ISSN: 2036-5330 (stampa); 2974-9174 (online)
DOI: 10.32076/RA16102

Unveiling peer assessment in teacher continuous professional development: perceptions and lessons from an experience with in-service teachers

Disvelare la valutazione tra pari nello sviluppo professionale continuo degli insegnanti: percezioni e indicazioni da un'esperienza con insegnanti in servizio

Laura Carlotta Foschi[1]

Abstract

This paper explores a peer assessment (PA) experience within teacher continuous professional development (CPD). The study investigates teachers' perceptions of giving (GA) and receiving assessments (RA) on the learning challenges they designed. Teachers participated in a survey of closed and open questions, analysed quantitatively and qualitatively. Results indicate that teachers found PA to be a valuable learning experience. Most reported learning from both GA and RA. Nevertheless, they perceived GA as more beneficial than RA. GA allowed teachers to reflect on their challenges, compare them with those of peers, understand different instructional strategies, and identify the positives and negatives of both peers' work and their own. RA helped teachers identify strengths and weaknesses in their challenges and offered improvement perspectives. Teachers also acknowledged peers' ability to give both quantitative scores and qualitative comments. This study contributes insights into PA within teacher CPD, highlighting its potential benefits and suggesting implementation directions.

Keywords: Peer assessment; In-service teachers; Continuous professional development; Teacher perceptions; Teacher training.

Sintesi

Il contributo esplora un'esperienza di valutazione tra pari (PA) nell'ambito dello sviluppo professionale continuo (CPD) degli insegnanti. Lo studio indaga le loro percezioni nel dare (GA) e ricevere (RA) valutazioni su sfide di apprendimento da loro progettate tramite un'indagine con domande aperte e chiuse, analizzate quantitativamente e qualitativamente. I risultati indicano che gli insegnanti hanno ritenuto il PA un'esperienza di apprendimento preziosa. La maggior parte ha riportato di aver imparato sia dal GA che dal RA. GA, tuttavia, è stato percepito come più vantaggioso di RA. GA ha permesso agli insegnanti di riflettere sulle proprie sfide, confrontarle con quelle dei pari, comprendere le diverse strategie didattiche e identificare gli aspetti positivi e negativi del lavoro dei pari e del proprio. RA li ha aiutati a identificare i punti di forza e di debolezza delle proprie sfide e ha offerto prospettive di miglioramento. Gli insegnanti hanno inoltre ritenuto i pari in grado di fornire sia punteggi quantitativi che commenti qualitativi. Questo studio fornisce spunti di riflessione sul PA nel CPD degli insegnanti, evidenziandone i potenziali benefici e suggerendo indicazioni per l'implementazione.

Parola chiave: Valutazione tra pari; Insegnanti in servizio; Sviluppo professionale continuo; Percezioni degli insegnanti; Formazione degli insegnanti.

1. Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata (FISPPA), Università degli Studi di Padova; lauracarlotta.foschi@unipd.it.

1. Introduction

Peer assessment (PA) is widely supported by educational research emphasising its benefits (Li *et al.*, 2020; Topping, 2018; 2022; Zhan *et al.*, 2023). Extensive literature shows its widespread application in the school context, especially in higher education, while in other contexts, such as school teacher education/training, particularly Continuous Professional Development (CPD) of in-service teachers, PA have been little explored (Topping, 2023). The present paper documents and examines the use of PA within a CPD activity. The aim is to investigate teachers' perceptions of the PA experience, particularly of giving and receiving assessments to and from peers.

This aim arises from the acknowledgement that the use of PA in the context of school teacher education/training remains largely unknown and needs further investigation. There are at least two areas of interest. The first area concerns the validity and reliability of PA, while the second area concerns teachers' perceptions of PA. As mentioned, PA studies in the context of teachers are limited. Moreover, there is a gap in the research literature between studies conducted with pre-service teachers and those with in-service teachers, where the former significantly exceeds the latter. In the context of pre-service and in-service school teacher education and development, PA validity and reliability have already been relatively investigated (e.g., Foschi & Cecchinato, 2019a, 2019b; Foschi *et al.,* 2019; Wen & Tsai, 2008; Yilmaz, 2017) and research evidence on PA effectiveness has also been examined (e.g., Foschi *et al.,* 2019; Grion & Restiglian, 2021; Lynch *et al.*, 2012; Restiglian, 2019; Sluijsmans *et al.*, 2002; Topping, 2023). This paper, therefore, aims to investigate the second area, i.e., the perceptions of teachers, specifically in-service teachers, regarding the PA experience, particularly of giving and receiving assessments to and from peers.

