Validity and reliability of peer-grading in in-service teacher training

Validità e affidabilità del peer-grading nella formazione di insegnanti in servizio

Laura Carlotta Foschi • PhD student, Department of Philosophy, Sociology, Education and Applied Psychology • University of Padova
Graziano Cecchinato • Researcher, Department of Philosophy, Sociology, Education and Applied Psychology • University of Padova

This paper aims to investigate the validity and reliability of peer-grading in in-service teacher training, a field where this practice has been little explored. The study has examined the peer-grading results in an in-service teacher training course involving high school teachers. The validity was measured by using the similarity between peer-grading and trainer-grading scores calculated by the Pearson correlation coefficient; while the reliability was measured by using the agreement of scores given by multiple peer graders calculated by the intraclass correlation coefficient. The empirical findings indicate that in-training teachers provided scores similar to those of the course trainers as well as fairly consistent grading results, thus highlighting that peer-generated grades seems to be valid and reliable in the field of in-service teacher training.

Keywords: Peer-grading, Self-grading, Peergrade, In-service teacher training, Validity, Reliability

Il presente contributo indaga la validità e l'affidabilità del peer-grading nell'ambito della formazione di insegnanti in servizio, contesto in cui questa pratica è poco esplorata. Lo studio ha esaminato l'attività di peer-grading svolta in un percorso di formazione che ha coinvolto insegnanti di scuola superiore. La validità è stata misurata analizzando la somiglianza tra i punteggi attribuiti dagli insegnanti e quelli dei docenti del corso utilizzando il coefficiente di correlazione di Pearson. L'affidabilità è stata misurata analizzando l'accordo tra i punteggi forniti dai diversi insegnanti utilizzando il coefficiente di correlazione intraclasse. I risultati indicano che gli insegnanti hanno fornito punteggi simili a quelli dei docenti del corso e punteggi abbastanza coerenti tra loro, evidenziando come il peer-grading sembri essere valido e affidabile nell'ambito della formazione degli insegnanti in servizio.

Parole chiave: Valutazione tra pari, Autovalutazione, Peergrade, Formazione degli insegnanti in servizio, Validità, Affidabilità

Although the contribution is the result of the joint work of the two authors, Laura Carlotta Foschi has written the paragraphs 3, 4, 5, 6 and the subparagraph 2.3; Graziano Cecchinato has written the paragraph 1 and the subparagraphs 2.1, 2.2.

© Pensa MultiMedia Editore srl – ISSN 2038-9744 (on line) - DOI 10.7346/SIRD-1S2019-P177

Validity and reliability of peer-grading in in-service teacher training

1. Context

Peer-assessment and self-assessment are widely supported by educational research underlining their formative benefits (Falchikov & Goldfinch, 2000; Liu & Carless, 2006). Extensive literature shows their widespread application in the school context, especially in higher education, while other contexts, such as teacher training, and in particular, in in-service teachers training, peer- and self-assessment have been little explored. The present work documents the use of peer-grading within a training activity for in-service teachers with the aim of evaluating the validity and reliability of this practice in this context.

The Authors of this study carried out training and research activities in schools to promote educational innovations inspired by the *flipped classroom* approach (Baker, 2000; Lage, Platt, & Treglia, 2000; Mazur, 1997). The approach combines the use of digital educational resources with active learning practices (Bishop & Verleger, 2013; Cecchinato, 2014; Keengwe, Onchwari, & Oigara, 2014), which are inspired by *Challenge Based Learning* (O'Mahony et al., 2012; Schwartz, Lin, Brophy, & Bransford, 1999).

In summary, we propose to replace the traditional teaching-learning cycle *Lesson – Study – Test*, with a learning-teaching cycle based on three phases: throwing down (*Challenge*), driving (*Reply*) and closing (*Closing*) the "Challenge" (Cecchinato & Papa, 2016). The training course takes approximately one school year to complete with alternating face to face meetings and online activities. At the end of the course, teachers will be able to design and conduct, in their classrooms, Lesson Plans (LPs) according to the proposed approach.

The training is a complex process because requires a significant conceptual change for each of the three phases: the conversion from a deductive to an inductive method in the *Challenge* phase; the transition from lecturing to a more constructivist approach in the *Reply* phase; the shift of emphasis from summative to formative assessment in the *Closing* phase.

To facilitate the design of LPs, teachers are involved in defining the features of a "good" LP with activities such as: the analysis of *exemplars* (Sadler, 1987), concrete examples of "good" LPs; the identification of the LPs corresponding to the proposed approach among others LPs; the sorting of them based on the their quality. These processes are particularly effective in clarifying the objectives and quality levels required, as well as offering a valid standard for comparison with one's own (Orsmond, Merry, & Reiling, 2002). Finally, the LPs designed by the trained teachers are peer-assessed, self-assessed and assessed by the trainers.

