1, when the oral presentation competence from the professor's point of view was low; 2, when it was medium; and 3, when it was high. Therefore, it was entered in the model in the form of two dummy variables.

Results

Figure 1 provides an initial approach to the level of agreement among the three collectives involved. The students' presentations are sorted from the worst to the best according to the grades awarded by the professors. In both, Engineering and Labour Relations and Human Resources, the professors' marks coincide more with those of the peers than with the self-assessments. In fact, the latter are higher than the other two in many cases.

Insert Figure 1 here

The descriptive statistics shown on Table 3 indicate that, among classmates, it is more common to give intermediate marks. However, the professors give a wider range of marks, as the difference between the professors' minimum and maximum marks is greater in both degree programmes than the range of marks awarded by the peers. In the case of self-assessment, the levels are always located above the minimum grades given by other raters. The same applies to the maximum grades, which indicates that self-assessments are systematically higher.

The ANOVA and subsequent Tukey test reveal that there are no significant differences between the means of the professors' and peers' assessments, while the differences between either of these two and self-assessment are statistically significant in both degrees, although more pronounced in Engineering than in Labour Relations and Human Resources. In fact, professors and peers could be considered as a homogeneous group according to their given scores in both degrees.

Insert Table 3 here

Breaking down the data by the students' gender, we analyse whether this differentiated self-assessment behaviour is maintained, regardless of gender. On average, peers scored their classmates' oral presentation skills higher than professors did, regardless of the student's gender. Moreover, the self-perception of this skill is, in general terms, higher than the classmates' perception, and this is more pronounced in men than in women. The difference is also greater in Engineering studies as opposed to Labour Relations and Human Resources, maybe because of the greater proportion of men in the former degree.

Insert Table 4 here

Analysing whether the apparent differences according to the gender of the speaker are statistically significant, the results show that in Engineering, the professors' ratings do not differ significantly by gender, and this is also true of the peers' ratings in both degrees. Even so, the differentiation by the students' gender is relevant because it becomes clear that, on average in both programmes, the men's

self-assessments are systematically higher than the women's, and these differences are statistically significant.

In light of the data in, there seems to be some agreement between the peers' and professors' ratings, but not between these values and self-assessment. To further examine this relationship, a correlation study was carried out and appears in Table 5.

Insert Table 5 here

A high linear correlation can be seen between the peers' and professors' ratings in Engineering (0.83 for men and women) and in Labour Relations and Human Resources (0.72 for men and 0.78 for women). However, the linear correlation between self-assessment and the other sources is only statistically significant for women in Labour Relations and Human Resources, although to a lesser degree (0.43 with professors and 0.47 with peers).

As almost no linear correlation was detected between self-assessment and peers' and professors' ratings, in spite of having seen the same presentation and used the same rubric, a non-parametric measure was also considered. Thus, we calculated Spearman's rho to check for any other type or monotonic relation, obtaining similar results (Table 6).

Insert Table 6 here

The analysis of the self-assessment behaviour of the students with better/worse oral presentation competence according to the professor was carried out by degree and, for Labour Relations and Human Resources, by gender. The sample of Engineering students, unlike the Labour Relations and Human Resources sample, combines both genders in each figure, as there were no statistically significant differences by gender (Figure 2).

Insert Figure 2 here

The presentations of the students who received the worst ratings from the professors, present greater disparity among the three assessment sources in both degree programmes. By contrast, for the students with higher marks, there is greater consensus among the raters. On the other hand, the behaviour of professors and peers doesn't show the same tendency: while on the presentations with low marks the professor is stricter than the peers, on those with high marks the professor is less demanding than the peers.

The multiple regression model that explores the determinants of the difference between professor and student marks reveals that men from Labour Relations and Human Resources with high oral presentation competence score themselves an average of 1.96 points higher than the professors (Table 7).

Insert Table 7 here

No statistically significant differences are detected between degrees. By contrast, the level of competence perceived by the professor does have an influence. In fact, regardless of the degree programme and gender, the students with a low competence level present a difference of 9.15 points more than the students with a high level of competence. This difference declines to 3.24 points for those who have an intermediate level compared to those with a high competence level.