2. Definition and forms of PA

A composite definition of PA is provided by Topping, who defines it as "an arrangement for learners to consider and specify the level, value, or quality of a product or performance of other equal-status learners, then learn further by giving elaborated feedback to and discussing their appraisals with those who were assessed to achieve a negotiated agreed outcome" (Topping, 2018, p. 1). The researcher expressly chose this definition as it is consistent with the perspective and the study addressed in the present paper. For this reason and because there are different kinds of PA, the fundamental aspects of PA, as understood here, are emphasised below.

Firstly, PA is understood here in terms of giving feedback and/or grades on peers' work and receiving feedback and/or grades from peers on one's own work. In other words, PA here is a reciprocal process and engages learners in both the roles of assessors and assessees (Topping, 1998; 2018; 2022; 2023).

Secondly, PA in the present study is both quantitative and qualitative. PA can take place through the giving of grades or marks

(quantitative), feedback (qualitative), or both (Topping, 2018; 2022).

Thirdly, closely related to the previous aspect, PA in the present study is both formative and summative. In short, when formative, PA helps students identify the strengths and weaknesses of their work and opportunities for improvement (Topping, 2009; 2018; 2022). When summative, PA shows the students how good or bad their work is without opportunities for improvement.

Fourthly, PA, as understood here, does not conceive of assessment as a one-way process but as an improvement-oriented comparison and dialogue between the assesses and the assessors. It is crucial that PA takes the form of a reciprocal dialogue, i.e., that the assessee can discuss the feedback/mark with the assessor (Nicol, 2010; Topping, 2018; 2022).

There would be further aspects to be analysed, for instance, the type of "work" subject to PA. In the present study, this is a product, precisely a written learning unit document. See subparagraph "4.3 The PA experience" for further choices made in the present context and Topping (2018) for a more detailed discussion of PA forms.

3. Giving and receiving assessments

The literature on studies conducted with students has recognised that both *giving* and *receiving* assessments are indispensable foundations in the practice of PA and documented the benefits for both parties. These two sides have interconnected and complementary purposes for facilitating learner learning (Topping, 1998; 2018; 2022; 2023). However, the benefits of PA differ depending on whether the learner is giving or receiving assessments (Felisatti *et al.*, 2020; Grion & Tino, 2018; Nicol *et al.*, 2014).

By receiving assessments (RA) from peers, students gain insights into their work strengths and weaknesses, which can guide their learning process (Nicol *et al.*, 2014; Topping, 1998). RA allows students to view their work from diverse perspectives - also realising the possible different interpretations of it - and to reflect on aspects they may have missed (Cho & MacArthur, 2010; Nicol *et al.*, 2014; Topping, 1998). The student learning processes typical of RA are reflection, feedback evaluation, awareness of effective feedback, and self-assessment (Li & Grion, 2019). RA can also help students develop their ability to accept and process constructive criticism (Topping, 1998).

By giving assessments (GA) to peers, students gain a deeper understanding of the subject matter (Davies, 2000) and are inspired by peers' work (Li & Steckelberg, 2006). When students assess their peers' work, they actively analyse and internalise the assessment criteria, leading to a more comprehensive understanding of the topic and the development of a deeper understanding of assessment criteria and standards (Li & Grion, 2019; Nicol *et al.*, 2014). The student learning processes typical of GA are active interpretation and application of assessment criteria, critical thinking and making critical judgements, feedback construction, reflection, comparison, and learning transfer (Li & Grion, 2019; Nicol *et al.*,

2014). In addition, GA helps students develop evaluative skills (Nicol *et al.*, 2014).

4. Study

4.1. Study objectives

This exploratory study aimed to provide insight into teachers' perceptions of the PA experience, particularly of GA and RA to and from peers. Specifically, the study aimed to explore teachers' subjective experiences of the PA activity, whether teachers perceive greater learning benefits from acting as assessors (i.e., GA to peers' work) or as assessees (i.e., RA from peers on their own work), or if they consider both roles equally contributory to their learning, why and how teachers believe they learn and benefit from GA or RA, or both.