2. Literature review

2.1. *Overview of peer-assessment*

Peer-assessment is defined by Topping (1998, p. 250) as «an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status». This is carried out by expressing grades and/or written feedback (Lu & Law, 2012; Strijbos, Narciss, & Dünnebier, 2010).

Peer-assessment is sometimes used simply to reduce teacher' workloads (Li, H. et al., 2016) and in some contexts, for example in MOOCs, it is the only solution that can efficiently provide feedback to students (Piech, Huang, Chen, Do, Ng, & Koller, 2013), but peer-assessment has potential educational benefits that make it a full-fledged practice that can foster learning (Black & Wiliam, 1998; Boud, 2000; Nicol, 2010). According to literature, peer-assessment increases student engagement (Bloxham & West, 2004; Brown & Harris, 2013) and motivation (Topping 2005; Vu & Dall'Alba, 2007), promotes critical thinking (Sims, 1989), develops metacognition (Vickerman, 2009; Wen & Tsai, 2006) and self-regulation (Nicol & Macfarlane-Dick, 2006; Panadero, Tapia, & Huertas, 2012).

Although the research highlights the educational benefits of peer-assessment, there are some concerns about its use, especially in the school context, based on the belief that peers do not have the ability to produce reliable and valid assessments (Li, H. et al., 2016; Liu & Carless, 2006; Magin, 2001). Two relevant meta-analyses summarized the outcomes of the studies conducted in this field taking into consideration the research carried out before 1999 (Falchikov & Goldfinch, 2000) and after 1999 (Li, H. et al., 2016). Although some of the studies were focused on the reliability of peer-assessment, i.e. the consistency between the different peer assessments, both meta-analyses have more specifically investigated the validity, i.e. the consistency between peers and teacher assessments, assuming the latter as exact (gold standard), (Falchikov & Goldfinch, 2000).

The analysed studies differed significantly for contexts, purposes and modalities, but common factors were extrapolated and analysed to establish their influence on the validity of peer-assessment. It was pointed out that the correlation between peer-assessment and teacher's assessment is higher when «(a) the peer assessment is paper-based rather than computer-assisted; (b) the subject area is not medical/clinical; (c) the course is graduate level rather than undergraduate or K-12; (d) individual work instead of group work is assessed; (e) the assessors and assessees are matched at random; (f) the peer assessment is voluntary instead of compulsory; (g) the peer assessment is non-anonymous; (h) peer raters provide both scores and qualitative comments instead of only scores; and (i) peer raters are involved in developing the rating criteria. » (Li, H. et al., 2016, p. 257).

Finally, as for the modalities of carrying out the peer-assessment, it is emphasized that a good organization, preparatory training activities and teacher assistance remarkably contribute to improve its validity.

2.2. *Peer-feedback and peer-grading in the context*

The adoption of peer-assessment in our training course has specific reasons. The main reason is to use potential educational benefits of this practice to promote meaningful learning and the development of specific skills in the proposed teaching approach (Lynch, McNamara, & Seery, 2012; Poon, McNaught, Lam, & Kwan, 2009; Sluijsmans, Brand-Gruwel, & van Merriënboer, 2002). Peer-assessment consists of providing and receiving feedback to and from peers. Both processes have strong educational benefits, but recent literature highlights how students learn more by providing feedback on peer work than receiving feedback from peers (Cho & MacArthur, 2011; Nicol, Thomson, & Breslin, 2014). The act of reviewing (i.e., providing feedback), indeed, activates a reflective process whereby a student compares their work with that of peers and thus realizes how they can improve their own work. Therefore, reviewing not only improves performance, but also the ability to self-regulate learning, as well as fostering a deeper understanding of the object of knowledge (Nicol et al., 2014).

Another reason that led us to use self- and peer-assessment, is promoting their adoption in teaching practices. As highlighted in the literature, experiencing at first hand these assessment practices, especially in training courses, makes teachers familiar with these practices and allows them to acquire the necessary skills for using them productively with their students (Cheng, M. M. H., Cheng, A. Y. N., & Tang, 2010;

Yilmaz, 2017). Through concrete experience of these practices we aim to highlight the educational potentials and the correct strategies in applying them. Self- and peer-assessment have proved to be effective educational strategies in fostering meaningful learning, redefining the role of the student in the evaluation process from passive to active subject. Hence they gather the metacognitive value that characterizes these assessment strategies, the opportunity for students to increase the awareness of their own knowledge, cognitive processes and learning experience, as well as the ability to identify strengths and weaknesses, in the perspective of "learning to learn". Moreover, from a broader educational perspective, allowing students to express complex evaluations about their and others' work, prepares them for making decisions in the complex and unpredictable socio-professional contexts to which they will relate in the future. This supports their ability to think independently, critically and thoughtfully, and their willingness to take responsibility for their actions.