Discussion

This study focuses on self-assessment and, more specifically, on students' realistic ratings of their own competencies, in this case, oral presentation skills. The literature review carried out shows that research on self-assessment accuracy offers inconclusive results and is lacking in rigour (Brown and Harris, 2013). There are studies that obtain high self-assessment accuracy (Boud et al., 2013; Panadero and Romero, 2014), while others (Kun, 2015; Marín-García, 2009) report a tendency towards self-assessment overevaluation. Therefore, given that students' academic and professional progress depends on making unbiased self-assessments, the present study follows the main recommendations found in the literature to correctly evaluate self-assessment accuracy and analyse the results by considering aspects that can influence them, such as the homogeneity of the assessment criteria (rubrics), and differences between raters (that is, gender, performance on the competency analysed).

The results indicate that, for the used sample, self-assessment is not accurate, which is in line with the studies of Kwan and Leung (1996), Magin and Helmore (2001) and Marín-García (2009), among others. Though professors and peers score in a similar way, students usually overrate themselves, even if all of them use the same scoring rubric. Various arguments can be used to explain this situation. First, self-assessment can be influenced by its effect on the final mark in the course, producing overrating compared to other assessment sources and rendering the rubric ineffective. The lack of a self-assessment habit and the fact that students do not participate in identifying the criteria are other possible explanations. Finally, coinciding with the conclusions of De Grez et al. (2012), the differences between professors' assessments and self-assessments may be due to the professors' greater experience in grading oral presentations. However, it should be kept in mind that the students are more accurate rating their classmates when performing peer-assessment (taking professors as a standard), than they are when performing their own self-assessment.

Regarding the differences among students, the results reveal that self-assessment accuracy is related to gender. Although professors' and peers' assessments are oriented in the same direction (in Engineering both groups find that the men present more oral presentation skills, while in Labour Relations and Human Resources it would be the women), the self-assessment behaviour is not as homogeneous (coinciding in the case of Engineering, but not in Labour Relations and Human Resources). In general, and regardless of the major, men present higher self-assessment scores than women. Moreover, no significant relationship is detected between self-assessment and assessments by professors and peers, except in the case of women in Labour Relations and Human Resources, although in this latter case these levels of agreement are lower than those found between peers and professors. It would be necessary to examine the reasons for this behaviour exhibited by male speakers, who systematically give themselves higher scores than those awarded by the other two collectives, in spite of belonging to branches as different as Industrial Engineering and Labour Relations and Human Resources.

In addition, and given that not all students are going to behave in the same way when assessing themselves, the findings show the existence of different behavioural patterns when dividing the sample according to the professors' assessment. It appears that the rubric makes it possible to unify the marks when the student doing the self-assessment is also a good speaker according to the professors' criteria. In the case of the worst speakers from the professors' point of view, their self-assessment is systematically higher than that of peers and professors. In spite of the rubric, the students with lower oral presentation skills balance their marks by rating themselves high. In the case of Labour Relations and Human Resources, this difference is even greater in men than in women. When students with high oral presentation skills assess themselves, the results for the Labour Relations and Human Resources degree deserve special attention: the men give themselves higher marks than those given by peers and professors, while the women tend to underrate themselves.

Regardless of the study findings, the incorporation of self-assessment and, above all, peer-assessment, has positive possibilities (Boud, 2007; De Grez et al., 2012; Langan et al., 2005; Kun, 2015). An important part of the literature considers self-assessment worthwhile, if only due to its positive influence on the students' learning (formative assessment). Self-assessment is an effective tool that enables students to combine various aspects of their learning, reflect on their achievements, and examine implications for their future training. Therefore, the most useful aspect of self-assessment would lie in its formative assessment dimension. It can improve skills and capabilities (Birenbaum and Dochy, 1996), including the capacity for lifelong learning, beyond the strictly academic realm (Boud, 2000; Boud and Falchikov, 2006). The study findings point out that peer-assessment is more accurate than self-

assessment. Therefore, peer-assessment could be introduced if summative assessment. A stronger involvement of students could be achieved. However, evaluations could be manipulated in order to benefit or harm certain classmates, regardless of their real performance.

In order to develop students' competence in performing accurate and realistic self-assessment, we propose a series of recommendations. One, increase students' training in self-assessment. Two, increase the number of self-assessment experiences, in order to improve their ability to assess themselves (Birenbaum and Dochy, 1996; Boud and Falchikov, 1989). Three, involve students in designing assessment scales (Falchikov, 2005) to increase their commitment to the system.

