Are Men and Women Really Different? The Effects of Gender and Training on Peer Scoring and Perceptions of Peer Assessment

José Carlos G. Ocampo[1], Ernesto Panadero[1,2], and Fernando Díez[1]

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

[2]Ikerbasque, Basque Foundation for Science, Bilbao, Spain

Correspondence concerning this article should be addressed to José Carlos G. Ocampo, ERLA Research Group, Universidad de Deusto, Bilbao, Spain. Email: jc.ocampo@deusto.es.

Recommended Citation

Ocampo, J. C. G., Panadero, E., Diez, F. (2022). Are men and women really different? The effects of gender and training on peer scoring and perceptions of peer assessment. *Assessment and Evaluation in Higher Education*, 1-18. https://doi.org/10.1080/02602938.2022.2130167

Funding: The first author of this project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska Curie grant agreement Nº 847624. In addition, a number of institutions backed and co-financed his project. Second and third authors funded by Spanish National R+D call from the Ministerio de Ciencia, Innovación y Universidades (Generación del conocimiento 2019), Reference number: PID2019-108982GB-I00

Acknowledgements: The authors would like to express their utmost gratitude to Dr. Ian Llenares, Dr. Rita Aringo, Ms. Nora Agpasa, and Mr. Raj Pallon for their assistance in data gathering. Thanks to Dr. Iván Sánchez for the advice in data analysis, and David Wind for the Eduflow access.

The authors declare to not having any conflict of interest regarding this manuscript.

Abstract

A number of studies have expressed that gender might be a source of difference and bias in peer assessment activities. However, evidence supporting this remains mixed and scant. The present study examined gender difference and accuracy bias between men and women assessors' peer scoring of same-sex or opposite-sex writing samples using a quasi-experimental approach in which we implemented peer assessment training to explore if it could minimise gender difference and bias. Additionally, we also explored the effects on participants' perceptions of trust and comfort in giving peer scores. A total of 145 (n_{Men} = 25, n_{Women} = 120) psychology students enrolled in four separate courses participated in this study. Two of the classes received peer assessment training, while the other two only received task instructions. Participants were divided into eight scoring subgroups where they peer scored three writing samples of varying quality (i.e., poor, average, and excellent) using a scoring rubric in *Eduflow*. We found that, regardless of their training condition, men and women assessors did not differ in their peer scores of men and women peers. Finally, only untrained men assessors showed less trust in their abilities and discomfort when peer scoring women assessees' writing samples.

Keywords: peer scoring, gender difference, training, interpersonal variables

Are Men and Women Really Different? The Effects of Gender and Training on Peer Scoring and Perceptions of Peer Assessment

Peer assessment is a collaborative learning activity where students "consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status" (Topping, 1998, p. 250). Several studies have documented the positive effects of peer assessment for students such as developing higher order thinking and deeper learning approaches (Cheng & Hou, 2015; Kim & Ryu, 2013; Panadero et al., 2016). A number of meta-analysis and reviews has also documented that peer assessment translates to academic performance (Double et al., 2020; Li et al., 2020; van Zundert et al., 2010). Also, since students interact with their peers to relay their assessments, they develop effective collaborative and communicative skills, an understanding of quality, and reflective habits which are all useful competencies even after they finish their studies (Kearney, 2013; Reinholz, 2016; Suñol et al., 2016; To & Panadero, 2019). These findings suggest that peer assessment has the potential to develop many facets of students' skills, which can transcend in different fields, if implemented consciously.

Since peer assessment is interpersonal in nature, students generate thoughts, emotions and actions as a result of their interaction with their peers and their work (Panadero, 2016). Some studies mention that students may also have different interpersonal reactions during peer assessment due to their gender (Rotsaert et al., 2017; Wen & Tsai, 2006, 2008; Zou et al., 2018). Since they vary interpersonally, it has been noted that gender has the potential to affect how they perform in peer assessment activities—including peer scoring (Falchikov, 2005). For example, some studies forward that gender is a source of bias for peer scoring (Espey, 2021; Langan et al., 2005, 2008), while other studies suggest that such a difference does not exist (Falchikov & Magin, 1997; Tucker, 2014). In spite of these findings, conclusive evidence that support or dispute gender difference in peer assessment is still

mixed and limited. More specifically, literature that examine the peer scores students give to same-sex or opposite-sex peers in the context of academic writing courses is limited. Thus, it is imperative to explore how gender plays a part in the peer assessment process. Therefore, the overall aim of this study is to compare men and women assessors in terms of their peer scores to same-sex or opposite-sex peer assessees' writing samples. We also examine how peer assessment training can affect peer scores, as well as perceptions of trust in one's self as an assessor and their perceptions of comfort after peer assessment were investigated.

Gender and Peer Scoring

Several reports have demonstrated that students are capable to give peer scores that are consistent to that of their teachers (Falchikov & Goldfinch, 2000; Topping, 1998; Van Zundert et al., 2010; Zhang et al., 2020). Also, since peer scoring can be a time-saving approach to judge the outputs of students, especially for big classes, it is not surprising that it has captured the interest of teachers and instructors from different academic programs and educational contexts (Ballantyne et al., 2002; Boud, 1995; Freeman & McKenzie, 2002). Due to the ease of using peer scoring as an educational practice to involve students in the assessment process, several studies have investigated its potential relationship with other variables—including gender.

One of the earliest influential studies that investigated gender bias and gender difference in peer scoring was conducted by Falchikov and Magin (1997) in science and technology and medicine students. In their study, they devised a method to detect gender bias and gender difference in men and women's peer scores to same-sex or opposite-sex group process contribution scores. Gender bias is parallel to the concept of accuracy bias (or peer marking bias or scoring bias), where the difference between the scores given by men or women assessor to same-sex or opposite-sex assessee is calculated, whereas gender difference is computed by getting the difference between men and women assessees scores.

Using this approach, they found the absence of accuracy bias in the peer scores awarded by men and women assessors to same- or opposite-sex peers. In a follow up study, Tucker (2014) also investigated the presence of accuracy bias in over 1500 university students from four degree programs in terms of their contributions to group assignments. Similar to Falchikov and Magin (1997), results showed the absence of accuracy bias between men and women assessors. But it was also observed that men were generous in giving peer scores than women. In spite of the findings that report the absence of accuracy bias in men and women assessors, there are also studies that report the presence of it in other contexts.

In a recent study of more than 1600 students enrolled in economics classes, results suggest that men and women assessors tend to be biased in favour of women assessees in giving peer scores in team-based activities (Espey, 2021). This was also found in Aryadoust's (2016) findings which showed that men assessors tended to score women assessees higher and that they were more inclined to be biased against men assessees in a presentation task. On the contrary, women assessors tended to be biased against women assessees and they were more inclined to score men assessee's presentation higher. This study also pointed out that women were stricter assessors, and that assessor's gender could be a factor in their level of scoring strictness. An opposite pattern can be seen in an earlier study which suggest that assessors tended to give higher scores (+1.9%) to presenters of the same gender in oral presentation tasks (Langan et al., 2005, 2008). Specifically, they found that men assessors tended to give higher scores to fellow men and lower scores to women, while women assessors tended to undervalue their own performance but were more accurate in scoring both genders.

Given these divergent findings, it is important to ascertain if a certain gender is more likely to give/receive higher/lower peer score to/from another gender, since both accuracy bias and gender difference were expressed to be threats to the validity and reliability of peer

assessment (Falchikov & Magin, 1997; Magin, 2001). Importantly, it is difficult to extract patterns from these studies due to the varied nature of peer assessment tasks performed in the reviewed studies (e.g., group/team contribution, oral communication). Thus, this signal the need to further examine gender as a potential deterrent of peer scoring accuracy in other subjects, such as writing focused courses, as there is limited evidence in this area of research. Importantly, there are no studies exploring how assessor's gender affects peer scoring, either to same-sex or opposite-sex peers' writing sample.