In performing the peer-assessment activity, we adopted some of the factors that research indicates are productive in obtaining a good validity. In particular: the activity was voluntary; the evaluators and assessed were matched at random; the peer-assessment required both scores and qualitative comments; the assessed tasks were individual, not group work; the rating criteria were shared and discussed with the teachers. According to the specifics of our context, however, we decided not to adopt the other factors indicated in the literature. In particular: the peer-assessment was computer-assisted due to the remarkable advantages provided by the adopted digital tool that will be highlighted below; the peer-assessment was anonymous because teachers wouldn't feel comfortable in openly assessing colleagues' work.

In our research we decided to involve the teacher in both component of peer-assessment (feedback and grading) relying on the learning benefits indicated above. Moreover we decide to use the peer-grading activity to carry out an analysis of the validity and the reliability of peer-grading in the context of in-service teacher training.

2.3. *Validity and reliability of peer-grading*

The validity and reliability of peer-grading have been researched primarily in the context of higher education (Cho, Schunn, & Wilson, 2006; Dochy, Segers, & Sluijsmans, 1999; Li, H. et al., 2016; Falchikov & Goldfinch, 2000; Stefani, 1994; Zhang, Johnston, & Kilic, 2008). Validity (i.e., Did the students provide accurate grading?) is commonly measured as the correlation coefficient between mean of peer-generated scores and instructor-generated scores, assuming that instructors can provide accurate and fair grades, while reliability (i.e., Did the students agree with one another? Or, in other words, did the students provide consistent grading?) is usually calculated by the consistency of scores given by multiple peer graders (inter-rater reliability).

The literature on peer-grading highlights that peer-grading appears to be a valid assessment method, indeed many studies have reported a high correlation between peer and instructor grading results. For example, Li and colleagues (2016) conducted a meta-analysis, based on studies since 1999, comparing peer and instructor ratings and found a significant moderately strong correlation between peer and instructor ratings (r = .63). Similar results (r = .69) have been found in a previous meta-analysis, based on studies before 1999, by Falchikov and Goldfinch (2000). The studies considered in these meta-analyses refer mostly to students in graduate or undergraduate courses and included only a small number of studies involving K-12 students. In general, among the available studies, only a few of them consider the context of teacher education and, in particular, most of these refers to pre-service teachers (e.g., Cheng, M. M. H., Cheng, A. Y. N., & Tang, 2010; Lynch, et al., 2012; Sluijsmans et al., 2002; Sluijsmans, Brand-Gruwel, van Merriënboer, & Martens, 2004; Yilmaz, 2017) while the studies related to in-service teachers are very rare (e.g., Wen & Tsai, 2008; Woolhouse, 1999), and in addition, only a few of these analyse validity. For example, regarding studies concerning in-service teachers, the study by Woolhouse (1999) doesn't take into account the issue of validity, while the study by Wen and Tsai (2008) found that two instructors' scores and peers' scores were in low to medium correlation. Concluding, more empirical research concerning validity is needed regarding in-service teachers.

Contrary to the large body of literature on peer-grading validity, there are few studies which take into account peer-grading reliability. The lack of such measurements can undermine the findings regarding peer-grading validity because a valid assessment should also be reliable (Zhang et al., 2008). Concerning the calculation of peer-grading reliability, researchers have used different metrics like Pearson product-moment correlation (e.g., Haaga, 1993), percentage of variance (e.g., Marcoulides & Simkin, 1995), Generalizability Theory (e.g., Yilmaz, 2017; Zhang et al., 2008) and intraclass correlation measuring in some case absolute agreement (e.g., Luo, Robinson, & Park, 2014), while in others consistency (e.g., Cho et al., 2006). Statistical results show peers can produce reliable grades. At the moment, however, there doesn't seem to be any study concerning peer-grading reliability in relation to in-service teachers, and thus empirical research is needed.

In summary, research findings in general support the validity and reliability of peer-grading. However, it is important to bear in mind that such findings are mostly based on the context of traditional college courses with relatively homogenous populations, and thus their use in the in-service teacher training context remains mainly unknown and needs to be further investigated.

3. Research context and questions

3.1. *Peer-grading in Peergrade*

The e-learning environment used for peer-grading was Peergrade (https://www.peergrade.io/), which offers specific advanced features developed by researchers in the field of assessment (e.g., Nicol, 2010). Specifically, Peergrade enhances the formative dimension of the assessment processes with a structured and in-depth dialogue, between all the involved participants, which is intentionally oriented towards improving learning. The peer-assessment activity does not end with the delivery of the outcomes, as usual, but it requires analysis, comparison and review of the assessments, with communicative exchanges among the involved people, structured according to well-defined procedures, and generating a profitable evaluation of the assessment.

