

Peer Assessment Reliability in an Organic Chemistry Activity: Do Students Overrate Their Peers?

Lorena Atarés Huerta* and Juan Antonio Llorens Molina




Cite This: *J. Chem. Educ.* 2023, 100, 3200–3208



Read Online

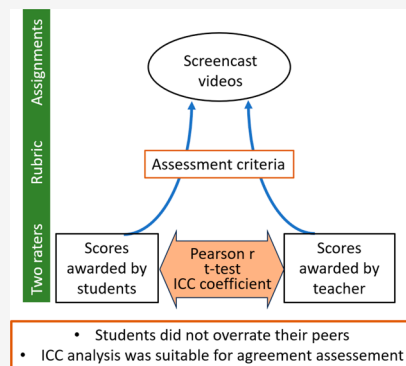
ACCESS |

 Metrics & More

 Article Recommendations

ABSTRACT: Despite the demonstrated learning benefits of peer evaluation, fears of teachers about its low reliability may restrict its use. In this study, the validity of peer assessment, in terms of agreement with the ratings of the teacher, has been tested in an organic chemistry course. The students were organized into small groups and commissioned to produce a screencast video on a molecule. Both students and teachers assessed the screencasts on five different dimensions. The internal consistency of the rating scale was confirmed. Comparing both data sets revealed fair correlations in all cases but statistically significant differences in four dimensions. The grades awarded by peers were lower than those granted by the teacher, which contradicts most of the results found in the literature. Statistically significant differences ($p < 0.05$) are compatible with the good agreement as reported by the Intraclass Correlation Coefficient and vice versa. Further research is necessary to elucidate the effect of diverse variables on the raters' agreement, improving the validity of peer evaluation.

KEYWORDS: *Chemical Education Research, First-Year Undergraduate, Organic Chemistry, Assessment, Molecular Properties, Peer Evaluation, Inter-rater Agreement*



Peer assessment (PA) is how students grade their classmates' activities using pre-agreed relevant criteria. A wide range of activities subject to PA is reported in the literature, often presented in the context of a team. Group activities have the ability to promote highly valued learning outcomes such as teamwork, leadership, communication, problem-solving, and critical thinking skills.¹ In particular, student-generated videos have proven to be an effective tool for learning. Before producing a screencast video, students have to research the information needed and generate the corresponding sequence and representations.² The interest in this type of assignment in the teaching and learning of organic chemistry is grounded in the visual nature of this discipline.

Feedback from different sources, such as mentors, lecturers, or peers, can significantly enhance the student learning process.³ PA has been repeatedly advocated in the literature as an effective pedagogical strategy for enhanced learning.^{1,3–5} PA increases students' engagement, promotes students' critical thinking, and increases students' motivation to learn.⁴ The positive formative effects of PA on student achievement and attitudes are as good as or better than the effects of teacher assessment.⁶ When students take the rater role, reviewing their peer's work allows them to reflect critically on their own understanding and performance.⁷ Therefore, PA is a reciprocal process in which the rater also benefits by expanding their own understanding of the matter.³ Bruffee⁸ points out that "conversation with people we regard as our peers—our equals, members of our own community—is almost always the most

productive kind of conversation. So students have to converse with their peers about writing both directly and indirectly". Research on the learning benefits of PA in chemistry courses is still scarce but very promising, as shown by previous studies^{9–13}

In addition to the learning benefits of PA, some practical advantages should also be pointed out.³ PA can greatly increase the efficiency of teachers' grading, both functioning as a formative pedagogical tool¹⁴ or a summative assessment tool.¹⁵ With increased student numbers and greater pressures on curriculum time, developments in PA can be an effective resource in the modern educational setting.³

On the other side, some practical restrictions have been pointed out to affect the successful implementation of this practice, such as constraints of time and classroom space³ or the fact that it might be difficult to manage for large classes. From a psychological perspective, students may consider that PA is challenging and socially inconvenient,^{16–18} while resenting some distrust in fellow students' abilities to peer-assess,¹⁹ especially if they have the attitude that they come to

Received: September 20, 2022

Revised: August 15, 2023

Published: August 30, 2023



Table 1. Recent Studies on the PA Validity, in Terms of the Agreement of Grades Awarded by Peers and Staff, that Have Been Reported in Higher Education Science and Technology Studies⁴

Reference	Context	Activity	Statistical Analyses	Main Result
De Meulemeester et al. ²⁸	Health science: medicine, dentistry and biomedical sciences (first year)	Research paper	Pearson's <i>r</i> , paired sample <i>t</i> tests	Strong correlations and overrating
Sánchez González ²⁹	Energy engineering (fourth year)	Workshop activity	None	Overrating
DuCoin et al. ³⁰	Medical education (third year)	Simulated procedures	Percentage agreement and kappa	Good agreement
Sealey ³¹	Clinical Exercise Physiology (fourth year) and Postgraduate Diploma in Clinical Exercise Physiology	Written assignment	Paired samples <i>t</i> test and correlations	Overrating
Arlianty and Febriana ³²	Physical chemistry (second semester)	Experiment	None	Overrating
Hassell and Lee ³³	Civil engineering and Computer science (first year)	Short speech	ANOVA and Spearman's correlation coefficient	Overrating
Davey and Palmer ³⁴	Chemical engineering (third year)	Problem solving	Dispersion plots	Good agreement
Gerczei ³⁵	Biochemistry (third and fourth years)	Mini-conference	Correlation coefficient	Similarity
Hamer et al. ³⁶	Software engineering programming	Project	Pearsons <i>r</i> , Paired samples <i>t</i> test	Good correlation
Verkade and Bryson-Richardson ³⁷	Genetics (final year)	Oral presentation	Pearsons <i>r</i> , Paired <i>t</i> tests	Strong correlation, overmarking
Atares Huerta et al. ¹³	Organic chemistry (first year)	Posters	Pearsons <i>r</i> , ICC	Overrating and underrating
English et al. ³⁸	Medicine (first year)	Paper	Bland–Altman plot	Underrating

⁴Overrating: the grades awarded by peers are higher than those awarded by the teacher. Underrating: the grades awarded by peers are lower than those awarded by the teacher.

university in order to receive feedback from experts. Overall, issues related to the reliability of PA seem to be the most restrictive against its use,^{4,5,18} hence depriving many students of its learning benefits. This article presents a literature overview on the validity of peer assessment in higher education and reports the findings of a project where the peer grades were compared to the teacher grades.