Peer Assessment Training

Salient evidence regarding the potential and positive effects of peer assessment training on students' assessment skills has been documented in a number of studies. For instance, students enrolled in a business writing class doing web-based peer assessment were trained through scoring exercises and a discussion of the discrepancies in scoring (against staff rating) of sample outputs resulted to improvement in scoring agreement and writing (Liu et al., 2018). Moreover, one of the studies that proposed a framework for conducting trainings in the peer assessment was done by Lam (2010). In his study, Lam offered a variation of an earlier training approach (Min, 2005, 2006) by dividing it into three stages, which are: (1) modelling—discussion of the rationale of doing peer assessment, and a demonstration of how to give peer assessment using a scoring rubric; (2) exploring— hands-on practice of giving peer assessment using exemplars, and (3) consciousness-raising—analysing the assessments received. Organized training opportunities like this, where opportunities are given to examine the assessment criteria (e.g., scoring rubric, checklists, exemplars), and assistance in assessing an output improves the quality of peer scores, and organizes the nature of the peer assessment process (Li et al., 2016; Miller, 2003). This is important since a recent meta-analysis found that training assessors' scoring skills was the strongest factor in influencing peer assessment, where the effect size of those who received training was significantly larger

than those who did not receive training (Li et al., 2020). Given these findings, training students' assessment capabilities to provide sound peer scores appears to be useful, especially since some studies have observed that students may give peer scores that are not reliable and valid (e.g., Cho & Schunn, 2007).

On the other hand, findings regarding the effects of peer assessment practice seems to be mixed. For example, teacher education students tended to use the scoring criteria more as a result of increased peer assessment experience (Sluijsmans et al., 2004). Moreover, a recent meta-analysis of 58 studies about peer assessment and student learning found a marginal significance with the frequency of giving peer assessment (i.e., number of times giving peer assessment) and learning gains (p = .054; d = 0.207) (Li et al., 2020). On the contrary, another recent meta-analysis of 54 studies investigated the effects of peer assessment on academic performance from primary to tertiary levels across different subject areas and domains found that the frequency of peer assessment practice (single time vs. multiple times) did not differ significantly (Double et al., 2020). Taken altogether, these findings generally point to the idea that repeated practice may improve students' peer assessment; however, this assertion should be taken with great caution.

Importantly, prior research has emphasized that students should be given proper training and practice before conducting peer assessment to become skilled assessors (Sluijsmans et al., 2002). Also, it can be assumed that giving sufficient scaffolding through training and constant exposure to assessment standards and different qualities of work can be instrumental in improving students' evaluative judgement (Tai et al., 2018). Additionally, exposing students with carefully crafted assessment materials (i.e., assessment rubrics) during training and practice may also minimize negative interpersonal outcomes (Panadero et al., 2013). However, studies exploring the effects of peer assessment training in minimising gender difference and accuracy bias towards same-sex or opposite-sex peers has been scarce

in peer assessment literature, in spite of previous recommendations (Torres-Guijarro & Bengoechea, 2017). Nonetheless, studies implementing peer assessment training which compared men and women assessors scoring found mixed patterns. Aryadoust (2016) found that women assessors were in favour of men assessees and were biased against women assessees, while men assessors were in favour of women assessees and were biased against men assessees. On the other hand, Yurdabakan (2011) found that men assessors tended to score men assessees less consistently even after implementing peer assessment training. Given this, it is important to establish further evidence on this area to explore how peer assessment training affects gender and peer assessment outcomes using a quasi-experimental approach.

Peer Assessment and Interpersonal Variables

In recent years, researchers have shown increased interest on the effects of interpersonal and intrapersonal variables on peer assessment (and vice versa) (Panadero, 2016). One of the variables that has been documented to affect peer assessment is the assessor's perceptions of self-efficacy when judging other's work. Bandura (1997) describes self-efficacy as the belief that one can perform a certain task successfully. The concept of efficacy can also be comparable to a student's trust in one's self as an assessor, where they hold the belief that he or she can be a good assessor (Panadero, 2016; van Gennip et al., 2009). Previous studies have seen that students who trust in their abilities as an assessor produced accurate judgements of peer's work (Alqassab et al., 2019). Similarly, assessor's trust in one's self (i.e., efficacy) is related to the scores they give to peers during peer assessment (De Grez et al., 2010). Meanwhile, the relationship between trust and peer assessment training seems to demonstrate that training may not have an effect on trust in one's abilities as an assessor. For instance, Sluijsmans et al. (2002) found students' trust in one's self did not change, even after an intensive peer assessment training. Similarly,

students still did not trust in their abilities as an assessor even after receiving peer assessment training and tutorials on how to assess peer's work (Kilickaya, 2017). On the other hand, findings regarding the difference between men and women in terms of their perceptions of trust in their abilities are scarce. The only exception is the cross-sectional study of Rotsaert et al. (2017) which found no gender difference in trust levels of more than 3000 high school students. Altogether, it is unclear if peer assessment training has an influence on students' trust in one's self as an assessor.

Similarly, assessor's perceptions of comfort when scoring peer's work is also important in peer assessment. In an earlier study, Topping et al. (2000) mentioned that students experience socio-emotional discomfort when assessing peer's work. Some also report that students feel uncomfortable when assessing their classmates' work since they do not like the idea of bringing the scores of their peers down, or some assessors feel that being critical towards peer's work is inappropriate because they are more knowledgeable in the area (Hanrahan & Isaacs, 2001). Furthermore, these studies suggested that in order to address the discomfort experienced during peer assessment, instructors should expose students to assessment training activities before the actual peer assessment (Hanrahan & Isaacs, 2001; Topping et al., 2000). On the other hand, findings revealed that assessors who received training on how to use a rubric to peer score a concept map did not differ in terms of their perceptions of comfort when compared to assessors who did not use a rubric (Panadero et al., 2013). In terms gender's relationship with perceptions of comfort in peer assessment, studies found consistent patterns where women may tend to feel less comfortable and that men feel more comfortable in peer assessment (Rotsaert et al., 2017; Wen & Tsai, 2006, 2008; Zou et al., 2018). Thus, it is important to assure that students, regardless of their gender, feel comfortable in peer assessment. As earlier studies mentioned, peer assessment training

appears to be one of the methods to minimise discomfort. However, studies comparing men and women assessor's comfort after peer assessment training remains scant.

As Panadero (2016, p. 247) puts it, peer assessment "does not happen in a vacuum; rather it produces thoughts, actions, and emotions as a consequence of the interaction of assessees and assessors." Thus, a more careful look at the interplay between peer assessment training, interpersonal variables, and gender in the context of peer assessment activities is essential, since one's gender may affect how they view and participate in various peer assessment activities (Rotsaert et al., 2017; Wen & Tsai, 2006, 2008; Zou et al., 2018).

The Present Study

Our aim in this study is to compare trained and untrained men and women assessors in terms of their peer scores to same- or opposite-sex peer writing samples, as well as their interpersonal reactions after giving peer scores. Hence, this research examines three independent variations ($2 \times 2 \times 2$) using a factorial design, namely: (a) with peer assessment training condition vs. without peer assessment training condition, (b) men assessor vs. women assessor, (c), men assessee writing sample vs. women assessee writing sample. The research questions (RQ) and hypotheses (H) are as follows:

RQ1: Does peer assessment training, assessor's gender and assessee's gender have an effect in peer scoring accuracy?

H1: There will be a positive effect of the training in increasing accuracy and no effect of gender.

RQ2: Does peer assessment training, assessor's gender and assessee's gender have an effect in perceptions of trust in one's self as an assessor?

H2: There will be no effect for training and gender in their perceptions of trust in one's self as an assessor.

RQ3: Does peer assessment training, assessor's gender and assessee's gender have an effect in perceptions of comfort?

H3: There will be no effect for training, and a positive effect for gender in their perceptions of comfort.

Method

Participants

An a priori power analysis for analysis of variance (ANOVA) to detect main effects and interactions was conducted via the G*Power (v. 3.1.9.7; Faul et al., 2009) software to determine the required sample size in this study. Based on accepted minimum standards (Murphy et al., 2014), a medium effect size of $f = .25$ and a power of .80 was assumed for eight groups and resulted to a minimum of 128 participants to achieve the assumed power. Additional participants were recruited to account for participant dropout.