Peergrade provides an advanced set of functionalities to make the anonymous review process effective and productive. For instance, the evaluators have to express their own considerations regarding the "usefulness" of the received assessment and feedback by choosing among 5 different choices. Moreover, useful feedback should have the following characteristics: constructiveness, specificity, justification, kindness; the same highlighted by the literature to evaluate the "goodness" of feedback (Hattie, 2012). It is also possible to comment with open text. These considerations expressed by the assessed peers, and shared with the corresponding evaluators, produce a score. Indeed, in Peergrade students are assessed not only on their submitted work ("Submission Score"), but also on the quality of the assessments they provide to peers ("Feedback Score"), to obtain the "Combined Score".

An additional feature is "Flags", which can be used to mark specific assessments or feedback received from peer or teacher of the course. With its activation changes or clarifications to the assessor and the intervention of the teachers of the course are required. The teachers can modify the assessment (either their own or that of the other), or confirm it, but have to explain in both cases their choice. The “Flags” and the teacher’s comments are visible to the evaluator. The evaluator and the assessed can then discuss the “Flags”, as well as express to the teacher their own considerations. The evaluator can also comment with an open text on the ratings and feedback by using the “Comment” button, as well as demonstrate their appreciation by using the “Like” button. As participants are aware that their assessments will be discussed they make the formative dimension of the assessment more concrete and productive. This acts as an incentive to provide more accurate and productive feedback. The evaluation of the assessments and the possibility of flagging requires reflection by the assessed; the subsequent discussion produces a negotiation of meaning that promotes learning (Carless, Salter, Yang, & Lam, 2011; Nicol, 2010; Price, Handley, & Millar, 2011).

The “Feedback Score” is a further element that contributes to empowering the participants in the peer review process. To improve the assessment skills and to make the “Feedback Score” more reliable, Peergrade recommends that each participant evaluate at least 3 peers. This is an element that also contributes to the overall improvement of learning, because receiving feedback from more peers can improve the quality of their work to a greater extent compared with receiving feedback from only one (e.g., Cho & MacArthur, 2010).

Finally, in Peergrade the self-assessment takes place only after evaluating those of the peers. It is important to highlight this aspect because the literature has shown how students learn more by giving feedback on their peer’s work, compared to receiving from them (e.g., Cho & MacArthur, 2011; Nicol et al., 2014). It follows that by first performing the peer-assessment than the self-assessment, the students should be more aware and competent in the latter.

3.2. *Peer-grading details*

The peer-grading assignment examined in this study is the final assignment of an in-service teacher training course, involving 42 Italian high school teachers and lasting about 6 months during the school year 2017-2018, with alternating in-presence meetings and online activities to promote, as stated above, innovation in the learning-teaching cycle. In particular, at the end of the course, teachers are involved in designing LPs according to the proposed approach.

The activity on Peergrade was articulated into three phases: submission, assessment and assessment review. During the submission phase teachers, individually, had to propose their LPs by filling out a document with predefined fields in which to detail, in addition to general information (e.g., school grade, subject), the three phases of the proposed teaching approach: throwing down (*Challenge*), driving (*Reply*) and closing (*Closing*) the challenge. During the assessment phase teachers were asked to anonymously assess, both providing scores and qualitative comments, the LPs of three colleagues randomly assigned by Peergrade, and to self-assess their own LP. Finally, during the assessment review phase, teachers could express their own considerations regarding the "usefulness" of the received assessments.

In the assessment phase, the grading scale used for the scoring of the LPs consisted of 12 criteria specifically prepared by the trainers. In relation to the prepared LPs, the criteria were articulated into 3 sections (*Challenge*: 5 criteria; *Reply*: 4; *Closing*: 3) and aimed to investigate if the LP was characterized by the three main dimensions of the proposed approach: inductive teaching (*Challenge*); active learning (*Reply*); formative assessment (*Closing*). The criteria for the grading scale were scored as 2 (yes), 1 (partially), and 0 (no), depending on weather or not the LP had the features proposed by the criterion, with the sum of the twelve criterion scores as the peer-grading score. As a result, the score for the LP ranged from 0 to 24.

Each teacher was required to grade three LPs submitted by their peers. They were also required to assess their own LPs using the same criteria and provide a self-grading score.