LITERATURE REVIEW

Despite the great potential of PA for meaningful long-term learning, both researchers and practitioners remain concerned about the students' ability to assign valid ratings to their peers' work.²⁰ Validity refers to the agreement between peer ratings and teacher ratings, assuming teacher ratings to be the gold standard.⁵ In the last 25 years, several reviews and meta-analyses have been published on PA,^{4–6,21–27} and concerning its validity, several of these works provide interesting conclusions based on aggregated data. In their meta-analysis, Van Zundert et al.²¹ explored the factors leading to satisfactory psychometric qualities in PA, such as a good correlation between the peers and the staff's marks, and found that, when expressed in terms of the agreement between PA and staff assessment, the psychometrics were generally satisfactory. Falchikov and Goldfinch⁵ reported a meta-analysis on forty-eight quantitative studies that included comparisons of numerical marks or grades awarded by peers and faculty and found a good mean correlation ($r = 0.69$). Li et al.⁴ and Sanchez et al.²⁷ reported similar aggregated results, and Sluijsman et al.²⁴ also found an overall good agreement between the two raters. Topping⁶ asserted that PA is of adequate reliability and validity in a wide variety of applications. In spite of the general optimistic conclusions normally reached from aggregated data, individual studies report inconsistent results. Table 1 compiles some papers published since 2012 where the ratings of teachers and students have been compared in a variety of science and

engineering higher education studies. Two different aspects should be commented on in this body of literature: the diversity of the conclusions reached on the validity of PA and the statistical methods used for comparison.

While some studies report overrating,^{17,33,39} many other studies report varying degrees of agreement.^{40,41,13} The correlation coefficients reported range from poor (0.21;³⁵ 0.29⁴²) to moderate (0.47;⁴¹ 0.60¹⁸) to high (0.98⁴³). Although individual cases of underrating are common, only one study was found where underrating is the general tendency observed.³⁸

The inconsistent outcome of individual studies is probably due to the great number of variables affecting the validity of PA. In their review, Li et al.⁴ list 17 variables, which were classified into two categories. On the one hand, those related to PA settings include the PA mode (paper- or computer-based), the subject area, and the task. On the other hand, those related to PA procedures include the constellation of assessors and assesseees, the number of peer raters per assignment, and the activity being compulsory or voluntary.⁴⁷ The quality of the peer rating improves when PA is supported by training, checklists, exemplification, teacher assistance, and monitoring,⁴ and when peer raters are familiar with explicit rating criteria.⁵ In this respect, providing students with an assessment tool that clearly establishes the assessment criteria is important to minimize disagreement. Previous research on rating instruments confirms that they increase peer assessment construct validity.⁴¹ The internal consistency of this instrument should constantly be tested to warrant consistent results from different parts of a measure.

When quantitative data are analyzed statistically, some studies merely present the mean values for bulk comparison. For instance, Davey and Palmer³⁴ present a dispersion plot and compare both grades by means of their ratio and difference. Correlation coefficients are frequently used to report raters' agreement.⁵ Even if it is a good measure of how two variables

correlate, it merely quantifies how close points lie to any straight line. In fact, two sets of ratings can strongly correlate while being significantly different, as observed previously.^{28,39} Measuring the agreement between two raters goes beyond checking for mere correlation and requires that the scores awarded by different raters are equal.⁵ The Intraclass Correlation Coefficient (ICC) has been used to test the inter-rater agreement in several studies.^{18,44,45} This coefficient can be interpreted as the total amount of variance in the score that can be attributed to the actual object. High agreement between raters would result in little variation of scores awarded by different raters to the same object and high ICC (close to 1) results.

■ RATIONALE FOR THIS STUDY

As shown in the previous sections, extensive research has been done on the validity of PA in Higher Education and its inconsistent outcome depending on a wide variety of factors. Previous research has stressed the necessity of conducting new studies to improve the quality and accuracy of PA.⁴

Since peer evaluation is the major tool to evaluate scientific research, students interested in science are well served by being introduced to this evaluation method.³⁵ In the specific field of physical sciences, peer assessment includes alternatives to the traditional tutor-marked methods for laboratory work, scientific group projects, and student poster presentations.³

In a previous study, we carried out a PA experiment based on poster presentations.¹³ In the present study, we decided to integrate screencasts into our practice, given the advantages of their use.² To the best of our knowledge, no studies on the validity of PA with this activity have been done. This work intends to make its contribution to the application of PA to this particular activity in our academic context: a first year undergraduate organic chemistry course.