4.2. The context

The PA experience examined in this study is the first assignment of an in-service teacher training course. The course involved 34 Italian primary and lower secondary school teachers and lasted approximately four months in the school year 2022-2023. It was a blended course that alternated face-to-face meetings with online activities to promote innovation in the learning-teaching cycle. In short, a learning-teaching cycle based on three phases - throwing down (Challenge), driving (Reply), and closing (Closing) the "Challenge" - is what the training course proposes to replace the traditional teaching-learning cycle based on Lecture - Study - Test (Cecchinato & Papa, 2016; Cecchinato *et al.,* 2019). Each month, teachers were involved in a face-to-face meeting and online activities designed to enable them to gain first-hand experience and thus design one of the three phases of the Learning Unit (LU) according to the proposed cycle.

To primarily support the design of LUs, but also their assessment, teachers were involved in multiple activities and were provided with multiple resources during both face-to-face meetings and online modules. These, for example, were the identification of LUs corresponding to the proposed learning-teaching approach among other LUs; the use of exemplars (Sadler, 1987; i.e., examples of LUs of different quality), their analysis and sorting according to their quality; exemplars assessed using the same criteria and questions that teachers would have used during the PA activity; concrete examples of "good" LUs for different subjects and school grades. See Foschi (2021; 2022) for a more detailed discussion of similar activity and the reasons underlying these choices.

4.3. The PA experience

The PA activity was carried out via an online platform (*www.eduflow.com*) and consisted of four phases: submission, peer assessment, self-assessment, and reflection. During the submission phase, teachers had to design their own Challenge by filling in a document and submitting it on the platform. During the PA phase, teachers anonymously had to assess the Challenges of three randomly assigned peers, providing both

scores/ratings and comments. Teachers had to self-assess their own Challenge using the same parameters during the self-assessment phase. Finally, during the reflection phase, teachers had to express their thoughts on the peer assessments their Challenge received.

The PA phase, as well as the self-assessment phase, included both a *quantitative* and a *qualitative* part. The former consists of two criteria and one closed-ended question. The latter consists of three open-ended questions. The two criteria aimed to investigate the two key characteristics of the Challenge, i.e., whether it was challenging and whether its approach was inductive. Teachers expressed their degree of disagreement/agreement with each criterion on a four-point scale. The closed-ended question asked teachers to judge the overall quality of the Challenge by choosing between poor, fair, good, and excellent. Of the three open-ended questions, the first asked teachers to indicate the reasons for their ratings on the two criteria. The second asked them to indicate what they liked about the Challenge and why. The third asked what could be improved in the Challenge and why. The prompts proposed for the PA phase (i.e., criteria and questions) were available in advance during the submission phase. They were also available before the teachers designed their Challenge. In fact, as mentioned in the previous section, the Challenge *exemplars* proposed to the teachers were also assessed (justifying the choices made) using the same prompts as the PA activity in which the teachers would take part.

The reflection phase consisted of two open-ended and one closed-ended question. The former asked teachers to express what they had learned from the feedback their Challenge had received from peers and how they planned to use it to improve their Challenge. The latter asked teachers to indicate how useful they found the feedback. Teachers had to choose between not at all useful, not very useful, somewhat useful, very useful, and extremely useful. In judging the usefulness of the feedback, teachers were invited to consider that useful feedback should have, in short, the following characteristics: clarity, specificity, justification, constructiveness, balance between positive aspects and areas of improvement, and kindness.

4.4. Data collection and participants

The data collection on the PA experience was carried out through an online survey. Teachers completed the anonymous survey after the PA activity had ended. The survey comprised seventeen questions and sought information relevant to the study objectives. Fourteen were closed-ended; half asked teachers to select an answer, and half asked them to rate their agreement using a five-point scale. Three questions were open-ended and required teachers to make written comments on a previously closed answer.

Of the 34 teachers in the training course, 32 participated in the PA experience analysed in the present study, and 28 responded to the survey described above. Teacher age ranged from 34 to 65 (M = 49.2; SD = 7.93), and years of teaching ranged from 3 to 40 (M = 10.8; SD = 9.75). The teachers were

75% in primary school and 25% in lower secondary school. Primary school teachers' subjects were Mathematics, Science, and Technology (12); Italian, History, Geography, and English (8); one was a special needs teacher. The second or third group teachers also taught Art, Civic Education, English, Music, and Physical Education. Secondary school teachers' subjects were: Technology (2); Foreign Language (2); Music (1); Mathematics and Science (1); Italian, History, and Geography (1).