In spite of the study's relevant contributions, some limitations should be pointed out. First, since, only quantitative data are used, further studies should be developed to add qualitative data to find out why self-assessment is not accurate enough. In a summative self-assessment context, low self-assessment accuracy might not be related to a lack of students' skills, but to the desire to increase their marks. Mixed methods as research approach (Scoles, Huxham and McArthur, 2012) could be useful to aboard this issue in order to find out the actual reasons for low accuracy self-assessment. Moreover, future work is needed to explore whether or not the results would be the same in a task that was not given marks/grades. Second, it would have been desirable to have a control group of students whose final mark would not be affected by the self-assessment. This would make it possible to observe possible changes in self-assessment accuracy depending on the motivation level, examining whether the results are the same when the self-assessment is not used for summative purposes. Given that peers show a high degree of accuracy when compared to professors, this could be a way of controlling their knowledge about the rubric and their ability to apply it, using similar criteria to those of the professors, and without a stimulus. Future research could study students' opinions about the usefulness of self-assessment in improving their skills and competencies, and how these opinions affect self-assessment accuracy (Moreno-Murcia et al., 2014). Another limitation is related to the large differences in sample size between peers, students and professors. Maybe, this could be corrected by bootstrapping to sub-sample the biggest population. However, given the sample size and scope of the study, it might not reveal much more.

There are also limitations in terms of sample size, as this research was carried out only in one university, in one particular discipline, with only undergraduates and in only one country. As results may vary from discipline to discipline and from culture to culture, future work needs to look at different levels of student, different disciplines and different cultures. It could also be advantageous to increase the number of degree programmes studied to try to draw common conclusions about all of them, given their differences in demographic composition, and even the learning styles present in them. This result is especially interesting because, in spite of having carried out the experiences in two degree programmes that could have quite different characteristics, no significant differences were found between them. Thus, the question arises of whether the use of rubrics can minimize differences produced by the context. Finally, although the influence of gender on self-assessment accuracy has been examined, we did not explore the possible reasons for the differences between self-assessment and peer- and teacher-assessment, and this could be a topic for future research. In addition, this research focused on rating oral presentations, so future work could continue deepening in the assessment of different types of tasks.

We conclude that self-assessment is less accurate than peer assesment. Students tend to overrate their scores. However, it is important to point out the existence of a significant relationship between gender and self-assessment: men's self-assessment is less accurate than women's, regarding to teachers' and peer's assessments. This requires to examine the reasons behind these differences and to look for measures to improve self-assesment accuracy.

Funding

This research was supported by the Universidad de Las Palmas de Gran Canaria under Grant number CPIE2013/14 – GIE14.