3.3. *Research questions*

To start to deepen the understanding of the validity and reliability of peer-grading in the in-service teacher training context, we wanted to investigate the following two research questions:

- Q1. Does peer-grading provide a valid assessment of teacher LPs in an in-service teacher training course?
- Q2. Does peer-grading provide a reliable assessment of teacher LPs in an in-service teacher training course?

4. Methods

4.1. *Data source*

The main data source in this study is Peergrade's database containing all of the trainer-provided and teacher-generated content regarding peer-grading, including copies of submitted LPs, peer-grading scores and feedback.

The submission_scores_data contains, besides the teacher ID, name, etc., the final peer-grading score. The feedback_data contains more detailed information regarding each LP, such as the feedback-giver ID, username and name; the submission-student ID, username and name; the scores (from the feedback-giver to the submission-receiver) of the 12 criteria; the peer-grading score that each LP received from each feedback-giver, and similarly for self-grading (total score and twelve criterion scores). Additional information such as the submission time, including late submissions, can be found in submission_data.

The trainers required each of 42 teachers to grade three LPs and to self grade their own LP. As a result each teacher received 4 grades, three peer grades and the self grade.

Besides data exported from Peergrade's database, the course trainers also download all LPs and manually[1] graded all of the peer graded LPs (N = 42) using the same grading scale. Each trainer assigned a score for each criterion, with the sum of the twelve criterion scores as the grading score for each trainer and the mean of these two scores as the final trainer-grading score. As a result, a LP has the following features: three individual peer-grading scores, one final peer-grading score using the mean, two individual trainer-grading scores, one final trainer-grading score using the mean, and one self-grading score, as shown in Figure 1.

[1] Even if there is the possibility for trainers to grade LPs directly in Peergrade, we preferred not to do it, in order not to be influenced by the grades assigned by peers.

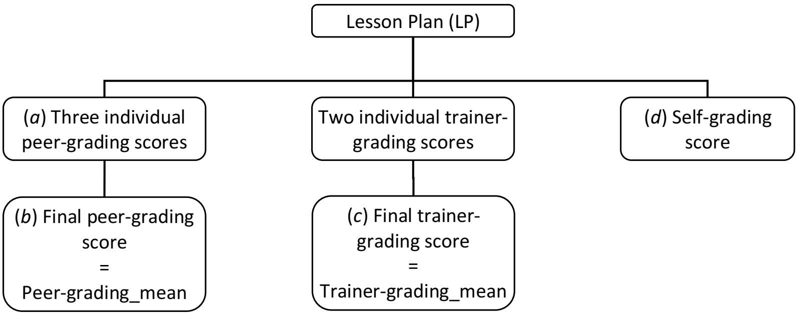
Lesson Plan (LP)
(a) Three individual peer-grading scores
Two individual trainer-grading scores
(d) Self-grading score
(b) Final peer-grading score = Peer-grading_mean
(c) Final trainer-grading score = Trainer-grading_mean

Fig.1: Features of a LP

4.2. *Data analysis*

In order to answer the two research questions proposed in this study, the data analysis focused on the following: calculating the validity of the final peer-grading scores (*b* in Figure 1), calculating the inter-rater reliability of peer-grading scores (*a* in Figure 1).

The validity of peer and self-grading in this study is measured by the similarity between the final peer-grading scores (*b*), the self-grading score (*d* in Figure 1) and the final trainer-grading scores (*c* in Figure 1), assuming trainers ratings to be the "gold standard" (Falchikov & Goldfinch, 2000). The similarity is calculated as the Pearson product-moment correlation coefficient (r). The computation was executed in SPSS by selecting two-tailed Pearson correlation coefficient for bivariate correlation.

The reliability of peer-grading in this study is inter-rater reliability, measured by the agreement among the teacher graders assigned to grade the same LP (*a*). According to Bartko (1966) and Koo and M. Y. Li (2016), because each LP was graded by a different set of three graders randomly selected from the given teacher population, one-way random-effect intraclass correlation coefficient (ICC) was selected as the appropriate statistical model to use in this situation. This model, one-way random-effect, only (unlike other ICC models that can also consider consistency) calculate the grader agreement (absolute agreement), that is based on the exact same scores among graders/recordings and takes into account systematic error among raters/recordings. The calculation of peer-grading reliability was conducted using SPSS by selecting one-way random for ICC scale reliability analysis.

5. Results

This study assumes the course trainers can provide an accurate score for a LP; therefore, the validity of final peer-grading scores (*b*) and self-grading scores (*d*) can be determined by their similarity to the final trainer-grading scores (*c*), measured by the strength of bivariate correlation. As shown in Table 1, there is a statistically significant very strong positive correlation (r = .890) between the trainer-grading_mean (*c*) scores and the peer-grading_mean scores (*b*), indicating teachers can provide similar scores to those assigned by the course trainers.