■ RESEARCH QUESTIONS

The outcome of PA depends on numerous variables, some of which are constrained in a specific academic context. In our case, the most important nonmodifiable features were the area (organic chemistry) and the academic level (first year). Keeping in mind the potential benefits of PA for meaningful learning, we wondered how much validity would be within reach in a particular PA activity involving a specific rating scale and assignment. We chose to assign the production of screencasts based on both the lack of studies where it is used and its visual nature, which makes it especially interesting for the teaching and learning of chemistry. Therefore, in this study, we address the following research questions:

RQ1: Does the rating scale used in this activity have good internal consistency to measure the quality of the screencast that the students produce?

RQ2: Considering our academic context and the constraints of the variables affecting the outcome of PA, is peer assessment a valid practice in terms of the agreement between the peers' and the teacher's ratings?

■ METHODOLOGY

Context and Project Description

This research was conducted in two consecutive academic years (2019–2020 and 2020–2021) in an introductory course in organic chemistry for Food Science and Technology. The mean age of the students in the class participating in this study

was 19 (47.3% male, 52.7% female). At the beginning of the semester, the students were introduced to the assignment of producing a 5 min screencast video on a molecule of interest in organic chemistry for Food Science and Technology. Poliformat, the university's course management system, was used to provide the students with concise instructions, the rubric that would be used for assessment, and example screencasts for guidance. Within 2 weeks, the three-member groups were formed, and the molecules were assigned by the teacher. One month later, the screencasts were submitted, and the assessment process began.

Assessment of the Screencasts

A total of 64 screencasts were submitted (29 in 2019–2020 and 35 in 2020–2021) and independently evaluated by both the instructor and the students over a 2 week period. The students were provided with the same rubric (a matrix describing scaled levels of achievement for a set of dimensions) that the teacher used, which is shown in Table 2. The five dimensions considered were

- (1) compliance with rules
- (2) technical assessment and formal aspects
- (3) oral and written expression
- (4) bibliographic references and sources of information
- (5) conceptual correctness and scientific vocabulary

Based on this rubric, both raters scored each screencast on each criterion on a 0 to 10 scale. This numeric scale was selected because it is normally used by the staff; hence, students can naturally assess their own performance on this numeric scale. Each screencast was graded by at least five groups of students (screencast submitted by group 1 was graded by groups 2 to 6, and so on), and their qualifications were averaged to get the mean student rate per item (MSR). These would be compared with the teacher ratings (TR), which were awarded by the teacher of the course. The granularity of the scores was 1.

Statistical Analysis

In order to answer RQ1, Cronbach's alpha was calculated. This parameter quantifies the internal consistency of a rating scale in terms of the agreement among the answers to the different items.

In accordance with the approach typically taken in the literature, we calculated the Pearson product-moment correlation between the TR and the MSR. Paired-sample *t* tests were conducted to look for differences between the grades awarded by both raters. When significant differences could not be confirmed at the 95% confidence level ($p > 0.05$), equivalence tests were run to confirm the equivalence. These three tests were carried out using a Statgraphics Centurion XVI (Manugistics Corp., Rockville, MD).

The fact that each group of students evaluates only a small subset of all the screencasts produced may cause a reviewer–screencast interaction, rendering correlation coefficients less trustworthy.¹⁸ The solution to this problem is to use the Intraclass Correlation Coefficients (ICC), which is a common measure of the reliability of either different judges or different items on a scale.⁴⁶ Essentially, the ICC increases as the mean square of the effect of the assignment increases, and it goes down as the mean square of the interaction of assignments with reviewers increases. This calculation was conducted using the Statistical Package for Social Sciences (SPSS) version 20.0.

Table 2. Rubric Used by Both the Students and the Teacher to Grade the Screencast Videos

Dimensions	0.0–2.5	2.5–5.0	5.0–7.5	7.5–10.0
compliance with rules	Some of the parts of the work have not been addressed.	Some part of the work is treated superficially or poorly.	All parts of the work have been adequately addressed, with some minor shortcomings.	All parts of the work have been treated in a balanced and complete manner.
technical assessment and formal aspects	The video and audio recording is clearly deficient, with excessive text on the slides and poor quality images.	Slides have the necessary text and images are adequate, but audio or video recording is deficient.	Slides (text and images) are correct, although not visually appealing. Audio and video are correct.	The presentation is correct and visually appealing. The audio contributes to capture attention.
oral and written expression	The written text presents important deficiencies in syntax and spelling. The oral text consists of the reading of the text of the slides.	The written text is correct, although it presents some syntax and spelling mistakes. The slides have too much text that is read as is.	The written text presents deficiencies but is schematic, so that the oral text explains and develops it and is not a simple reading.	The text does not present syntax or spelling errors. The oral text, although based on the slides, is elaborated with autonomy.
bibliographic references and sources of information	Only 1 or 2 general and informative references are cited, without applying any type of regulation. The references of the images are not cited.	Several general and informative references are cited, although they are not complete or do not comply with the regulations.	All content is supported by bibliographic references, complying with the regulations. They include some university manuals or Web sites of university and scientific institutions.	All content is supported by bibliographic references, complying with the regulations. Specialized texts or Web sites of university or scientific institutions predominate.
conceptual correctness and scientific vocabulary	There are important conceptual errors and little or erroneous use of scientific vocabulary.	Some conceptual errors or confusions. Sometimes inadequate vocabulary taking as a reference the contents of the subject.	There are no significant errors. In the vocabulary, every day or imprecise expressions are used instead of the subject vocabulary.	There are no conceptual errors or confusion and the vocabulary used corresponds to the scientific terminology introduced in the subject.

FINDINGS

General Development of the Activity

More than 90% of the enrolled students participated in this voluntary activity. While they were producing the screencasts, the teacher's ordinary weekly office hours were sufficient to provide the necessary guidance. At this stage of the activity, it did not involve a significant time requirement from the teacher. During the assessment period, no difficulties were reported, which probably relates to the suitability of the scaffolded peer review process.