4.5. Data analysis

Both quantitative and qualitative data analyses were used to achieve the study objectives.

Chi-square test. The researcher used the one-sample chi-square test (Pearson, 1900) to examine the proportions of teachers' responses in the case of categorical data. When test assumptions were unmet, the exact significance was calculated instead of the asymptotic significance. In addition, post-hoc tests with Bonferroni corrected alpha level were performed.

One-sample/Paired-samples Wilcoxon test. Due to the limited sample size and the non-normal variable distributions (*p*-values of Shapiro-Wilk < .05), the researcher used nonparametric tests. One-sample Wilcoxon tests were performed to test the differences between the median calculated for the teachers' responses and the scale midpoint. Paired-samples Wilcoxon tests were used to examine the within-teacher differences. Since *N* was greater than 16 (Gibbons, 1993), it was possible to obtain an approximation of the probability distribution of the *T* statistic to the normal one and calculate the *Z*-scores. The critical alpha level was .05. The effect size (ES), *r*, was calculated by dividing the observed value of *Z* by the square root of *N* (Pallant, 2007) and interpreted using Cohen's guidelines (Cohen, 1988).

Thematic analysis. The researcher conducted the thematic analysis using a circular and recursive process similar to that of Braun and Clarke (2006). At first, she familiarised herself with the data by reading and re-reading the teachers' answers in search of patterns and meanings, as well as generating and noting initial ideas about what the data contained and what was interesting in it. She then created initial codes using the teachers' language (*in vivo codes*). The researcher subsequently interpreted and grouped these codes into potential overarching themes. Finally, she reviewed the themes to find patterns and to examine the most recurrent themes. The researcher developed "semantic" themes (ibid.) as her main interest was what the teachers had written explicitly, not identifying latent meanings. Moreover, answers to free-text survey questions often generate data that are not "rich enough" to support deeper forms of analysis (LaDonna *et al.*, 2018). When the researcher finished analysing the data, she described the results in summary terms. As highlighted by Hammersley and Atkinson (1983, p. 178), the concepts used were "[...] a combination of those (concepts, *ed.*) derived from the data themselves and those inferred by the researcher". Indeed,

as Braun and Clarke (2019, p. 591) argue, "For us, the final analysis is the product of deep and prolonged data immersion, thoughtfulness and reflection, something that is active and generative."

5. Results

All teachers indicated that they both assessed the peer Challenges and viewed the assessments received from peers on their Challenge. Their responses to the survey are then analysed below. The results are analysed concerning teachers' perceptions of PA, GA, RA, GA and/or RA.

5.1. PA

From which aspect of PA do you think you learned the most? Why? Concerning the question "From which aspect of PA do you think you learned the most?", 57.1% of the teachers answered "from both: from assessing peer Challenges and from receiving their assessments on my Challenge", 28.6% "from assessing peer Challenges", 10.7% "from receiving peer assessments on my Challenge" and only one teacher (3.6%) "from neither"[2]. The result of the one-sample chi-square test (X^2 = 19.1, *df* = 3, p[3] < .001) suggests that there is no equidistribution in teachers' responses. Subsequent post-hoc tests (Tab. 2), performed with Bonferroni correction, showed that the response *Both* received significantly more preferences than expected. In contrast, the other responses did not differ from what was expected. Moreover, post-hoc pairwise comparison tests (Tab. 3) with the Bonferroni corrected alpha level showed a difference in the teacher responses' proportions in the following comparisons: *Both* vs *Neither* and *Both* vs *RA*. In both cases, the first proportion was greater than the second.

The reasons (i.e., the answers to the question "Why?") given by the teachers who indicated having learned from both aspects refer to a multiplicity of aspects. These can be summarised in the words of one teacher: "They allowed me a 360 degrees comparison". Teachers pointed out that both aspects of the PA allowed them to reflect on their own work, how they had designed their own Challenge, and to better understand how to progress. At the same time, teachers learned something from both peers' work and their suggestions. They also reconsidered some of their initial positions and reviewed their own Challenges. Moreover, teachers reported that by giving assessments, they recognised the strengths and weaknesses of peers' Challenges and, by receiving assessments, learned those of their own. They also indicated that by assessing peers' Challenges, they could confront ways of working that differed from their own. In contrast, by receiving peer assessments, they questioned some aspects of their own Challenges.