Compared to the peer-grading scores (*b*), teachers' self-grading (*d*) scores seem to be a less valid assessment of the LP, as the correlation between the self-grading scores (*d*) and the trainer-grading_mean scores (*c*) show a statistically significant moderately strong positive correlation (r = .622). The descriptive analysis also reveals that the mean of self-grading scores (*M* = 19.452) is, albeit slightly, higher than the mean of trainer-grading scores (*M* = 18.047) but lower than the mean of peer-grading scores (*M* = 19.688). In addition this result shows that, in general, teachers tend to give higher scores than trainers both in assessing their own LPs and in assessing their peers.

	Peer-grading_mean	Trainer-grading_mean	Self-grading
Peer-grading_mean	1	.890**	.551**
Trainer-grading_mean		1	.623**
Self-grading			1
** Correlation is significant at the 0.01 level (2-tailed)			

Tab.1: Pearson's correlation coefficient between trainer, peer and self-grading scores (N = 42)

The inter-rater reliability of peer-grading scores (*a*) was calculated using ICC one-way random and the statistical results are presented in Table 2. The Single Measures ICC calculates the inter-rater agreement among the three randomly selected teacher graders when grading the same LP. Regarding the Single Measures, although the obtained ICC value is .514 that, according to Koo and M. Y. Li (2016), indicates a fair reliability, its 95% confidence interval ranges between 0.335 and 0.677, meaning that there is 95% chance that the true ICC value lands on any point in this range, therefore it would be more appropriate to conclude the level of reliability is between poor and fair. This result indicates that peer-grading scores tend to vary among individual teachers and that a single teacher's grading score is not very reliable. Compared to the Single Measures, the Average Measures ICC (.761) shows, according to Koo and M. Y. Li (2016), good reliability, also considering its 95% confidence interval that ranges between good (.602) and excellent (.863). This result suggest that the reliability of peer-grading scores can be enhanced if the mean of the three individual scores is used as an index of measurement.

	Intraclass Correlation	95% Confidence Interval		F Test with True Value 0			
		Lower bound	Upper bound	Value	df1	df2	Sig
Single measures	.514	.335	.677	4.176	41	84	.000
Average measures	.761	.602	.863	4.176	41	84	.000

Tab.2: Intraclass Correlation Coefficient one-way random for peer-grading scores (N = 42).

6. Discussion and conclusion

The empirical findings in this study seem to support the validity of peer-grading in the context of in-service teacher training. The .850 correlation coefficient between final peer-grading scores (named "Submission Score" in Peergrade) and the final trainer-grading score shows that peer-grading in general can provide grading results similar to what a trainer would provide. Even if peer-grading results might never be as accurate as trainer-grading, they seem to yield much higher validity than simply having teachers assess their own LPs, since self-grading scores are found to only moderately correlate with the trainer-grading scores (r = .623). The peer-grading result is consistent with what we have found, using a different data analysis, in another study involving primary and lower secondary school teachers (Foschi, Cecchinato, & Say, 2019) and with what has been found in most of the literature regarding college students and pre-service teachers, as stated above. We can also note that our empirical findings are higher when compared with generally found in literature concerning higher education. This could be due to our specific target and context in which it is probable that the problems that typically can be found regarding peer-assessment (Falchikov, 1995; Kaufman & Schunn, 2011; Liu & Carless, 2006; Magin, 2001) do not occur. Indeed, we deal with adults and specifically teachers who are used to grade and who, therefore, would not have experienced a reluctance to grade others; the anonymity (both of the evaluator and of the assessed) ensured by the online environment would have excluded the possible "friendship bias"; the randomized assignment of evaluators and assessed, as well as the fact that the evaluation has no formal value, would keep the pressure and competition, sometimes induced by peer-grading, under control. With regards to the self-grading result, opposing results can be found in literature, but, consistent with our finding, some studies highlight that self-assessment is generally less accurate than peer assessment (e.g., Stefani, 1994; Dochy et al., 1999). Self-grading was not the main topic of this paper, but, in order to analyse this result in more detail and provide an adequate interpretation, future analyses may not consider the self-grading results as a whole, but categorise them into quartiles, since part of the literature (regarding college students) has highlighted how students with high grades from the teacher tended to assign themselves a lower grade and students with low grades from the teacher assign themselves a higher grade (Boud & Falchikov, 1989; Stefani, 1994).