Evidence of Validity

While developing the rating scale, the authors focused on creating an instrument that would show strong evidence of consistent scores among raters. First, we consulted bibliography on prior rubrics used to compare scores provided by different raters, such as those shown in Table 1. Although no studies on screencast assessment were found, an initial version of the instrument was produced, which we tested in a previous study where PA was compared with teacher assessment on a poster assignment (Atares Huerta et al.¹³). Both for that study and for the present study, we had the instrument reviewed by colleagues to verify that it represented the actual quality traits being included. In the present case, we outlined a set of criteria to assist the raters, both teachers and students, in recognizing mastery attributes in the screencasts. Lastly, in the present study, we report that the raters were in fact able to use the rating scale consistently.

Internal Consistency of the Rating Scale

Despite the fact that each of the items of the rubric constitutes by itself a quality factor of the screencasts produced, we wondered whether a general indicator of *screencast quality* could be the outcome of our rubric. Under this hypothesis, we used Cronbach's alpha index as a useful estimate of single-administration reliability conceptualized as item consistency,⁴⁸ that is "the proportion of the test variance due to all common factors among the items".⁴⁹ Although an alpha threshold value of 0.7 is often reported in the literature for acceptable consistency, there is no standard criterion and the value of alpha should be interpreted in the specific context.⁴⁸ In the present study, Cronbach's alpha values were 0.7825 and 0.8533 for the ratings awarded by the instructor and the students, respectively. These high values of the parameter confirm the internal consistency of the rating scale aiming to measure the quality of the screencast videos produced. Moreover, they demonstrate that the ratings awarded per item can be used to calculate an average score for the general quality of the screencast, which will be onward referred to as *global*.

Peer Evaluation vs Instructor Evaluation

Figure 1 plots the individual grades awarded by both raters, which reflect the high quality of the screencasts. Most scores ranged between 7 and 10. Table 3 shows the incidences of overrating and underrating. Out of the 320 ratings assigned (64 screencasts times five items each), the students overrated their peers on 92 occasions, while on 215 occasions, the opposite was true. This resulted in 45 and 18 underrated and overrated global grades, respectively. Underrating was noticeably relevant for item 3 (oral and written expression), whereas item 5 (conceptual correctness and scientific vocabulary) showed the most balanced trend.