A total of 145 second year undergraduate students majoring in psychology participated in this study. The participants belong to four experimental psychology classes at a mid-size private university in Manila, Philippines. There were 25 men and 120 women participants whose age ranged from 19 to 24 years (M= 20.17; SD= 0.89). Two of the classes attended the peer assessment training workshop (n= 74; n_{Mens}= 13, n_{Women}= 61), while the other two classes did not attend the peer assessment training workshop (n= 71; n_{Men}= 12, n_{Women}= 59). Since there are only few men psychology students enrolled in the course we decided to divide them equally among the two conditions. Among the participants, 60.7% have experience giving peer scores (52.7% within the training condition, and 69% within the no training condition) in their previous courses. All participants received extra credit for their participation.

Course Context

The chosen course to run our study was experimental psychology as it is a writing focused course. Enrolled students were required to produce an experimental research paper in groups as a major requirement, and instructors also use peer assessment as one of the assessment strategies in the course. In the present study, four instructors taught the subject online due to pandemic restrictions. Instructors met with students synchronously (5 hours a week) for lectures and discussions, and asynchronously (2 hours a week) for research activities and exercises. The study was carried out from the 7th through the 10th week of a 13-week semester.

Materials

Writing Samples

The researchers asked the instructors for three writing samples of previous work. The three writing samples provided by the instructors were of poor quality (about online learning), average quality (about passion in teachers), and excellent quality (about student burnout). Three independent expert raters (i.e., psychological research instructor, English writing instructor, and educational assessment expert) scored the writing samples using the same writing rubric the participants used. The expert ratings confirmed the quality difference between the three writing samples, with an excellent final interrater reliability (ICC= 0.98; Hallgren, 2012).

Writing Rubric

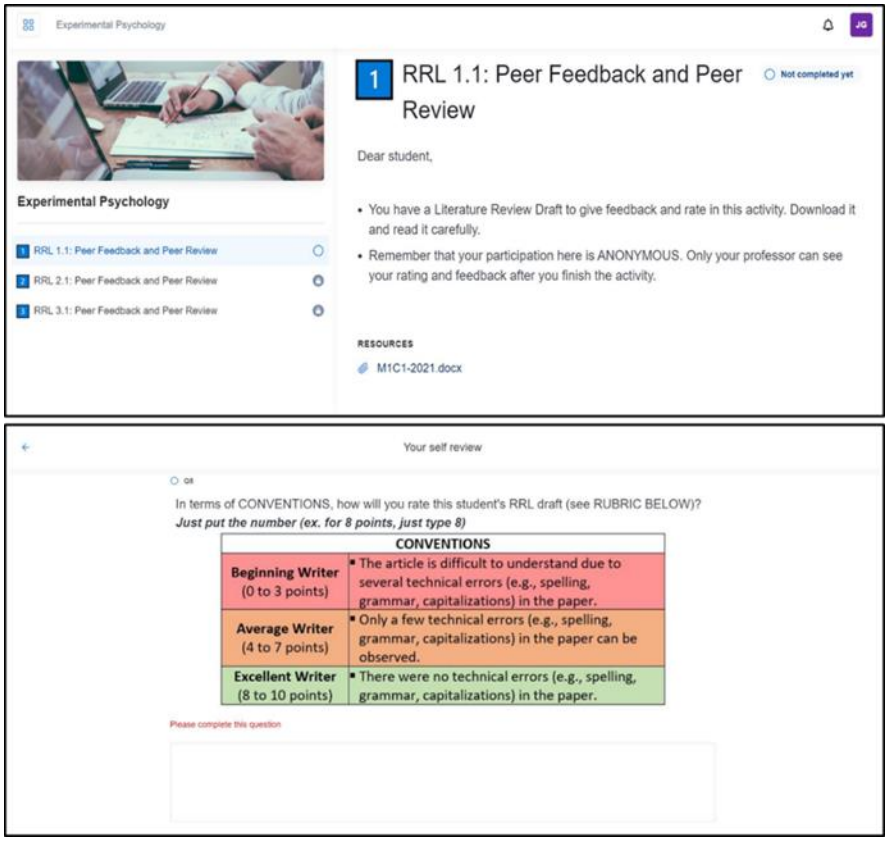
This study adapted a rubric recommended by the psychology department where the data collection took place. It contains four dimensions: content, organization and fluency, conventions, and intext citations and references. Some dimensions of the rubric (i.e., methodology, results, and discussion) were omitted because they did not apply for the selected writing samples.

Online Peer Assessment Platform

Eduflow was utilised to assess the writing samples in this study. Eduflow is a learning management system, where students can submit and self- and peer assess work using any web-enabled device (e.g., laptop, tablet, smartphone). Eduflow's self-assessment function was used to facilitate the activity. Participants turned in their assessments in the review page of each of the writing sample (see figure 1).

Figure 1

Participants' PA Page in Eduflow.



Measures

Peer Scores

Participants scored each dimension of the rubric from 0 (lowest) to 10 (highest), where scores from 0 to 3 points were considered as poor, 4 to 7 points as average, and 8 to 10 points as excellent. The scores of the participants in each dimension of the rubric were

summed to obtain their peer score of a writing sample. Scores of: 0 to 15 would mean that the output is poor, 16 to 31 would mean that the output is average, and 32 to 40 would mean that the output is excellent. The highest possible score a writing sample could get is 40 points.

Peer Assessment Experience and Perceptions Scale

Participants answered a 3-item post peer assessment scale about peer scoring after all writing samples were turned in. The first question was about prior peer scoring experience: "Have you experienced giving peer scores before?" (Yes or No). The second question was about trust in one's ability: "How effective did you feel in giving peer scores to your peer's review of related literature drafts?" (1 = Extremely Ineffective to 5 = Extremely Effective). The last question was about their perceptions of comfort: "How comfortable did you feel in giving peer scores to your peer's review of related literature draft?" (1 = Extremely Uncomfortable to 5 = Extremely Comfortable).

Procedure

Prior to the implementation of the intervention, the first author met with the instructors to explain the nature of the project. The instructors volunteered their classes to be in either the peer assessment training or no peer assessment training group. Subsequently, the instructors told the students that they will have a feedback training activity in preparation for the assessments they will give to their peers' experimental research paper. It was also mentioned that an external speaker will facilitate this activity.

This study is part of a bigger project looking at the effects of gender on peer assessment, only the information about peer scoring will be discussed in this section. Since we did not want the participants to get a hint that they will participate in a gender experiment, the training we facilitated did not cover gender-related issues. Instead, students in the training group received peer assessment training based on a modified version of Lam's (2010) training framework, which lasted for around two and a half hours. First, the first author

discussed each of the dimensions of the scoring rubric to the participants. Exemplars were also shown to the participants to assist their understanding of different dimensions of the rubric. Second, the participants practiced what they learned in the previous step on a writing sample using the same rubric. Then, participants' scores were discussed in unison to clarify their confusions in the text or the rubric. Lastly, participants were given a walkthrough of Eduflow and specific instructions on how they will deliver their peer scores in the website. The third stage (i.e., consciousness raising) of Lam's (2010) training framework was not used since it was not applicable to the design of the study. On the other hand, participants in the no training group were only given a walkthrough of Eduflow and instructions on how to turn in their peer scores, which lasted for about 30 minutes.