The empirical findings in this study also seem to support the reliability of peer-grading in the context of in-service teacher training. In this study, LPs were graded by three teachers and the results highlight that the reliability of peer-grading scores can be improved when all three grading scores were averaged to create a composite score, as Average Measures ICC is higher than Single Measures ICC. This suggests that the joint efforts of multiple teacher graders could lead to fairly consistent grading results or, in other words, that in general multiple graders should be used to ensure good reliability. This finding is consistent with what has been found in literature. For example the study of Cho and Colleagues (2006) suggests that the use of multiple graders (four to six) allows for the achievement of very high levels of reliability. In general, it has been found that the number of graders is a key factor for reliability, as reliability is positively correlated with the number of graders.

In conclusion, our work has tried to start to deepen the understanding of validity and reliability of peer-grading in a context where little has been explored, such as in-service teacher training, and the results obtained have highlighted that peer-generated grades seem to be valid and reliable to be used in this context. Finally, it is important to acknowledge that this study involved only a small number of high school teachers, thus these results may not be generalizable or may not reflect all school grades. This study also only focused on peer-grading without examining peer-feedback, as well as having not taken into consideration the impact or the role of the two perspectives of the peer-assessment, namely providing ratings/feedback or receiving ratings/feedback, therefore further studies should investigate each of these issues.

References

Baker W. J. (2000). The "classroom flip": Using web course management tools to become the guide by the side. *Cedarville University: Communication Faculty Publication*, pp. 9-17.

Bartko J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports, 19*(1), pp. 3-11.

Bishop J. L., & Verleger M. A. (2013). The flipped classroom: A survey of the research. In *Proceedings – 120*th *ASEE Annual Conference & Exposition. American Society for Engineering Education.* Atlanta.

Black P., & Wiliam D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), pp. 7-74.

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

Boud D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education, 22*(2), pp. 151-167.

Boud D., & Falchikov N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education, 18*(5), pp. 529-549.

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

Carless D., Salter D., Yang M., & Lam J. (2011). Developing sustainable feedback practices. *Studies in Higher Education, 36*(4), pp. 395-407.

Cecchinato G. (2014). Flipped classroom: Innovare la scuola con le tecnologie digitali. *Italian Journal of Educational Technology*, *22*(1), pp. 11-20.

Cecchinato G., & Papa R. (2016). *Flipped classroom: un nuovo modo di insegnare e apprendere*. Torino: UTET.

Cheng M. M. H., Cheng A. Y. N., & Tang S. Y. F. (2010). Closing the gap between the theory and practice of teaching: Implications for teacher education programmes in Hong Kong. *Journal of Education for Teaching, 36*(1), pp. 91-104.

Cho K., & MacArthur C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction, 20*(4), pp. 328-338.

Cho K. & MacArthur C. (2011). Learning by reviewing. *Journal of Educational Psychology, 103*(1), pp. 73-84.

Cho K., Schunn C. D., & Wilson R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98*(4), pp. 891-901.

Dochy F., Segers M., & Sluijsmans D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education, 24*(3), pp. 331-350.

Falchikov N. (1995). Peer feedback marking: Developing peer assessment. *Programmed Learning, 32*(2), pp. 175-187.

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), pp. 287-322.

Foschi L.C., Cecchinato G., & Say F. (2019). Quis iudicabit ipsos iudices? Analisi dello sviluppo di competenze in un percorso di formazione per insegnanti tramite la valutazione tra pari e l'autovalutazione. *Italian Journal of Educational Technology*, *27*(1), pp. 49-64.

Haaga D. A. F. (1993). Peer review of term papers in graduate psychology courses. *Teaching of Psychology, 20*(1), pp. 28-32.

Hattie J. (2012). *Visible learning for teachers: Maximizing impact on learning*. New York: Routledge.

Kaufman J. H., & Schunn C. D. (2011). Students' perceptions about peer assessment for writing: Their origin and impact on revision work. *Instructional Science, 39*(3), pp. 387-406.

Keengwe J., Onchwari G., & Oigara J. (2014). *Promoting active learning through the flipped classroom model*. Hershey: IGI Global.

Koo T. K., & Li M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of chiropractic medicine*, *15*(2), pp. 155-63.

Lage M. J., Platt G. J., & Treglia M. (2000). Inverting the classroom: A gateway to creating an inclusive learning environment. *Journal of Economic Education, 31*(1), pp. 30-43.

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

Liu N., & Carless D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education, 11*(3), pp. 279-290.

Lu J., & Law N. (2012). Online peer assessment: effects of cognitive and affective feedback. *Instructional Science*, *40*(2), pp. 257-275.

Luo H., Robinson A., & Park J-Y. (2014). Peer grading in a MOOC: Reliability, validity, and perceived effects. *Online Learning, 18*(2).

Lynch R., McNamara P. M., & Seery N. (2012). Promoting deep learning in a teacher education programme through self- and peer-assessment and feedback. *European Journal of Teacher Education, 35*(2), pp. 179-197.