References

- AlFallay I (2004) The role of some selected psychological and personality traits of the rater in the accuracy of self- and peer-assessment. *System* 32(3): 407–425.
- Andrade H and Du Y (2005) Student perspectives on rubric-referenced assessment. *Practical Assessment, Research & Evaluation* 10(3): 1-11.
- Archer J (1992) Sex bias in evaluations at college and work. *The Psychologist: Bulletin of the British Psychological Society* 5(5): 200-204.
- Beyer S (1990) Gender differences in the accuracy of self-evaluations of performance. *Journal of Personality and Social Psychology* 59(5): 960-970.
- Birenbaum M and Dochy F (1996) *Alternatives in assessment of achievements, learning processes and prior knowledge*, Boston, Dordrecht y Londres: Kluwer.
- Bolívar-Cruz A, Verano-Tacoronte D and González-Betancor SM (2015) Is University Students' Self-Assessment Accurate? In: Peris-Ortiz M and Merigó Lindahl JM (eds) *Sustainable Learning in Higher Education*. Springer International Publishing, 21-35.
- Boud D (1989) The role of self-assessment in student grading. *Assessment in Higher Education* 14(1): 20-30.
- Boud D (2000) Sustainable assessment: rethinking assessment for the learning society. *Studies in continuing education* 22(2): 151-167.
- Boud D (2007) Reframing assessment as if learning were important. In: Boud D and Falchikov N (eds) *Rethinking assessment in higher education: Learning for the longer term*. Londres: Routledge, 14-25.
- Boud D and Falchikov N (1989) Quantitative studies of student self-assessment in higher education - A critical analysis of findings. *Higher Education* 18(5): 529-549.
- Boud D and Falchikov N (2006) Aligning assessment with long-term learning. *Assessment and Evaluation in Higher Education* 31(4): 399-413.
- Boud D, Lawson R and Thompson DG (2013) Does student engagement in self-assessment calibrate their judgement over time? *Assessment & Evaluation in Higher Education* 38(8): 941-956.
- Brown GTL and Harris LR (2013) Student self-assessment. In: McMillan JH (ed) *The SAGE handbook of research on classroom assessment*. Thousand Oaks, CA: SAGE, 367-393.
- Brown GTL and Harris LR (2014) The future of self-assessment in classroom practice: Reframing self-assessment as a core competency. *Frontline Learning Research* 2(1): 22-30.
- Cortina JM (1993) What is coefficient alpha - An examination of theory and applications. *Journal of Applied Psychology* 78(1): 98-104.
- De Grez L, Valcke M and Roozen I (2012) How effective are self- and peer assessment of oral presentation skills compared with teachers' assessments? *Active Learning in Higher Education* 13(2): 129-142.
- Dochy F, Segers M and Sluijsmans D (1999) The use of self-, peer and co-assessment in higher education: a review. *Studies in Higher Education* 24(3): 331-350.
- Falchikov N (2005) *Improving Assessment Through Student Involvement: Practical solutions for aiding learning in higher and further education.*, London & New York: RoutledgeFalmer.
- Falchikov N and Boud D (1989) Student self-assessment in higher education - A metaanalysis. *Review of Educational Research* 59(4): 395-430.
- Falchikov, N and Goldfinch, J (2000). Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks. *Review of Educational Research* 70(3): 287–322.
- Falchikov N and Magin D (1997) Detecting gender bias in peer marking of students' group process work. *Assessment and Evaluation in Higher Education* 22(4): 385-396.
- García-Ros R (2011) Analysis and validation of a rubric to assess oral presentation skills in university contexts. *Electronic Journal of Research in Educational Psychology* 9(3): 1043-1062.
- Gessa Perera A (2011) La coevaluación como metodología complementaria de la evaluación del aprendizaje: Análisis y reflexión en las aulas universitarias. *Revista de Educación* (354): 345-346.
- Gil Flores J and Padilla Carmona MT. (2009) Participation of university students in the learning assessment. *Educación XX1* (12): 43-65.
- Ibarra Sáiz MS and Rodríguez Gómez G (2007) El trabajo colaborativo en las aulas universitarias: reflexiones desde la autoevaluación. *Revista de Educación* (344): 229-230.
- Jonsson A and Svingby G (2007) The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review* 2(2): 130-144.
- Kun AI (2016) A comparison of self versus tutor assessment among Hungarian undergraduate business students. *Assessment & Evaluation in Higher Education* 41(3): 350-367.
- Kwan K-P and Leung RW (1996) Tutor versus peer group assessment of student performance in a simulation training exercise. *Assessment and Evaluation in Higher Education* 21(3): 205-214.