On the other hand, the primary motivation given by teachers who indicated that they learned the most from assessing peer Challenges is that they could see other and different ways of throwing down

2. The responses were abbreviated as follows: Both, GA, RA, Neither.
3. Unless otherwise specified, *p* refers to asymptotic significance.

the Challenge and thus had a variety of ideas and a wider range of Challenges. Teachers also indicated that, by comparing their own Challenge with that of peers, they understood the quality of their own (e.g., "[...] I realised that, on the whole, my Challenge was structured fairly well"). Teachers then reported figuring out what might be unstimulating and unchallenging for students. Alongside these motivations, some teachers also reported that they could not claim to have learned from the assessments from peers as these were either confirmatory and thus did not highlight anything that could be improved, or "unfortunately" were not constructive and thus did not allow for productive comparison.

Teachers who otherwise reported learning the most from receiving peer assessments on their Challenge emphasised the usefulness of having different points of view on their work, the possibility of getting feedback from peers on the effectiveness of their Challenge and whether it was adequately described, as well as the enrichment generated by criticism - as long as it was constructive - in recognising the limitations of their work and fixing them.

Finally, the teacher who indicated "from neither" reported that because she was unsure whether her own work met the characteristics of the Challenge, she was neither confident in her ability to assess peers' work nor to grasp peer reviews.

Did you modify (or do you intend to modify) your Challenge as a result of the PA activity? Considering the above question, 60.7% of the teachers answered "no", 17.9% "yes, as a result of the assessments given and received", 14.3% "yes, as a result of the assessments received from my peers", and 7.1% answered "yes, as a result of the assessments given to my peers". Subsequent post-hoc tests, conducted after one-sample chi-square test (X^2 = 19.71, df = 3, p < .001), showed that the response "no" was statistically the most preferred.

In general, I believe that the PA process (as a whole) is a valuable learning experience. Considering the teachers' responses to item E (Tab. 1), the one-sample Wilcoxon test showed that the responses' median was statistically significantly higher than the scale midpoint (Z = 3.144, p = .002) with an ES of .59.

	Percentages (N = 28)					Descriptive statistics				
Items	1	2	3	4	5	M	SD	25° (Q1)	50° (Me)	75° (Q3)
(A) I would have liked to have assessed more than 3 Challenges.	0%	32.1%	39.3%	17.9%	10.7%	3.07	.98	2	3	4
(B) I would have liked to have received more than 3 assessments of my Challenge.	0%	10.7%	35.7%	32.1%	21.4%	3.64	.95	3	4	4
(C) My knowledge and understanding of the topic (i.e., throwing down the Challenge) improved as a result of my role as ASSESSOR and of having GIVEN assessments.	0%	17.9%	42.9%	35.7%	3.6%	3.25	.80	3	3	4
(D) My knowledge and understanding of the topic (i.e., throwing down the Challenge) improved as a result of my role as ASSESSEE and of having RECEIVED assessments.	10.7%	21.4%	46.4%	17.9%	3.6%	2.82	.98	2	3	3
(E) In general, I believe that the PA process (as a whole) is a valuable learning experience.	0%	7.1%	39.3%	35.7%	17.9%	3.64	.87	3	3	4
(F) In general, I think my peers are able to give qualitative feedback/comments on my work.	0%	7.1 %	53.6%	35.7%	3.6%	3.36	.68	3	3	4
(G) In general, I think my peers are able to give quantitative scores on my work using a scale/rubric.	0%	7.1%	64.3%	25.0%	3.6%	3.25	.65	3	3	4

Response scale: 1 (Strongly disagree) to 5 (Strongly agree).

Tab. 1 - *Descriptive statistics of item responses.*

5.2. Giving assessments

While assessing the peers' Challenges, did you happen to think about your own Challenge? If yes, in what way? Almost all the teachers (92.9%) indicated that, while assessing their peers' Challenges, they happened to think about their own Challenge, while only two teachers answered no. The "In what way?" answers revealed different aspects and modalities. Teachers indicated that they compared how peers designed their Challenge with how they designed their own. Through this comparison, they looked for possible similarities and differences and analysed the positive and negative aspects. They reflected on their own work by asking themselves whether their Challenge was good and/or if it had something to improve or could be done differently, as well as if it had - or not - stimulating and motivating elements. They thought about and grasped what they could do differently in their Challenge and what they could modify and improve. They realised that, overall, they had done fair work.

Did assessing peers' Challenges make you think about how you could improve your Challenge? When asked the above question, 96.4% of the teachers answered yes, while only one answered no.