Magin D. (2001). Reciprocity as a source of bias in multiple peer assessment of group work. *Studies in Higher Education, 26*(1), pp. 52-63.

Marcoulides G. A., & Simkin M. G. (1995). The consistency of peer review in student writing projects. *Journal of Education for Business, 70*(4), pp. 220-223.

Mazur E. (1997). *Peer instruction: A user's manual.* Upper Saddle River: Prentice Hall.

Nicol D. (2010). From monologue to dialogue: Improving written feedback processes in mass higher education. *Assessment and Evaluation in Higher Education, 35*(5), pp. 501-517.

Nicol, D. & MacFarlane-Dick D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), pp. 199-218.

Nicol D., Thomson A., & Breslin C. (2014). Rethinking feedback practices in higher education: A peer review perspective. *Assessment and Evaluation in Higher Education, 39*(1), pp. 102-122.

O'Mahony T. K., Vye N. J., Bransford J. D., Sanders E. A., Stevens R., Stephens ... Soleiman M. K. (2012). A comparison of lecture-based and challenge-based learning in a workplace setting: Course designs, patterns of interactivity, and learning outcomes. *Journal of the Learning Sciences, 21*(1), pp. 182-206.

Orsmond P., Merry S., & Reiling K. (2002). The use of exemplars and formative feedback when using student derived marking criteria in peer and self-assessment. *Assessment and Evaluation in Higher Education, 27*(4), pp. 309-323.

Panadero E., Tapia J. A., & Huertas J. A. (2012). Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education. *Learning and Individual Differences, 22*(6), pp. 806-813.

Piech C., Huang J., Chen Z., Do C., Ng A., & Koller D. (2013). Tuned models of peer assessment in MOOCs. In *Proceedings of the 6th International Conference on Educational Data Mining*, Memphis.

Poon W., McNaught C., Lam P., & Kwan H. S. (2009). Improving assessment methods in university science education with negotiated self and peer assessment. *Assessment in Education: Principles, Policy & Practice, 16*(3), pp. 331-346.

Price M., Handley K., & Millar J. (2011). Feedback: Focusing attention on engagement. *Studies in Higher Education, 36*(8), pp. 879-896.

Sadler D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education, 13*(2), pp. 191-209.

Schwartz D. L., Lin X., Brophy S., & Bransford J. D. (1999). *Toward the development of flexibly adaptive instructional designs*. Hillsdale: Erlbaum.

Sims G. K. (1989). Student peer review in the classroom: A teaching and grading tool. *Journal of Agronomic Education, 8*(2), pp. 105-108.

Sluijsmans D. M. A., Brand-Gruwel S., & van Merriënboer J. J. G. (2002). Peer assessment training in teacher education: Effects on performance and perceptions. *Assessment and Evaluation in Higher Education, 27*(5), pp. 443-454.

Sluijsmans D. M. A., Brand-Gruwel S., van Merriënboer J. J. & Martens R. L. (2004). Training teachers in peer-assessment skills: Effects on performance and perceptions. *Innovations in Education and Teaching International, 41*(1), pp. 59-78.

Stefani L. A. J. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education, 19*(1), pp. 69-75.

Strijbos J. W., Narciss S., & Dünnebier K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: are they critical for feedback perceptions and efficiency? *Learning and instruction*, *20*(4), pp. 291-303.

Topping K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), pp. 249-276.

Topping K. J. (2005). Trends in peer learning. *Educational Psychology, 25*(6), pp. 631-645.

Vickerman P. (2009). Student perspectives on formative peer assessment: An attempt to deepen learning? *Assessment & Evaluation in Higher Education, 34*(2), pp. 221-230.

Vu, T. T. & Dall'Alba G. (2007). Students' experience of peer assessment in a professional course. *Assessment & Evaluation in Higher Education, 32*(5), pp. 541-556.

Wen M. L., & Tsai C. (2006). University students' perceptions of and attitudes toward (online) peer assessment. *Higher Education, 51*(1), pp. 27-44.

Wen M. L., & Tsai C. (2008). Online peer assessment in an inservice science and mathematics teacher education course. *Teaching in Higher Education, 13*(1), pp. 55-67.

Woolhouse M. (1999). Peer assessment: The participants' perception of two activities on a further education teacher education course. *Journal of further and Higher Education, 23*(2), pp. 211-219.

Yilmaz F. M. (2017). Reliability of scores obtained from self-, peer-, and teacher-assessments on teaching materials prepared by teacher candidates. *Educational Sciences: Theory & Practice, 17*(2), pp. 395-409.

Zhang B., Johnston L., & Kilic G. B. (2008). Assessing the reliability of self and peer rating in student group work. *Assessment & Evaluation in Higher Education, 33*(3), pp. 329-340.