- Langan AM, Shuker DM, Cullen WR, et al. (2008) Relationships between student characteristics and self-, peer and tutor evaluations of oral presentations. *Assessment & Evaluation in Higher Education* 33(2): 179-190.
- Langan AM, Wheeler CP, Shaw EM, et al. (2005) Peer assessment of oral presentations: Effects of student gender, university affiliation and participation in the development of assessment criteria. *Assessment and Evaluation in Higher Education* 30(1): 21-34.
- Lew MDN, Alwis WAM and Schmidt HG (2010) Accuracy of students' self-assessment and their beliefs about its utility. *Assessment & Evaluation in Higher Education* 35(2): 135-156.
- Magin D and Helmore P (2001) Peer and teacher assessments of oral presentation skills: how reliable are they? *Studies in Higher Education* 26(3): 287-298.
- Marín-García JA (2009) Students and lecturers as markers. Application in assessing oral presentations. *Revista Espanola De Pedagogia* 67(242): 79-98.
- Moreno-Murcia JA, Aracil A and Reina R. (2014) Assignment of responsibility in evaluation: A strategy adapted to the European higher education area. *Educación XX1* 17(1): 183-199.
- Mulder RA, Pearce JM and Baik C (2014) Peer review in higher education: Student perceptions before and after participation. *Active Learning in Higher Education* 15(2): 157-71.
- Pallier G (2003) Gender differences in the self-assessment of accuracy on cognitive tasks. *Sex Roles* 48(5-6): 265-276.
- Panadero E, Brown G and 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.
- Panadero E and Jonsson A (2013) The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review* 9 (June): 129-144.
- Panadero E and 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.
- Patri M (2002) The influence of peer feedback on self-and peer-assessment of oral skills. *Language Testing* 19(2): 109-131.
- Ramos-Alvarez M-M and Luque G (2010) A competence-based constructivist tool for evaluation. *Cultura y Educación* 22(3): 329-344.
- Ritchie, SM (2016) Self-assessment of video-recorded presentations: Does it improve skills? *Active Learning in Higher Education* 17(3): 207-221.
- Scoles J, Huxham M and McArthur J (2014) Mixed-methods research in education: exploring students' response to a focused feedback initiative. Sage Research Methods Cases, 19-24.
- Sellnow DD and Treinen KP (2004) The role of gender in perceived speaker competence: An analysis of student peer critiques. *Communication Education* 53(3): 286-296.
- Stefani LAJ (1994) Peer, self and tutor assessment - relative reliabilities. *Studies in Higher Education* 19(1): 69-75.
- Suñol JJ, Arbat G, Pujol J, et al. (2015) Peer and self-assessment applied to oral presentations from a multidisciplinary perspective. *Assessment & Evaluation in Higher Education* 41(4): 622-637.
- Taras M (2010) Student self-assessment: processes and consequences. *Teaching in Higher Education* 15(2): 199-209.
- Topping K. (2003) Self and peer assessment in school and university: Reliability, validity and utility. In: Segers M, Dochy F and Cascallar E (eds) *Optimising new modes of assessment: In search of qualities and standards*. Dordrecht, The Netherlands: Kluwer Academic, 55-87.
- Topping K (2009) Peer Assessment. *Theory into Practice* 48(1): 20-27.
- Verano-Tacoronte D, González-Betancor SM, Bolívar-Cruz A, et al. (2016) Valoración de la competencia de comunicación oral de estudiantes universitarios a través de una rúbrica fiable y válida. *Revista Brasileira de Educação* 21(64): 39-60.
- Ward M, Gruppen L and Regehr G (2002) Measuring self-assessment: Current state of the art. *Advances in Health Sciences Education* 7(1): 63-80.

Table 1. Number of assessments by type of rater

	Engineering	Labour Relations and Human Resources
Professors	126	184
Peers	3843	8280
Self-assessment	63	92

Table 2. Reliability of the rubric

		Cronbach's Alpha
Engineering	Professors	0.8578
	Peers	0.8003
	Self-assessment	0.7520
Labour Relations and Human Resources	Professors	0.8296
	Peers	0.8704
	Self-assessment	0.7965

Table 3. Descriptive statistics of the total score on the presentations by degree

		Engineering	Labour Relations and Human Resources
Min. – Max.	Professors	12 – 29	11 – 29
	Peers	17 – 27	18 – 26
	Self-assessment	18 – 30	18 – 30
Mean	Professors	21.29	21.49
	Peers	22.62	22.18
	Self-assessment	25.39*	24.29*
Standard Deviation	Professors	4.41	3.78
	Peers	2.14	1.83
	Self-assessment	2.94	3.10

Note: * = Mean difference by source of assessment statistically significant at 1%

Table 4. Descriptive statistics of the total score awarded to the presentations, by the degree programme and gender of the speaker

		Engineering		Labour Relations and Human Resources	
		Men	Women	Men	Women
Number of presentations		47	16	27	65
Min. – Max.	Professors	14 – 29	12 – 28	14 – 25	11 – 29
	Peers	17 – 27	17 – 26	19 – 25	18 – 26
	Self-assessment	18 – 30	20 – 27	18 – 30	18 – 30
Mean	Professors	21.60	20.38	19.46*	22.25*
	Peers	22.81	22.06	22.07	22.23
	Self-assessment	26.00*	23.86*	25.55*	23.84*
Standard Deviation	Professors	4.11	5.26	3.13	3.74
	Peers	2.06	2.35	1.66	1.91
	Self-assessment	3.02	2.14	2.61	3.16

Note: *= Mean difference by gender of the speaker statistically significant at 1%

Table 5. Linear correlation among sources of assessment by gender and degree programme

		Men			Women		
		Professors	Peers	Self-ass.	Professors	Peers	Self-ass.
Engineering	Professors	1			1		
	Peers	0.83	1		0.83*	1	
	Self-assessment	0.20	0.24	1	-0.09	0.32	1
Labour Relations and Human Resources	Professors	1			1		
	Peers	0.72	1		0.78*	1	
	Self-assessment	0.03	0.10		0.43*	0.47*	1