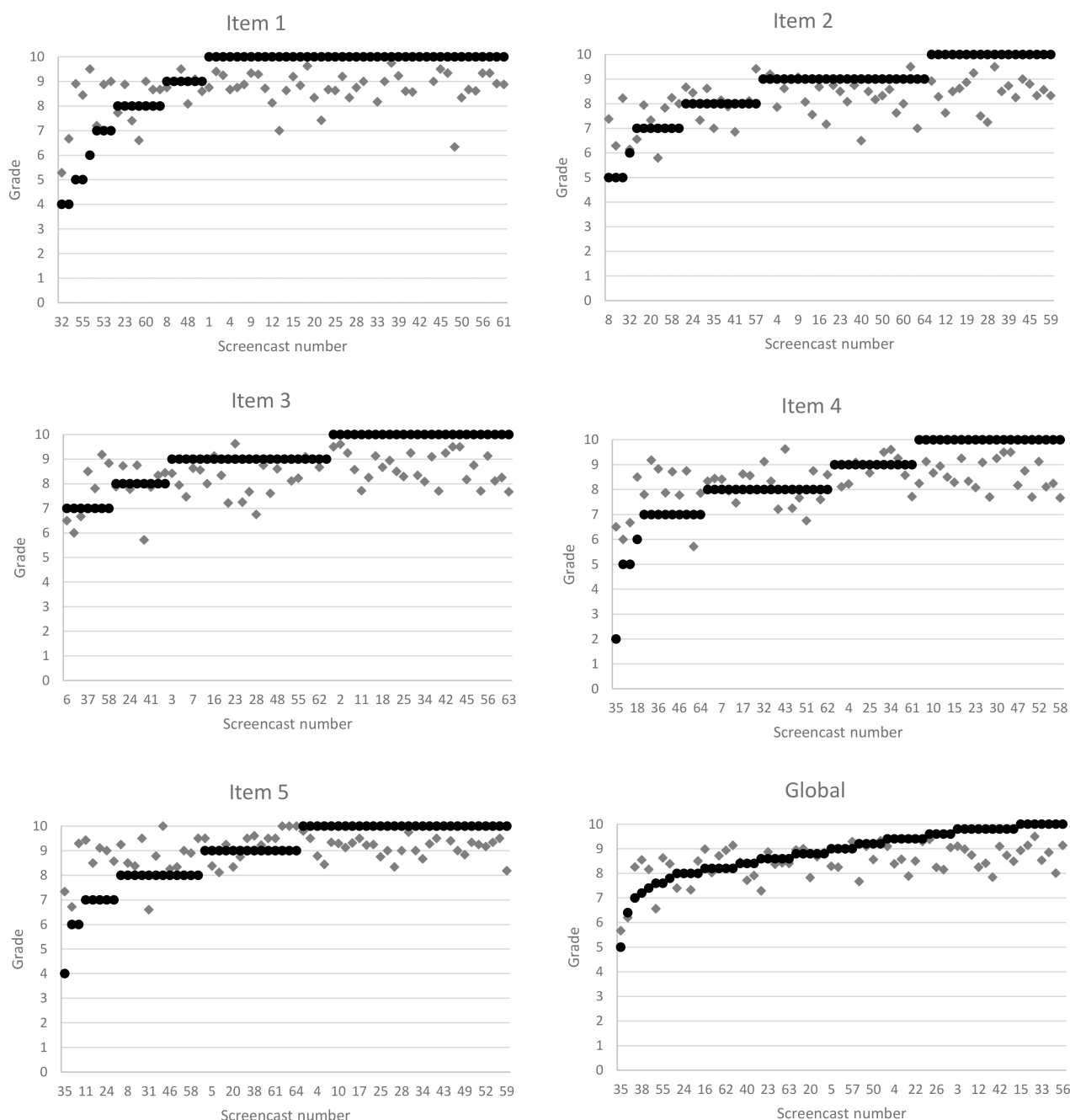


Figure 1. Teacher grades (black) and mean student grades (gray) on each item and globally, on the 64 screencasts, labeled on the *x* axis. Item 1: compliance with rules; item 2: technical assessment, formal aspects; item 3: oral and written expression; item 4: bibliographic references and sources of information; item 5: conceptual correctness and scientific vocabulary.

Table 3. Incidence of Underrating and Overrating on Each Item and Globally

Dimension	Underrating	Overrating
Global	45	18
(1) compliance with rules	46	14
(2) technical assessment, formal aspects	42	18
(3) oral and written expression	50	11
(4) bibliographic references and sources of information	43	20
(5) conceptual correctness and scientific vocabulary	34	29

Correlation Tests

Prior to the statistical analyses, the normality of the data was examined. Table 4 shows the standard skewness and kurtosis of the data sets, which were within the +3 and −3 criteria in most cases.⁴¹ The normality of the data supports the quality of subsequent statistical analyses. The two variables were subject to a Pearson correlation analysis, and the resulting *r* coefficients of the individual items and the global grade are shown in Table 4, along with their corresponding linear equations related to the two variables. All correlations were significantly positively linear and statistically significant at the 99.0% confidence level.

Table 4. Standard Skewness and Kurtosis, Linear Equations, Pearson Correlation Coefficients, Regression Model, and Determination Coefficients^a

Dimension	Standard Skewness	Standard Kurtosis	Linear Equations	<i>r</i>	Regression Model	<i>R</i> ²
Global	0.78	−0.99	MSR = 0.4516TR + 4.4360	0.6078 ^b	MSR = 0.947TR	0.992 ^b
1	3.81	3.50	MSR = 0.2483TR + 6.3881	0.4293 ^b	MSR = 0.928TR	0.976 ^b
2	2.15	1.68	MSR = 0.3189TR + 5.3740	0.5012 ^b	MSR = 0.926TR	0.983 ^b
3	1.88	0.68	MSR = 0.3425TR + 5.2214	0.4060 ^b	MSR = 0.913TR	0.987 ^b
4	1.11	−0.95	MSR = 0.4848TR + 3.9292	0.5832 ^b	MSR = 0.934TR	0.977 ^b
5	2.55	0.26	MSR = 0.2195TR + 7.0569	0.4136 ^b	MSR = 0.992TR	0.983 ^b

^aTR: teacher ratings; MSR: mean students ratings. ^b*p* < 0.01.

The results of *r* ranged between 0.41 and 0.61, similar to previous studies.^{41,18} The best correlation was observed for item 4 (bibliographic references and sources of information), probably because the acceptability of information sources is a relatively objective item to assess. The second-best correlation coefficient was that of the global grade, in agreement with the observations of Falchikov and Goldfinch,⁵ who stated that global judgments with clearly stated criteria showed better agreement than judgments on separate dimensions. The worst correlation was found for item 3 (oral and written expression), where underrating was most prevalent.

Once moderate correlations had been found, it was hypothesized that the score given by the students could be predicted from the score given by the instructor. Linear regressions with no intercept were performed with TR as the independent variable to predict MSR. The resulting equations and determination coefficients (*R*²) are shown in Table 4. Over 97% of the variability of MSR is explained by the TR variable. Figure 2 is a scatterplot of global MSR vs global TR, including the correlation line and the regression model.

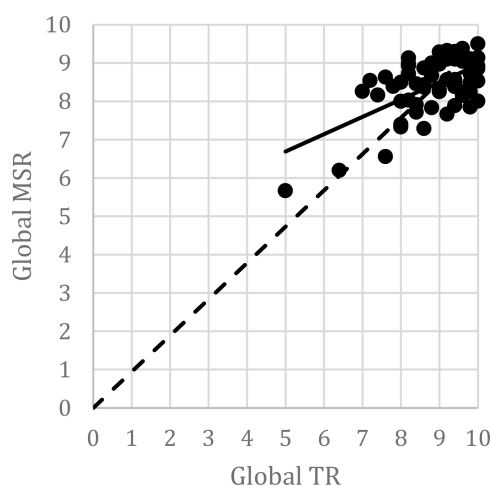


Figure 2. Scatterplot of global MSR vs global TR (*n* = 64). Linear correlation (solid line) and regression model (dashed line)

The results of correlation and regression analyses demonstrated the fairly linear shape of the dispersion plots, even though the data were clustered between 7 and 10. Despite the linear trends, the agreement between both raters should not be blindly trusted on the value of a correlation coefficient. Further analyses were carried out to study the differences and agreement between the raters.

Dependent Samples *t* Test

As commented above, good correlation results are not incompatible with significant differences between the grades awarded by the two raters. Table 5 shows the mean and standard deviation of TR and MSR on each item and globally, along with the results of the dependent sample *t* test.