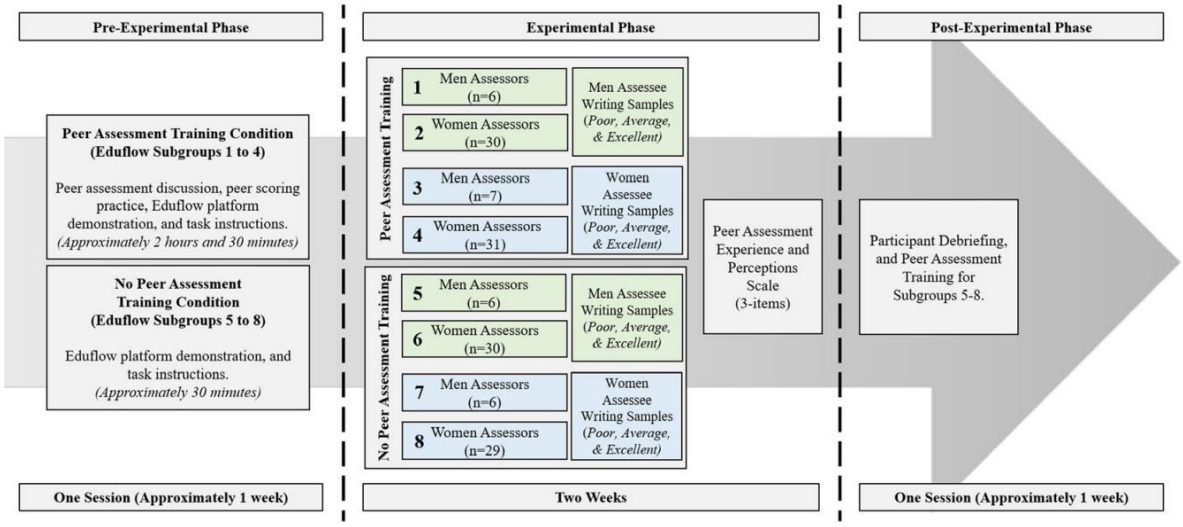
In order to separate students assigned in the peer assessment training and no peer assessment training groups in Eduflow, the researcher created two identical classes with four subgroups in each class (i.e., total of eight subgroups). The first and fifth subgroups were for trained or untrained men assessors peer scoring men writing samples ($n_{Training} = 6$; $n_{No\ Training} = 6$), the second and sixth subgroups were for trained or untrained women assessors peer scoring men writing samples ($n_{Training} = 30$; $n_{No\ Training} = 30$), the third and seventh subgroups were for trained or untrained men assessors peer scoring women writing samples ($n_{Training} = 7$; $_{n}No\ Training = 6$), and the fourth and eighth subgroups were for trained or untrained women assessors peer scoring women writing samples ($n_{Training} = 31$; $n_{NoTraining} = 29$).

Once all instructions were delivered to the participants, the researcher activated each students' account to start the activity. The instructors gave the students two-weeks to complete the activity. After scoring the third writing sample, the participants answered the peer assessment experience and perceptions scale. Once everyone has completed the activity, the researcher met with the participants to debrief them about the design the study. Students

in the no peer assessment training group were also given the same peer assessment training during the debriefing session. Figure 2 summarizes the flow of procedures in this study.

Figure 2

Study procedure flow



Research design and data analysis

Our research design was a randomized controlled trial with 2 training conditions (intervention vs. control) × 2 assessor gender (men vs. women) × 2 assessee gender (men vs. women). Our dependent variables were the three writing samples and the participants' perceptions of trust in one's self and comfort. We ran factorial ANOVA to answer our research questions. All pairwise comparisons for the factorial ANOVA were Bonferroni corrected. No post hoc test was done because the three independent variables only has two levels.

Since we have an unbalanced number of men and women respondents in this study, we followed previous gender and peer assessment study's strategy for unequal number of men and women by conducting a Levene's test to check for variance between the groups (Noroozi et al., 2022). Our results showed no difference between men and women in the poor ($F= 1.970$; $p=.163$), average ($F= 0.226$; $p=.635$), and excellent ($F= 0.003$; $p=.958$) writing

samples; as well as for perceptions of comfort (F= 2.306; *p*=.131). This suggests that the unbalanced number of men and women assessees were not a problem (Sweet & Grace-Martin, 2011). However, we found a significant difference for perceptions of trust (F=6.091; *p*=.015). Nonetheless, we still proceeded with factorial ANOVA, as it is robust enough for violations of this assumption (Blanca et al., 2017).

Results

The means and standard deviations of peer scores given by trained or untrained men or women assessor for a men or women assessee for the three writing samples are summarized in table 1.

Table 1

Means, Standard Deviations, and Total Peer Scores for the Three Writing Samples

			Poor Writing Sample		Average Writing Sample		Excellent Writing Sample	
Condition	Assessor Gender	Assessee Gender	*M*	*SD*	*M*	*SD*	*M*	*SD*
Training	Men	Men	13.50	6.98	25.33	7.58	33.00	5.40
		Women	15.00	9.10	27.86	10.81	35.14	5.21
		Total	14.31	7.89	26.69	9.16	34.15	5.19
	Women	Men	17.19	9.16	27.45	7.64	33.68	5.10
		Women	15.97	8.03	30.80	6.06	33.57	5.70
		Total	16.59	8.57	29.10	7.06	33.62	5.36
	Total	Men	16.59	8.87	27.11	7.57	33.57	5.08
		Women	15.78	8.11	30.24	7.10	33.86	5.58
		Total	16.19	8.45	28.68	7.46	33.72	5.30
No Training	Men	Men	21.17	12.91	28.33	5.89	34.50	5.47
		Women	21.67	7.45	23.67	8.62	29.50	6.06
		Total	21.42	10.0	26.00	7.45	32.00	6.09
	Women	Men	14.07	7.02	27.38	8.38	32.41	5.36
		Women	14.60	7.29	25.20	8.45	30.57	7.12
		Total	14.34	7.09	26.27	8.41	31.47	6.33
	Total	Men	15.29	8.51	27.54	7.94	32.77	5.36
		Women	15.78	7.68	24.94	8.37	30.39	6.88
		Total	15.54	8.05	26.23	8.21	31.56	6.25
Total	Men	Men	17.33	10.67	26.83	6.66	33.75	5.24
		Women	18.08	8.74	25.92	9.70	32.54	6.12
		Total	17.72	9.51	26.36	8.22	33.12	5.63
	Women	Men	15.68	8.27	27.42	7.94	33.07	5.22
		Women	15.28	7.63	28.00	7.82	32.07	6.57

	Total	15.48	7.93	27.71	7.85	32.57	5.93
Total	Men	15.96	8.65	27.32	7.70	33.18	5.20
	Women	15.78	7.85	27.63	8.15	32.15	6.45
	Total	15.87	8.23	27.48	7.90	32.66	5.87

The three expert rater scores were averaged for each writing sample, being $M = 11.00$ ($SD = 4.58$) for the poor writing sample, $M = 20.00$ ($SD = 5.19$) for the average writing sample, and $M = 33.00$ ($SD = 5.29$) for the excellent writing sample. We also computed for the difference between assessor scores and expert scores for each writing sample (see table 2). Overall, majority of the assessors over scored the poor and average writing samples, regardless of training condition or gender. However, trained and untrained assessors showed opposite scoring patterns for the excellent writing sample.

Table 2

Difference Between Expert Scores and Peer Assessor Scores

Condition	Assessor Gender	Poor Writing Sample	Average Writing Sample	Excellent Writing Sample
Training	Men	+3.31	+6.69	+1.15
	Women	+5.59	+9.10	+0.62
	Total	+5.19	+6.68	+0.72
No Training	Men	+10.42	+6.00	-1.0
	Women	+3.34	+6.27	-1.53
	Total	+4.54	+6.23	-1.44
Total	Men	+6.62	+6.36	+0.12
	Women	+4.48	+7.71	-0.43
	Total	+4.87	+7.48	-0.34

Note. Positive value indicates that the peer assessor over scored, negative value indicates that the peer assessor under scored.

RQ1: Does peer assessment training, assessor's gender and assessee's gender have an effect in peer scoring accuracy?

For the poor writing sample, we found that the interaction between condition and assessors' gender was significant, $F(1, 137) = 6.83$, $p = .010$, partial $\eta 2 = .048$ (small effect; Cohen, 1988). Specifically, the simple effects analysis showed that the peer scores of men assessors in the no peer assessment training group were significantly higher than the men

assessors in the peer assessment training group, $F(1, 137) = 4.784$, $p = .030$, partial $\eta 2 = .034$ (small effect), while the peer scores of trained and untrained women assessors were not statistically different. This means that the untrained men scored significantly higher the poor writing sample. On the other hand, no significant difference was observed in the triple interaction of condition × assessor gender × assessee gender for the poor writing sample, $F(1, 137) = .147$, $p = .702$, partial $\eta^2 = .001$ (small effect). This implies that there were no differences in the peer scores awarded by trained or untrained men and women assessors to either men or women assessee's poor writing sample. Table 3 shows the *p*-values for all effects for the poor writing sample.