Note: * = Correlation coefficient statistically significant at 1%

Table 6. Rank correlation among sources of assessment by gender and degree programme

		Men			Women		
		Professors	Peers	Self-ass.	Professors	Peers	Self-ass.
Engineering	Professors	1			1		
	Peers	0.76*	1		0.80*	1	
	Self-assessment	0.16	0.17	1	-0.12	0.24	1
Labour Relations and Human Resources	Professors	1			1		
	Peers	0.84*	1		0.78*	1	
	Self-assessment	-0.01	0.15	1	0.40*	0.43*	1

Note: * = Correlation coefficient statistically significant at 1%

Table 7. Multiple linear regression

	Coefficient	
Constant	-1.96 *	(-1.97)
Degree (Ref. Labour Relations and Human Resources)		
Engineering	1.08	(1.37)
Gender (Ref. Men)		
Women	3.85 **	(5.06)
Oral presentation competence (Ref. High)		
Low	-9.15 **	(-9.06)
Medium	-3.24 **	(-4.21)
$R^2 = 0.52$		$F = 30,57$ ($p = 0.000$)

Note 1: Response variable = Score given by the professor – students' self-assessment

Note 2: ** = Significant at 1%; * = Significant at 10%

Note 3: t-values between brackets

Figure 1: Grades given to the presentations by each type of rater, ranked according to grades given by professors

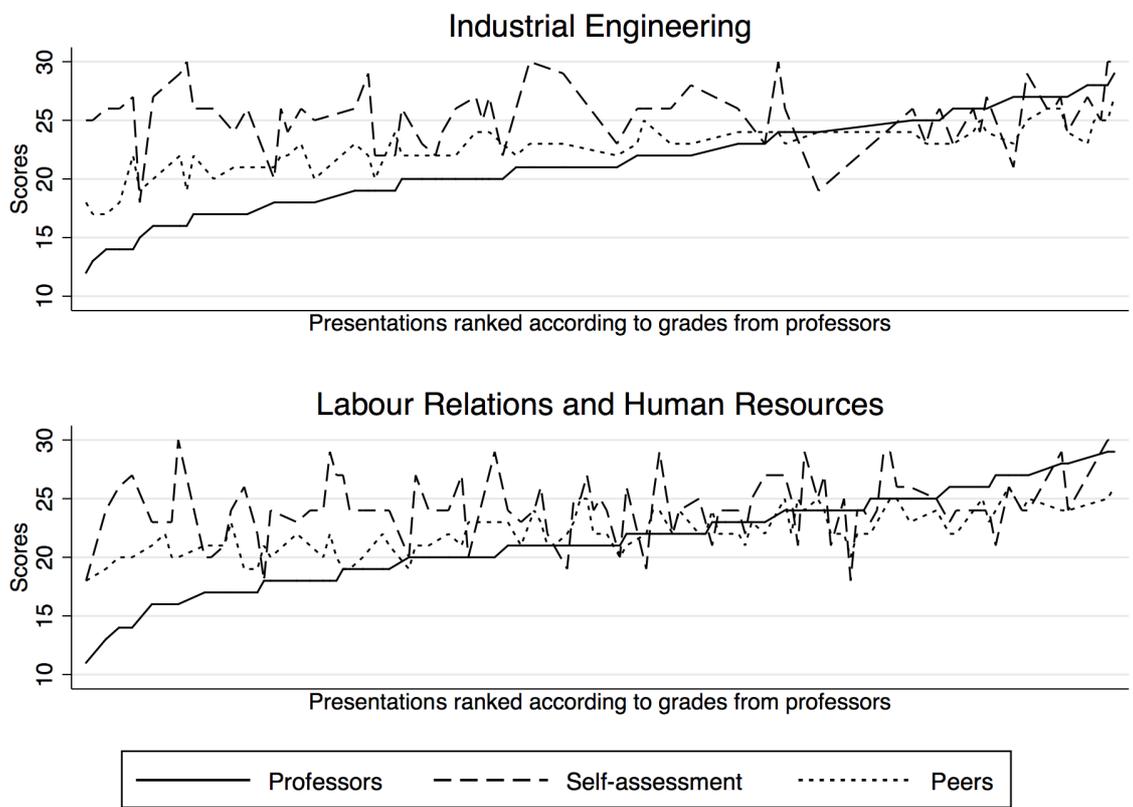


Figure 2: Marks given by the raters to the students rated best and worst by the professors