Except for item 5 (conceptual correctness and scientific vocabulary), the *p*-values were lower than the usually chosen cutoff of 0.05, which demonstrates that, with that exception, the students were harsher than the expert assessor. This is coherent with the high incidence of underrating compared to overrating and with the average scores provided by the two raters. To the best of our knowledge, these results are incoherent with most of the previous research on PA, since underrating has very rarely been the main conclusion of previous studies. The most differing averages of TR and MSR, and hence the most remarkable underrating, were found in item 3 (oral and written expression). This is probably because students may be unable to recognize correct statements, both oral and written, from their peers. On the contrary, by having greater understanding and confidence in the subject, the teacher is more able to recognize the student's expressions and understanding.

In the case of item 5, no significant differences between TR and MSR were detected, and the null hypothesis could not be rejected at the 95% confidence level (*p* > 0.05). An equivalence test (the two one-sided tests, TOST procedure) was run, where upper and lower equivalence bounds have to be specified based on the smallest effect of interest. The boundaries set in this case were −0.5 and 0.5, and the equivalence between both raters was demonstrated at the 95% confidence level. The equivalence of both sets of ratings is probably due to the nature of item 5, which deals with

Table 5. Average Values and Standard Deviations (in Parentheses) of TR, MSR, *t* statistics, *p*-Values, and Effect Size on Each Item and Globally

Dimension	Item 1	Item 2	Item 3	Item 4	Item 5	Global
TR	9.1 (1.6)	8.7 (1.3)	9.1 (1.0)	8.5 (1.5)	9.0 (1.3)	8.9 (1.0)
MSR	8.7 (0.9)	8.1 (0.8)	8.3 (0.8)	8.0 (1.3)	9.0 (0.7)	8.4 (0.7)
<i>t</i> statistics	−2.70	−3.65	−5.80	−2.69	0.49	−4.22
<i>p</i>	0.0088	0.0005	0.0000	0.0090	0.6291	0.0001
Effect size	−0.38	−0.51	−0.82	−0.38	0.07	−0.60

conceptual correctness and scientific vocabulary. These aspects are strongly related to basic concepts, which are very clearly defined. More specifically, this item relates to the correct denomination of the molecule's functional groups. Hence, a screencast video in which this is done would be awarded a good grade by both raters.

Agreement Tests (ICC)

Considering the different models available to calculate ICC in SPSS, we used those that fit to our experimental design and the absolute agreement type. In this study, the ICC represents the total amount of variance in the scores, which is attributable to the actual quality of each screencast. Hence, the highest value of ICC (1) would be obtained if the two raters fully agreed on the score awarded to every screencast, and ICC would decrease to 0 as the agreement between the raters diminishes. Table 6 shows the results of ICC.

Table 6. Intraclass Correlation Coefficient (ICC) for Average Measures of Peer Grading and Instructor Grading Scores ($n = 64$)

Dimension	ICC	p
Global	0.687 (good)	0.000
Item 1	0.518 (sufficient)	0.001
Item 2	0.585 (sufficient)	0.000
Item 3	0.470 (sufficient)	0.000
Item 4	0.709 (good)	0.000
Item 5	0.515 (sufficient)	0.003

The results showed moderate strength, and the agreement reached between the two raters depended on the item considered. The best agreement was reached for item 4 (bibliographic references and sources of information), which was even higher than the agreement on the global grade. Probably, this stems from the same cause that led to the high r coefficient for item 4, that is, the objectivity of the criterion, as stated above.

On the contrary, the worst agreement was reached for item 3 (oral and written expression), coherently with the high incidence of underrating observed. As stated above, this may relate to the student's inability to perceive the correct statement. This is coherent with that found in a previous study on poster presentations, despite the different nature of the communication tools used in both activities.¹³

The results of ICC reveal how this analysis may involve a step further in inter-rater agreement studies. On the one hand, no significant differences between means were found for TR and MSR on item 5, even though the agreement shown by the ICC value is only sufficient (0.515). On the contrary, the good agreement found for item 4 (ICC = 0.709) is not incompatible with significant differences between raters at the 95% confidence level.

CONCLUSIONS

This study sheds light on best practices for designing peer assessment activities in a first-year undergraduate course. To study the validity of peer assessment, a screencast assignment was put forward and the agreement between the grades awarded by the teacher and the students, while using the same grading instrument, was quantified. In response to Research Question 1, the internal consistency of the rating scale used to measure the quality of the screencast videos was verified. In

addition to the actual experiment, we intend to make a contribution concerning the use of correlation coefficients to study the inter-rater agreement. A correlation coefficient only describes how two variables correlate, whereas the ICC looks for actual agreement between the grades awarded by different raters.

The findings on the inter-rater agreement were dependent on the dimension assessed. Where a statistically significant difference was found ($p < 0.05$), students did not overrate but repeatedly underrated their classmates. Overrating was not observed in any of the dimensions tested. This difference in marking is likely the result of a different perspective between the lecturer and students, and the possible sources of these differences have been discussed. Coherently with our results, and going back to RQ2, peer assessment is a promising practice that does not necessarily imply overrating between peers. Coherently with this, we can state that concerns about the reliability of PA are probably not a good reason for preventing instructors from implementing this approach, as long as appropriate scaffolds for peer review are implemented.

LIMITATIONS

The present work aimed to study the validity of peer assessment based on the agreement of the scores given by the two raters. It did not look into the consistency across peer raters or the effect of peer assessment on students' learning. Both of these issues need to be addressed in future research. We also focused on the product of the collaboration of the group rather than the actual collaboration, which could also be the focus of future studies to widen the knowledge of PA benefits. Further research into student engagement with peer feedback would also be valuable.

We found good internal consistency of the scores provided to the items in the rating scale, and hence, the average of the scores was used as a measure of the *global quality* of the screencasts. Further research should focus on the specific items of the rubric, as well as on additional items that would also have an impact on screencast quality. The fact that reliability has been tested by a single-administration coefficient also constitutes a limitation of this work.