When it comes to the average writing sample, none of the interactions nor the simple effects were significant (Table 3). The only effect that came close to significance was the condition × assessee gender interaction, $F(1, 137) = 3.404$, $p = .067$, partial $\eta 2 = .024$ (small effect), but it was still not significant. The triple interaction effect of condition × assessor gender × assessee gender on the peer scores was not significant either, $F(1, 137) = .058$, $p = .810$, partial $\eta 2 = .000$ (small effect). Similar to the outcome of the poor writing sample, there were no differences in the peer scores awarded by trained or untrained men and women assessors to either men or women assessee's average writing sample.

Finally, regarding the excellent writing sample, the triple interaction of condition × assessor gender × assessee gender on the peer score was also not significant $F(1, 137) = 1.112$, $p = .294$, partial $\eta 2 = .008$ (small effect). Like the results for the poor and average writing samples, this forwards that there were no differences in the peer scores awarded by trained or untrained men and women assessors to either men or women assessee excellent writing sample. Just like with the average writing sample, the only effect that came close to significance was the condition × assessee gender interaction, $F(1, 137) = 2.999$, $p = .086$,

partial η2 = .021 (small effect). Table 3 shows the *p*-values for all effects for the excellent writing sample.

Across the three writing samples, peer assessment training did not increase assessor's peer scoring accuracy, and gender had no effect on peer scoring accuracy. Thus, our findings partially supported our first hypothesis.

RQ2: Does peer assessment training, assessor's gender and assessee's gender have an effect in perceptions of trust in one's self as an assessor?

From the intervention effects on trust, two interactions were significant. First, the condition × assessee gender interaction was significant, $F(1,137) = 8.816$, $p = 0.004$, partial η2 = .06 (small effect), peer assessment training $M = 4.04$ ($SD = 0.86$), and no peer assessment training $M = 3.80$ ($SD = 0.86$). The simple effects analysis showed that assessors trusted themselves less when peer scoring women assessees, $F(1,137) = 13.528$, $p = .000$, partial η2 = .090 (medium effect) in the no peer assessment training condition $F(1,137) = 10.718$, $p = .001$, partial η2 = .073 (medium effect). Additionally, the triple interaction of training condition × assessor gender × assessee gender revealed statistical significance, $F(1,137) = 5.44$, $p = 0.021$, partial η2 =.038 (small effect). Simple effects analysis revealed that untrained assessors trusted themselves less than trained assessors only for men assessors $F(1, 137) = 12.120$, $p = .001$, partial η2 =.081 (medium effect), and only when peer scoring women assessees $F(1,137) = 8.658$, $p = .004$, partial η2 =.059 (small effect). Table 4 shows the *p*-values for all effects for perceptions of trust.

Our findings revealed that training and gender had an effect on assessor's perceptions of trust in one's self as an assessor after peer scoring the writing samples. Thus, our second hypothesis was rejected.

Table 3

Factorial ANOVA Results for the Poor, Average, and Excellent Writing Samples

	Poor Writing Sample			Average Writing			Excellent Writing		
	F (1,137)	*p*	partial η^2	*F* (1,137)	*p*	partial η^2	*F* (1,137)	*p*	partial η^2
Condition	1.868	.174	.013	0.991	.321	.007	2.688	.103	.019
Assessor Gender	1.742	.189	.013	0.669	.415	.005	0.140	.709	.001
Assessee Gender	0.033	.857	.000	0.020	.888	.000	0.882	.349	.006
Condition × Assessor Gender	6.833	.010	.048	0.423	.517	.003	0.001	.981	.000
Condition × Assessee Gender	0.011	.916	.000	3.404	.067	.024	2.999	.086	.021
Assessor Gender × Assessee Gender	0.140	.709	.001	0.231	.632	.002	0.031	.861	.000
Condition × Assessor Gender × Assessee Gender	0.147	.702	.001	0.058	.810	.000	1.112	.294	.008

RQ3: Does peer assessment training, assessor's gender and assessee's gender have an effect in perceptions of comfort?

From the intervention effects on comfort, two interactions were significant. First, the condition × assessee gender interaction was significant, $F(1,137) = 5.861$, $p = .017$, partial η2 =.041 (small effect), peer assessment training $M = 3.72$ ($SD = 0.91$), and no peer assessment training $M = 3.54$ ($SD = 0.94$). The simple effects analysis further revealed that assessors felt more comfortable in peer scoring women assessees, $F(1,137) = 7.294$, p = .008, partial η2 =.051 (small effect), when they received peer assessment training, $F(1,137) = 5.659$, $p = .019$, partial η2 =.040 (small effect). Also, the interaction between condition × assessor gender × assessee gender was found to be significant, $F(1,137) = 4.353$, $p = .039$, partial η2 =.031 (small effect). Specifically, simple effects analysis showed that only untrained men assessors felt uncomfortable when peer scoring women written writing sample $F(1,137) = 4.522$, $p = .035$, partial η2 =.032 (small effect). Regardless of peer assessment training condition received, women assessors were comfortable in peer scoring men or women writing samples. On the other hand, only trained men assessors were comfortable in peer scoring writing samples from both genders. Untrained men assessors only felt comfortable when they peer score men written writing samples. Also, it should also be highlighted that trained men assessors who peer scored men assessee writing samples may have also felt uncomfortable than trained men assessors peer scoring women assessee writing samples since the simple effects analysis *p*-value was close to significance ($p = .056$). Table 4 shows the *p*-values for all effects for perceptions of comfort.

Altogether, peer assessment training and gender had an effect on assessor's perception of comfort. Thus, our third hypothesis was partially supported.

Table 4

Factorial ANOVA Results for Interpersonal Variables

	Trust			Comfort		
	F (1,137)	*p*	partial η^2	*F* (1,137)	*p*	partial η^2
Condition	4.629	.033	.033	1.793	.183	.013
Assessor Gender	1.427	.234	.010	1.047	.308	.008
Assessee Gender	3.020	.084	.022	0.786	.377	.006
Condition × Assessor Gender	1.795	.183	.013	0.459	.499	.003
Condition × Assessee Gender	8.816	.004	.060	5.861	.017	.041
Assessor Gender × Assessee Gender	1.226	.270	.009	0.281	.597	.002
Condition × Assessor Gender × Assessee Gender	5.441	.021	.038	4.353	.039	.031

Discussion

Previous research on the effects of gender on peer assessment were mixed. Therefore, our aim was to compare trained and untrained men and women assessors in terms of their peer scores to men or women assessees' writing samples of varying quality (i.e., poor, average, excellent) in a quasi-experimental approach, while exploring also the effects on interpersonal variables. Next, we discuss our results organized around the RQs.

In RQ1, we investigated if peer assessment training, assessor's gender, and assessee's gender had an effect in peer scoring accuracy. Our results were non-significant, trained or untrained men and women assessor did not differ in their peer scores of men or women assessees' writing sample. Importantly, this is contrary to previous findings that found men may tend to over score their peers (Kaufman et al., 2000; Tucker, 2014) and that women may tend to receive higher scores than men counterparts (Espey, 2021; Sasmaz Oren, 2012; Tucker, 2014; Yurdabakan, 2011); but also consistent with previous studies that found no gender difference (e.g., Aryadoust, 2016; Langan et al., 2005, 2008) and accuracy bias (e.g., Falchikov & Magin, 1997; Tucker, 2014) in peer scores. A plausible explanation that can be deduced from this finding was previously described by Falchikov and Magin (1997), and was also found in Tucker's (2014) follow up investigation. They found that conducting multiple

peer assessments appear to cancel the effects of same-sex or opposite-sex bias in peer assessment. In our study, since participants peer scored three writing samples, this may have also cancelled potential biases during peer assessment.