I would have liked to have assessed more than 3 Challenges. Finally, considering the teachers' responses to item A (Tab. 1), the one-sample Wilcoxon test showed that the responses' median was not statistically significantly different than the scale midpoint ($p > .05$).

5.3. Receiving assessments

Did the peer assessments received make you reflect on how you could improve your Challenge? If yes, how? When asked "Did the peer assessments received make you reflect on how you could improve your Challenge?", 60.7% of the teachers answered yes, while 39.3% answered no. The "In what way?" answers revealed different aspects and modalities. Above all, teachers indicated that peers offered them useful and interesting hints and suggestions, which they reflected on and considered to think about how they could improve their Challenge. Other teachers stated that the peer reviews helped them revise certain aspects of their Challenge. These teachers also indicated that, due to the peer feedback, they modified aspects or passages of their Challenge (e.g., peer assessments "allowed me to apply different ways of interacting with the group and the individual") or specified them better. One teacher instead realised that her Challenge might have been too complex.

I received ... assessments of my Challenge. Select the option that corresponds to the number of peer assessments you received. Almost all teachers (92.9%) reported receiving three peer assessments on their Challenge, while the remaining two (7.1%) reported receiving two.

I would have liked to have received more than 3 assessments of my Challenge. Tab. 1 shows the teachers' response percentages to item B. Looking specifically at the two teachers who received two assessments, one

indicated "3" and one "4". The one-sample Wilcoxon test showed that the responses' median was statistically significantly higher than the scale midpoint (Z = 2.99, p = .003) with an ES of .57.

In general, I think my peers are able to give qualitative feedback/comments on my work. & *In general, I think my peers are able to give quantitative scores on my work using a scale/rubric.* Considering the teachers' responses to items F and G (Tab. 1), one-sample Wilcoxon tests showed that the responses' medians were statistically significantly higher than the scale midpoint (respectively: Z = 2.5, p = .012; Z = 1.941, p = .026[4]) with ESs of .47 and .37 respectively. The paired-samples Wilcoxon test showed no statistically significant difference between the responses' medians (p > .05).

5.4. Giving and/or receiving assessments

Did assessing peers' Challenges make you think about how you could improve your Challenge? X *Did the peer assessments received make you reflect on how you could improve your Challenge?* From the intersection (i.e., contingency table) of the responses to the following two questions, "Did assessing peers' Challenges make you think about how you could improve your Challenge?" and "Did the peer assessments received make you reflect on how you could improve your Challenge?"[5], emerged four categories (Tab. 4): *No GA-No RA* (0%), *No GA-Yes RA* (3.6%), *Yes GA-No RA* (39.3%), and *Yes GA-Yes RA* (57.1%). The one-sample chi-square test result (X^2 = 26, df = 3, p < .001) suggests no equidistribution in these four response categories. Subsequent post-hoc tests (Tab. 5), performed with Bonferroni correction, showed that the category *Yes GA-Yes RA* received significantly more preferences than expected, and *No GA-No RA* received fewer preferences than expected. In contrast, the other responses did not differ from what was expected. Moreover, post-hoc pairwise comparisons tests (Tab. 6) with the Bonferroni corrected alpha level showed that there was a difference in the teacher responses' proportions in the following comparisons: Yes GA-No RA vs No GA-Yes RA, Yes GA-Yes RA vs No GA-Yes RA, Yes GA-No RA vs No GA-No RA, and Yes GA-Yes RA vs No GA-No RA. In all cases, the first proportion was greater than the second.

My knowledge and understanding of the topic (i.e., throwing down the Challenge) improved as a result of my role as ASSESSOR and of having GIVEN assessments. & *My knowledge and understanding of the topic (i.e., throwing down the Challenge) improved as a result of my role as ASSESSEE and of having RECEIVED assessments.* Considering the teachers' responses to items C and D (Tab. 1), one-sample Wilcoxon tests showed that the responses' medians were not statistically significantly different from the scale midpoint (ps > .05). The paired-samples Wilcoxon test showed that there was a statistically significant difference between the responses' medians (Z = 1.713,

4. Right-sided one-tailed test. If not specified, it is a two-tailed test.
5. In this paragraph, GA will be used to name the first item and RA the second.

p = .043[6]) with an ES of .32. Specifically, the responses' median to the first item (i.e., concerning the assessor role) was higher than that of the responses to the second (i.e., concerning the assessee role).