It is worth mentioning that the high quality of the screencasts, resulting in ratings mostly varying between 7 and 10, resulted in data clustering and probably played against the accuracy of the correlations. Had the study resulted in a wider dispersion of the grades, which implies having some low grades, a better accuracy of the correlations would probably have been attained. Despite this drawback, we consider this clustering as a positive outcome of the activity, which reflects the engagement of the students in the production of good-quality screencast videos.

The different ways in which students and faculty interpret and apply the evaluation criteria may cause disagreement between the raters. The reasons behind the underrating observed should be further explored. It might be worthy of inquiry in a follow-up study to interview the participants for depth analysis to find detailed reasons for underrating. In any case, given the potential value of PA, future work is required to understand how peer marking can be used more effectively.

IMPLICATIONS

Some teaching implications related to the activity carried out in this study should be pointed out, which affects the practical

relevance of this study for instructors. Starting the activity reported here required relevant organizational effort from the teacher. Diverse aspects demanded time and attention, such as the rating instrument, the need to make regular announcements to the students, the organization of the groups, the video submission, the rating organization, and finally the assessment itself. In some cases, since this study was performed in large first-year pregraduate courses, its organization was more necessary and time-consuming.

AUTHOR INFORMATION

Corresponding Author

Lorena Atarés Huerta – *Universitat Politècnica de València, 46022 Valencia, Spain*; orcid.org/0000-0002-7172-3666; Email: loathue@tal.upv.es

Author

Juan Antonio Llorens Molina – *Universitat Politècnica de València, 46022 Valencia, Spain*

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acs.jchemed.2c00945>

Notes

The authors declare no competing financial interest.

REFERENCES

- (1) Wenzel, T. J. (2007). Evaluation tools to guide students' peer-assessment and self-assessment in group activities for the lab and classroom. *Journal of chemical education*. **2007**, *84* (1), 182–186.
- (2) Gallardo-Williams, M.; Morsch, L. A.; Paye, C.; Seery, M. K. Student-generated video in chemistry education. *Chemistry Education Research and Practice* **2020**, *21* (2), 488–495.
- (3) Chin, P. Peer assessment. *New Directions in the Teaching of Physical Sciences* **2007**, *3*, 13–18.
- (4) Li, H.; Xiong, Y.; Zang, X. L.; Kornhaber, M.; Lyu, Y.; Chung, K. S.; Suen, H. K. Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education* **2016**, *41* (2), 245–264.
- (5) Falchikov, N.; Goldfinch, J. Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of educational research* **2000**, *70* (3), 287–322.
- (6) Topping, K. Peer assessment between students in colleges and universities. *Review of educational Research* **1998**, *68* (3), 249–276.
- (7) Roberts, T. S. *Self, peer, and group assessment in e-learning*; Information Science Pub: Hershey, PA, 2006; pp 1–16.
- (8) Bruffee, K. A. *Collaborative Learning: Higher Education, Interdependence, and the Authority of Knowledge*; Johns Hopkins Press: Baltimore, MD, 1993.
- (9) Glaser, R. E.; Carson, K. M. Chemistry is in the News: Taxonomy of authentic news media-based learning activities. *International Journal of Science Education* **2005**, *27* (9), 1083–1098.
- (10) Glaser, R. E.; Poole, M. J. Organic chemistry online: Building collaborative learning communities through electronic communication tools. *J. Chem. Educ.* **1999**, *76* (5), 699–703.
- (11) Shibley, I. A., Jr; Milakofsky, L. K.; Nicotera, C. L. Incorporating a substantial writing assignment into organic chemistry: library research, peer review, and assessment. *J. Chem. Educ.* **2001**, *78* (1), 50–53.
- (12) Wimpfheimer, T. Peer-evaluated poster sessions: An alternative method to grading general chemistry laboratory work. *J. Chem. Educ.* **2004**, *81* (12), 1775–1776.
- (13) Atarés Huerta, L.; Llorens Molina, J. A.; Marin García, J. A. La evaluación por pares en Educación Superior. *Educación química* **2021**, *3* (1), 112–121.
- (14) Topping, K. J. Peer Assessment. *Theory into Practice* **2009**, *48* (1), 20–27.
- (15) Tsai, P. V. Peering into Peer Assessment: An Investigation of the Reliability of Peer Assessment in MOOCs. Master thesis, Princeton University, Princeton, NJ, 2013.
- (16) Raes, A.; Vanderhoven, E.; Schellens, T. Increasing anonymity in peer assessment by using classroom response technology within face-to-face higher education. *Studies in Higher Education* **2015**, *40* (1), 178–193.
- (17) Kilic, D. An Examination of Using Self-, Peer-, and Teacher-Assessment in Higher Education: A Case Study in Teacher Education. *Higher Education Studies* **2016**, *6* (1), 136–144.
- (18) Cho, K.; Schunn, C. D.; Wilson, R. W. Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology* **2006**, *98* (4), 891–901.
- (19) Planas Lladó, A.; Soley, L. F.; Fraguell Sansbelló, R. M.; Pujolras, G. A.; Planella, J. P.; Roura-Pascual, N.; Moreno, L. M. Student perceptions of peer assessment: an interdisciplinary study. *Assessment & Evaluation in Higher Education* **2014**, *39* (5), 592–610.
- (20) Liu, N. F.; Carless, D. Peer Feedback: The Learning Element of Peer Assessment. *Teaching in Higher Education* **2006**, *11* (3), 279–290.
- (21) Van Zundert, M.; Sluijsmans, D.; Van Merriënboer, J. Effective peer assessment processes: Research findings and future directions. *Learning and instruction* **2010**, *20* (4), 270–279.
- (22) Adiyani, P. N. EFL student's attitude towards writing peer assessment: a systematic review. *RETAIN (Research on English Language Teaching in Indonesia) (e-Journal)* **2021**, *9* (3), 1–10.
- (23) Topping, K. J. Digital peer assessment in school teacher education and development: a systematic review. *Research Papers in Education* **2023**, *38*, 472–498.
- (24) Dochy, F.; Segers, M.; Sluijsmans, D. The use of self-, peer- and co-assessment in higher education. *Studies in Higher Education* **1999**, *24* (3), 331–350.
- (25) Dochy, F. J. R. C.; Segers, M.; Sluijsmans, D. The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher education* **1999**, *24* (3), 331–350.
- (26) Double, K. S.; McGrane, J. A.; Hopfenbeck, T. N. The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review* **2020**, *32* (2), 481–509.
- (27) Sanchez, C. E.; Atkinson, K. M.; Koenka, A. C.; Moshontz, H.; Cooper, H. Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology* **2017**, *109* (8), 1049–1066.
- (28) De Meulemeester, A.; Peleman, R.; Pauwels, N. S.; Buysse, H. Peer Assessment, Self-assessment and Teacher Scoring within an Information Literacy Course. In *The Seventh European Conference on Information Literacy (ECIL)*; p 42.
- (29) Sánchez González, A. Peer-review to promote learning and collaboration between students of "Energy in Buildings. *Advances in Building Education* **2020**, *4* (1), 9.
- (30) DuCoin, C.; Zuercher, H.; McChesney, S. L.; Korndorffer, J. R., Jr Peer Assessment in Medical Student Education: A Study of Feasibility, Benefit, and Worth. *American Surgeon* **2022**, *88*, 2361–2367.
- (31) Sealey, R. M. Peer assessing in higher education: perspectives of students and staff. *Education Research and Perspectives* **2013**, *40*, 276–298.
- (32) Arlianty, W.; Febriana, B. Understanding peer and teacher assessment about laboratory skill on formative assessment through scientific approach. *Proceeding of the 3rd International Conference on Education* **2017**, *3*, 106–113.
- (33) Hassell, D.; Lee, K. Y. Evaluation of multi-peer and self-assessment in higher education: A Brunei case study. *International Journal of Innovative Teaching and Learning in Higher Education (IJITLHE)* **2020**, *1* (1), 37–53.