When it comes to the peer assessment training's effects on peer scores assessors give, the main effect results for the three writing samples showed non-significant results. This indicates that trained and untrained assessors scored the writing samples almost similarly. This finding is in contrast to the study of Liu et al. (2018) which found differences in accuracy in peer scoring a writing draft after receiving peer assessment training. However, this should be taken with caution since the interaction between peer assessment training condition and assessor gender was significant for the poor writing sample, where untrained men assessors peer scored the poor writing sample higher than the trained men assessor. Additionally, although not significant, trained and untrained assessors showed opposite scoring patterns for the excellent writing sample, as evidenced in the discrepancy between expert and assessor scores. Where trained assessors slightly over scored, and untrained assessor underscored the writing sample. There are at least two viable explanations for this. First, since the scoring rubric used in this study was a rubric recommended by the department, our participants in both conditions may have been exposed to the rubric in their other courses. Second, since the percentage of participants with peer scoring experience in previous courses were greater for the untrained assessors (i.e., 69% vs. 52.7%) it can be assumed that their prior experience and exposure to peer scoring may have affected how they scored the writing samples in this study. Also, although not significant, the interaction between peer assessment training condition and assessee gender were the closest to significance for both average (p = .067) and excellent (p = .087) writing samples. This may give us a hint that there might be gender differences in the peer scores received by either men

or women assessees for more complex outputs, but assessor's peer assessment training (and untrained assessor's prior experience of peer assessment) might have mitigated this.

Regarding RQ2, we sought to determine if peer assessment training, assessor's gender, and assessee's gender have an effect in perceptions of trust in one's self as an assessor. We found a significant triple interaction: men assessors who did not receive peer assessment training trusted themselves less when they peer scored women assessee's writing samples. Untrained men assessors may have felt less equipped to peer score women writing samples due to the global view that women perform better academically (Ellis et al., 2008; Gibb et al., 2008), and exhibit better writing abilities (Al-Saadi, 2020; Reynolds et al., 2015). This is similar with findings that mention that assessors' lack of peer assessment training (or experience) affects how they feel towards peer assessment activities (Kilickaya, 2017). Furthermore, our study is in contrast with the result of Rotsaert et al. (2017) which found no difference in men and women's levels of trust in peer assessment. In our study, trained assessors may have trusted their abilities to accurately score the writing outputs since the scoring rubric was explained to them in detail and they practiced on similar writing samples during the peer assessment training.

For RQ3, we wanted to determine if peer assessment training, assessor's gender, and assessee's gender have an effect in their perceptions of comfort. Our results showed that regardless of the training condition, women assessors felt comfortable peer scoring the writing samples from both genders. This is in contrast with the salient evidence that report that women feel less comfortable during peer assessment activities (Rotsaert et al., 2017; Wen & Tsai, 2006, 2008; Zou et al., 2018). Also, this partially supports Panadero et al.'s (2013) findings that trained and untrained assessors do not differ in their perceptions of comfort. Similar to the findings in perceptions of trust, untrained women assessors in this study may have drawn comfort from their previous experience of peer assessment in other

courses. Only trained men assessors were comfortable in peer scoring writing samples from both genders, while untrained men assessors only felt comfortable when they peer scored men writing samples.

In summary, our findings suggest that peer assessment training has an effect on assessor's perceptions of trust and comfort, especially for men students. Since results show that untrained men assessors trusted their abilities less and felt less comfortable when peer scoring women assessees, we recommend to pay more attention to men students as they are more prone to experience interpersonal problems during peer assessment. Specifically, we recommend a multifaceted training approach to address potential problems that may affect peer assessment. First, instructors should carefully plan and organise how peer assessment is to be implemented, and provide students with sufficient assessment scaffolds (e.g., exposure and explanation of rubrics, providing prompts, exposure to exemplars, etc.), as well as multiple opportunities practice performing peer assessment to enhance peer assessment skills (see Panadero et al., 2016). Second, we give importance to the suggestions by prior studies to implement gender-awareness workshops to mitigate unconscious gender issues (e.g., attitudes and perceptions) that students may hold in peer assessment (Langan et al., 2008; Torres-Guijarro & Bengoechea, 2017). Third, recent studies have explored the potential of training students' interpersonal and intrapersonal reactions (e.g., pressure, safety, confidence) during peer assessment activities (Li, 2017; Senden et al., 2022). Therefore, implementing programs that alleviate students' negative perceptions towards peer assessment that encourage them to feel more positive about their own capabilities and affect is also recommended.

Limitations and future lines of research

The first limitation is the low number of men participants, to enhance the comparability, we divided equally in the two training conditions. Nonetheless, it is recommended for future research to have gender balanced conditions if exploring gender

effects. The second limitation is that, due to pandemic constraints, the peer assessment training was only delivered in one online session. We recommend that incoming studies implement longer interventions, though we have found effects of our intervention—even if short—probably due to the intensity of it (e.g. use of rubric, exemplars, practice). Third, we only utilised peer assessment training as the primary intervention that could limit gender difference and peer scoring bias in this study. Other interventions, such as different anonymity conditions, is also an area worth exploring since evidence has mentioned that girls may tend to hold negative interpersonal processes and may tend to prefer anonymous peer assessment over boys (Rotsaert et al., 2017).

Conclusion

In summary, we compared trained and untrained men and women assessors in terms of their peer scoring accuracy to men or women assessees' writing samples of varying quality, as well as their perceptions of trust and comfort using a quasi-experimental design. Our results suggest that across the three writing samples, peer assessment training did not increase assessor's peer scoring accuracy, and gender had no effect on peer scoring accuracy. On the other hand, we found that peer assessment training and gender had an effect on assessor's perceptions of trust and comfort. Our study responds to the methodological dilemma in peer assessment research presented more than a decade ago, by initiating a quasi-experimental study on the differential effects of gender in peer assessment (Topping, 2010). Also, our study provides an insight on gender's effects in peer scoring accuracy in a writing focused context. Thus, in clear contrast to previous research, our study used a quasi-experimental research design from which we can extract stronger conclusions. Also, this extends the study of gender in peer assessment outside oral presentation and group contribution activities, which were the typical focus of previous studies. In conclusion, the gender of the assessor and assessee only showed marginal influences so teachers can be

confident that gender difference and peer scoring bias does not seem to be a problem in peer scoring given similar circumstances to the ones in this study (e.g. rubric use).

References

Adachi, C., Tai, J., & Dawson, P. (2018). A framework for designing, implementing, communicating and researching peer assessment. *Higher Education Research & Development*, *37*(3), 453–467. https://doi.org/10.1080/07294360.2017.1405913

Alqassab, M., Strijbos, J.-W., & Ufer, S. (2019). Preservice mathematics teachers' beliefs about peer feedback, perceptions of their peer feedback message, and emotions as predictors of peer feedback accuracy and comprehension of the learning task. *Assessment & Evaluation in Higher Education*, *44*(1), 139–154. https://doi.org/10.1080/02602938.2018.1485012

Al-Saadi, Z. (2020). Gender differences in writing: The mediating effect of language proficiency and writing fluency in text quality. *Cogent Education*, *7*(1), 1770923. https://doi.org/10.1080/2331186X.2020.1770923

Aryadoust, V. (2016). Gender and Academic Major Bias in Peer Assessment of Oral Presentations. *Language Assessment Quarterly*, *13*(1), 1–24. https://doi.org/10.1080/15434303.2015.1133626

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing Procedures for Implementing Peer Assessment in Large Classes Using an Action Research Process. *Assessment & Evaluation in Higher Education*, *27*(5), 427–441. https://doi.org/10.1080/0260293022000009302

Bandura, A. (1997). *Self-efficacy: The exercise of control* (pp. ix, 604). W H Freeman/Times Books/ Henry Holt & Co.

Blanca, M. J., Alarcón, R., & Arnau, J. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema*, *29.4*, 552–557. https://doi.org/10.7334/psicothema2016.383

Boud, D. (1995). Self and peer marking in a large technical subject. In *Enhancing Learning Through Self-assessment*. Routledge.

Cheng, K.-H., & Hou, H.-T. (2015). Exploring students' behavioural patterns during online peer assessment from the affective, cognitive, and metacognitive perspectives: A progressive sequential analysis. *Technology, Pedagogy and Education*, *24*(2), 171–188. https://doi.org/10.1080/1475939X.2013.822416

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education*, *48*(3), 409–426. https://doi.org/10.1016/j.compedu.2005.02.004

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

De Grez, M., Valcke, M., & Berings, L. (2010). Peer assessment of oral presentation skills. *Procedia - Social and Behavioral Sciences*, *2*(2), 1776–1780. https://doi.org/10.1016/j.sbspro.2010.03.983

Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2020). The Impact of Peer Assessment on Academic Performance: A Meta-analysis of Control Group Studies. *Educational Psychology Review*, *32*(2), 481–509. https://doi.org/10.1007/s10648-019-09510-3

Ellis, L., Karadi, K., Hershberger, S., Field, E., Wersinger, S., Pellis, S., Geary, D., Palmer, C., Hoyenga, K., & Hetsroni, A. (2008). *Sex differences: Summarizing more than a century of scientific research* (pp. xvii, 972). Psychology Press.

Espey, M. (2021). Gender and peer evaluations. *The Journal of Economic Education*, *0*(0), 1–10. https://doi.org/10.1080/00220485.2021.2004277

Falchikov, N. (2005). *Improving Assessment through Student Involvement: Practical Solutions for Aiding Learning in Higher and Further Education*. https://www.routledge.com/Improving-Assessment-through-Student-Involvement-Practical-Solutions-for/Falchikov/p/book/9780415308212

Falchikov, N., & 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. https://doi.org/10.3102/00346543070003287

Falchikov, N., & Magin, D. (1997). Detecting Gender Bias in Peer Marking of Students' Group Process Work. *Assessment & Evaluation in Higher Education*, *22*(4), 385–396. https://doi.org/10.1080/0260293970220403

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Freeman, M., & McKenzie, J. (2002). SPARK, a confidential web–based template for self and peer assessment of student teamwork: Benefits of evaluating across different subjects. *British Journal of Educational Technology*, *33*(5), 551–569. https://doi.org/10.1111/1467-8535.00291

Gibb, S. J., Fergusson, D. M., & Horwood, L. J. (2008). Gender Differences in Educational Achievement to Age 25. *Australian Journal of Education*, *52*(1), 63–80. https://doi.org/10.1177/000494410805200105

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Hanrahan, S. J., & Isaacs, G. (2001). Assessing Self- and Peer-assessment: The students' views. *Higher Education Research & Development*, *20*(1), 53–70. https://doi.org/10.1080/07294360123776

Kaufman, D. B., Felder, R. M., & Fuller, H. (2000). Accounting for Individual Effort in Cooperative Learning Teams. *Journal of Engineering Education*, *89*(2), 133–140. https://doi.org/10.1002/j.2168-9830.2000.tb00507.x

Kearney, S. (2013). Improving engagement: The use of 'Authentic self-and peer-assessment for learning' to enhance the student learning experience. *Assessment & Evaluation in Higher Education*, *38*(7), 875–891. https://doi.org/10.1080/02602938.2012.751963

Kilickaya, F. (2017). Peer assessment of group members in tertiary contexts. In M. Sowa & J. Krajka (Eds.), *Innovations in languages for specific purposes—Present challenges and future promises* (pp. 329–343). Peter Lang.

Kim, M., & Ryu, J. (2013). The development and implementation of a web-based formative peer assessment system for enhancing students' metacognitive awareness and performance in ill-structured tasks. *Educational Technology Research and Development*, *61*(4), 549–561. https://doi.org/10.1007/s11423-012-9266-1

Lam, R. (2010). A Peer Review Training Workshop: Coaching Students to Give and Evaluate Peer Feedback. *TESL Canada Journal*, 114–114. https://doi.org/10.18806/tesl.v27i2.1052

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

Langan, A. M., Wheater, C. P., Shaw, E. M., Haines, B. J., Cullen, W. R., Boyle, J. C., Penney, D., Oldekop, J. A., Ashcroft, C., Lockey, L., & Preziosi, R. F. (2005). Peer assessment of oral presentations: Effects of student gender, university affiliation and participation in the development of assessment criteria. *Assessment & Evaluation in Higher Education*, *30*(1), 21–34. https://doi.org/10.1080/0260293042003243878

Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education*, *45*(2), 193–211. https://doi.org/10.1080/02602938.2019.1620679

Li, H., Xiong, Y., Zang, X., L. Kornhaber, M., Lyu, Y., Chung, K. S., & K. Suen, H. (2016). Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education*, *41*(2), 245–264. https://doi.org/10.1080/02602938.2014.999746

Li, L. (2017). The role of anonymity in peer assessment. *Assessment & Evaluation in Higher Education*, *42*(4), 645–656. https://doi.org/10.1080/02602938.2016.1174766

Liu, X., Li, L., & Zhang, Z. (2018). Small group discussion as a key component in online assessment training for enhanced student learning in web-based peer assessment. *Assessment & Evaluation in Higher Education*, *43*(2), 207–222. https://doi.org/10.1080/02602938.2017.1324018

Magin, D. (2001). Reciprocity as a Source of Bias in Multiple Peer Assessment of Group Work. *Studies in Higher Education*, *26*(1), 53–63. https://doi.org/10.1080/03075070020030715

Miller, P. J. (2003). The Effect of Scoring Criteria Specificity on Peer and Self-assessment. *Assessment & Evaluation in Higher Education*, *28*(4), 383–394. https://doi.org/10.1080/0260293032000066218

Min, H.-T. (2005). Training students to become successful peer reviewers. *System*, *33*(2), 293–308. https://doi.org/10.1016/j.system.2004.11.003

Min, H.-T. (2006). The effects of trained peer review on EFL students' revision types and writing quality. *Journal of Second Language Writing*, *15*(2), 118–141. https://doi.org/10.1016/j.jslw.2006.01.003

Murphy, K. R., Myors, B., & Wolach, A. (2014). *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests, Fourth Edition* (4th ed.). Routledge. https://doi.org/10.4324/9781315773155

Noroozi, O., Banihashem, S. K., Taghizadeh Kerman, N., Parvaneh Akhteh Khaneh, M., Babayi, M., Ashrafi, H., & Biemans, H. J. A. (2022). Gender differences in students' argumentative essay writing, peer review performance and uptake in online learning environments. *Interactive Learning Environments*, 1–15. https://doi.org/10.1080/10494820.2022.2034887

Panadero, E. (2016). Is It Safe? Social, Interpersonal, and Human Effects of Peer Assessment: A Review and Future Directions. In *Handbook of Human and Social Conditions in Assessment* (pp. 247–266). Routledge.

Panadero, E., Jonsson, A., & Strijbos, J.-W. (2016). Scaffolding Self-Regulated Learning Through Self-Assessment and Peer Assessment: Guidelines for Classroom Implementation. In D. Laveault & L. Allal (Eds.), *Assessment for Learning: Meeting the Challenge of Implementation* (pp. 311–326). Springer International Publishing. https://doi.org/10.1007/978-3-319-39211-0_18

Panadero, E., Romero, M., & Strijbos, J.-W. (2013). 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*, *39*(4), 195–203. https://doi.org/10.1016/j.stueduc.2013.10.005

Reinholz, D. (2016). The assessment cycle: A model for learning through peer assessment. *Assessment & Evaluation in Higher Education*, *41*(2), 301–315. https://doi.org/10.1080/02602938.2015.1008982

Reynolds, M. R., Scheiber, C., Hajovsky, D. B., Schwartz, B., & Kaufman, A. S. (2015). Gender Differences in Academic Achievement: Is Writing an Exception to the Gender Similarities Hypothesis? *The Journal of Genetic Psychology*, *176*(4), 211–234. https://doi.org/10.1080/00221325.2015.1036833

Rotsaert, T., Panadero, E., Estrada, E., & Schellens, T. (2017). How do students perceive the educational value of peer assessment in relation to its social nature? A survey study in Flanders. *Studies in Educational Evaluation*, *53*, 29–40. https://doi.org/10.1016/j.stueduc.2017.02.003

Sasmaz Oren, F. (2012). The effects of gender and previous experience on the approach of self and peer assessment: A case from Turkey. *Innovations in Education and Teaching International*, *49*(2), 123–133. https://doi.org/10.1080/14703297.2012.677598

Senden, M., De Jaeger, D., & Coertens, L. (2022). *Can students feel safe and confident during peer feedback? A quasi-experiment testing a training efficiency*. Paper Presented at the EARLI SIG1+4 Conference, Cadiz, Spain.