(34) Davey, K. R.; Palmer, E. Student Peer Assessment: A research study in a level III core course of the bachelor chemical engineering program. *Education for Chemical Engineers* **2012**, *7* (3), e85–e104.

(35) Gercezi, T. Impact of an in-class biochemistry mini-conference on students' perception of science. *J. Chem. Educ.* **2016**, *93* (9), 1521–1527.

(36) Hamer, J.; Purchase, H.; Luxton-Reilly, A.; Denny, P. A comparison of peer and tutor feedback. *Assessment & Evaluation in Higher Education* **2015**, *40* (1), 151–164.

(37) Verkade, H.; Bryson-Richardson, R. J. Student acceptance and application of peer assessment in a final year genetics undergraduate oral presentation. *Journal of Peer Learning* **2013**, *6* (1), 1–18.

(38) English, R.; Brookes, S. T.; Avery, K.; Blazeby, J. M.; Ben-Shlomo, Y. The effectiveness and reliability of peer-marking in first-year medical students. *Medical Education* **2006**, *40* (10), 965–972.

(39) Han, Y.; James, D. H.; McLain, R. M. Relationships between student peer and faculty evaluations of clinical performance: a pilot study. *Journal of Nursing Education and Practice* **2013**, *3* (8), 170–178.

(40) Bloxham, S.; West, A. Understanding the rules of the game: marking peer assessment as a medium for developing students' conceptions of assessment. *Assessment & Evaluation in Higher Education* **2004**, *29* (6), 721–733.

(41) Panadero, E.; Romero, M.; Strijbos, J. W. The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation* **2013**, *39* (4), 195–203.

(42) Kovach, R. A.; Resch, D. S.; Verhulst, S. J. Peer Assessment of Professionalism: A Five-year Experience in Medical Clerkship. *Journal of General Internal Medicine* **2009**, *24* (6), 742–746.

(43) Harris, J. Peer Assessment in Large Undergraduate Classes: An Evaluation of a Procedure for Marking Laboratory Reports and a Review of Related Practices. *Advances in Physiology Education* **2011**, *35* (2), 178–187.

(44) Luo, H.; Robinson, A.; Park, J. Y. Peer grading in a MOOC: Reliability, validity, and perceived effects. *Online Learning Journal* **2014**, *18* (2), 1–14.

(45) Li, H.; Zhao, C.; Long, T.; Huang, Y.; Shu, F. Exploring the reliability and its influencing factors of peer assessment in massive open online courses. *British Journal of Educational Technology* **2021**, *52*, 2263–2277.

(46) Shrout, P. E.; Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* **1979**, *86*, 420–428.

(47) Piccinno, T. F.; Basso, A.; Bracco, F. Results of a Peer Review Activity Carried out Alternatively on a Compulsory or Voluntary Basis. *J. Chem. Educ.* **2023**, *100*, 489–495.

(48) Barbera, J.; Naibert, N.; Komperda, R.; Pentecost, T. C. Clarity on Cronbach's alpha use. *J. Chem. Educ.* **2021**, *98* (2), 257–258.

(49) Cronbach, L. J. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* **1951**, *16*, 297–334.