Sluijsmans, D., Brand-Gruwel, S., & van Merriënboer, J. J. G. (2002). Peer Assessment Training in Teacher Education: Effects on performance and perceptions. *Assessment & Evaluation in Higher Education*, *27*(5), 443–454. https://doi.org/10.1080/0260293022000009311

Sluijsmans, D., Brand-Gruwel, S., van Merriënboer, J. J. G., & Martens, R. L. (2004). Training teachers in peer-assessment skills: Effects on performance and perceptions. *Innovations in Education and Teaching International*, *41*(1), 59–78. https://doi.org/1470329032000172720

Suñol, J. J., Arbat, G., Pujol, J., Feliu, L., Fraguell, R. M., & Planas-Lladó, A. (2016). Peer and self-assessment applied to oral presentations from a multidisciplinary perspective. *Assessment & Evaluation in Higher Education*, *41*(4), 622–637. https://doi.org/10.1080/02602938.2015.1037720

Sweet, S. A., & Grace-Martin, K. (2011). *Data Analysis with SPSS: A First Course in Applied Statistics*. Pearson College Division.

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education*, *76*(3), 467–481. https://doi.org/10.1007/s10734-017-0220-3

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

Topping, K. (1998). Peer Assessment Between Students in Colleges and Universities. *Review of Educational Research*, *68*(3), 249–276. https://doi.org/10.3102/00346543068003249

Topping, K. J. (2010). Methodological quandaries in studying process and outcomes in peer assessment. *Learning and Instruction*, *20*(4), 339–343. https://doi.org/10.1016/j.learninstruc.2009.08.003

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative Peer Assessment of Academic Writing Between Postgraduate Students. *Assessment & Evaluation in Higher Education*, *25*(2), 149–169. https://doi.org/10.1080/713611428

Torres-Guijarro, S., & Bengoechea, M. (2017). Gender differential in self-assessment: A fact neglected in higher education peer and self-assessment techniques. *Higher Education Research & Development*, *36*(5), 1072–1084. https://doi.org/10.1080/07294360.2016.1264372

Tucker, R. (2014). Sex does not matter: Gender bias and gender differences in peer assessments of contributions to group work. *Assessment & Evaluation in Higher Education*, *39*(3), 293–309. https://doi.org/10.1080/02602938.2013.830282

van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2009). Peer assessment for learning from a social perspective: The influence of interpersonal variables and

structural features. *Educational Research Review*, *4*(1), 41–54. https://doi.org/10.1016/j.edurev.2008.11.002

van Zundert, M., Sluijsmans, D., & van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction*, *20*(4), 270–279. https://doi.org/10.1016/j.learninstruc.2009.08.004

Wen, M. L., & Tsai, C.-C. (2006). University Students' Perceptions of and Attitudes Toward (Online) Peer Assessment. *Higher Education*, *51*(1), 27–44. https://doi.org/10.1007/s10734-004-6375-8

Wen, M. L., & Tsai, C.-C. (2008). Online peer assessment in an inservice science and mathematics teacher education course. *Teaching in Higher Education*, *13*(1), 55–67. https://doi.org/10.1080/13562510701794050

Yurdabakan, I. (2011). The investigation of peer assessment in primary school cooperative learning groups with respect to gender. *Education 3-13*, *39*(2), 153–169. https://doi.org/10.1080/03004270903313608

Zhang, F., Schunn, C., Li, W., & Long, M. (2020). Changes in the reliability and validity of peer assessment across the college years. *Assessment & Evaluation in Higher Education*, *45*(8), 1073–1087. https://doi.org/10.1080/02602938.2020.1724260

Zou, Y., Schunn, C. D., Wang, Y., & Zhang, F. (2018). Student attitudes that predict participation in peer assessment. *Assessment & Evaluation in Higher Education*, *43*(5), 800–811. https://doi.org/10.1080/02602938.2017.1409872

Supplementary Materials

Peer Scoring Rubrics

Criteria	Poor (0 to 3 points)	Average (4 to 7 points)	Excellent (8 to 10 points)
Content	▪ Paragraphs do not introduce the topic or theories needed in the research. ▪ Information discussed has little or nothing to do with the main topic.	▪ Paragraphs introduce the topic or theories needed in the research. ▪ Theories and ideas need to be stated more clearly and be better supported.	▪ Paragraphs introduce the topic or theories needed in the research. ▪ Theories and ideas are stated clearly and supported by examples.
Organization & Sentence Fluency	▪ Many details are not in a logical or expected order. ▪ The RRL is written in a non-scholarly manner.	▪ Details are placed in a logical order. ▪ The RRL is written in a scholarly manner, but improvements can still be made in terms of its fluency.	▪ Details are placed in a logical order. ▪ The RRL is fluently written in a scholarly manner.
Conventions	▪ The article is difficult to understand due to several technical errors (e.g., spelling, grammar, capitalizations) in the paper.	▪ Only a few technical errors (e.g., spelling, grammar, capitalizations) in the paper can be observed.	▪ There were no technical errors (e.g., spelling, grammar, capitalizations) in the paper.
In-Text Citations and References	▪ No APA style in-text citations used throughout document. ▪ Reference page contains no scholarly academic resources, only internet webpages or no reference page.	▪ Some APA style in-text citations used in the document. ▪ Reference page contains some scholarly academic resource and text reference.	▪ APA style in-text citations used throughout document. ▪ Reference page contains more than 10 scholarly academic reference and text reference.

Peer Assessment Design Elements of the Study Based on Adachi et al. (2018)

Cluster	Design Element	Peer Assessment Training Condition	No Peer Assessment Training Condition
Cluster I: decisions concerning the use of peer assessment	(1) Subject area	Experimental psychology course.	
	(2) Intended learning outcomes (for students)	Practice peer assessment prior to the actual peer assessment of experimental research papers.	
	(3) Intended objectives (for staff)	Develop students' peer assessment skills.	
	(4) Timing	First week: 2 hours and 30 mins of training and Eduflow walkthrough Second to third week: Peer assessment activity Fourth week: Debriefing	First week: 30 mins of Eduflow walkthrough Second to third week: Peer assessment activity Fourth week: Debriefing and training
	(5) Assessment type	Three writing samples of varying quality (i.e., poor, average, excellent).	
	(6) Formality and weighting	Students were told they will peer score drafts (formative). Extra credit was given for participation.	
Cluster II: link between peer assessment and other elements in the learning environment	(7) Relationship to other assessments	None. Practice of peer scoring prior to actual peer assessment of experimental research papers.	
	(8) Link to self-assessment	None.	
Cluster III: interaction between peers	(9) Anonymity	Anonymous.	
	(10) Feedback information type	Quantitative (peer scores).	
	(11) Feedback utilisation	Not applicable.	
Cluster IV: composition of assessment groups	(12) Peer configuration	Individual.	
	(13) Peer matching	Matched based on assessor gender and training (See figure 1).	
Cluster V: management of the assessment procedure	(14) Standards use	Department recommended rubric for research outputs.	
	(15) Calibration/task scaffolding	Peer assessment training based on a modified version of Lam (2010).	None.

Cluster VI: contextual elements	(16) Moderation of feedback	None.
	(17) Technology use	Eduflow (web-based peer assessment platform).
	(18) Resources required	Peer assessment training materials, writing rubrics, and writing samples.
	(19) Policy	None.