6. Discussion

6.1. Learning from PA: Close-ended responses

Teachers generally believe the PA process is a very valuable learning experience (item E: large ES). This belief is corroborated by the responses to the question, "From which aspect of PA do you think you learned the most?". They show that almost all teachers believed they learned from some aspect of the PA activity. Specifically, most teachers reported that they learned more from both aspects of PA: GA and RA. Similar results emerged when cross-tabulating the teachers' responses to questions about whether GA and RA made them think about how they could improve their Challenge. Indeed, most teachers responded yes to both questions. Moreover, a very interesting result is the statistically significant difference in the proportions of teachers' responses in the comparison of *Yes GA-No RA* vs *No GA-Yes RA*. This result suggests that the PA aspect that made teachers think more about how they could improve their own work was GA than RA. Ultimately, these results indicate that teachers believe that there was value, both in terms of learning and improving their work, in both PA processes. Nonetheless, GA is particularly beneficial for teachers, both in general and compared to RA. In addition, teachers felt that their knowledge and understanding of the topic (i.e., throwing down the Challenge) improved more because of their role as assessor and of GA than because of their role as assessee and of RA (comparison between items C and D: medium ES). Overall, these results overlap with those found in studies involving students, especially the study of Nicol *et al.* (2014). Similar findings, particularly concerning GA being more beneficial than RA, can also be found in Carlsson Hauff and Nilsson (2021), Cho and Cho (2011), and Grion and Tino (2018).

6.2. Learning from PA: Open-ended responses

The reasons why teachers reported learning from both aspects of PA referred to different aspects. These can be summarised in that both GA and RA allowed them an all-around comparison, learning both from peers' work and their suggestions. The PA process allowed teachers to reflect on their own Challenge, understand how to improve it, and review it.

Effects of GA. The reasons why teachers reported learning the most from GA can be summarised in two main effects. On the one hand, GA has improved teachers' understanding of the subject matter, i.e. throwing down a Challenge, allowing teachers to figure out what might be unstimulating and unchallenging. On the other hand, GA has allowed teachers a multiple comparative

6. Right-sided one-tailed test. If not specified, it is a two-tailed test.

process. The first comparison was between one's own work and that of peers, and the second was between the work of different peers. GA allowed teachers to see others' work and diverse ways of throwing down a Challenge, enabling them to confront ways of working different from their own and different from others. Through GA, teachers acknowledged the strengths and weaknesses of their peers' Challenges, learning to recognise those well done from those not. By GA, teachers compared their own Challenge with that of peers, understanding the quality of their own and the design differences. Alongside these motivations, some teachers also reported learning from GA because RA was useless. The latter result seems consistent with the fact that some teachers indicated that the peer assessments they received did not make them reflect on how to improve their Challenge. These results are not dissimilar to those of Nicol *et al.* (2014) and Li and Grion (2019). Indeed, these authors emphasised that GA allows students to gain a deeper understanding of the topic and to experience learning processes such as reflection and comparison.

The results regarding teachers' reflections on their own Challenge while GA allow to deepen the results just highlighted. Almost all teachers reported thinking about their own work and reflecting on how to improve it while assessing that of peers. By GA, teachers engaged in a comparative analysis of how their peers designed their Challenges compared to their own. This comparative process allowed them to identify possible similarities and differences, as well as strengths and weaknesses, providing valuable insights into different instructional strategies and approaches. By considering alternative perspectives and comparing them with their own, teachers engaged in self-reflection, critically analysed their Challenges, evaluated their quality, and looked at areas for improvement and change. Ultimately, teachers realised what they could do differently.

Effects of RA. RA helped teachers identify the strengths and weaknesses of their own Challenge, prompting critical evaluation and questioning of their work. Teachers valued the diverse perspectives provided by peers, finding them beneficial in assessing the effectiveness of their Challenge, identifying areas for improvement, and acknowledging limitations that could be addressed for enhancement. It is important to emphasise that, as mentioned regarding the uselessness of RA, for teachers, criticism only generated enrichment if it was constructive. These benefits are similar to those students derive from RA from peers (e.g., Cho & MacArthur, 2010; Nicol *et al.*, 2014; Topping, 1998). Indeed, Nicol et al. (2014) and Topping (1998) pointed out that RA allows students to gain insights into the strengths and weaknesses of their work, as well as view their work from different perspectives and reflect on aspects they may have missed.

The results regarding the impact of RA on teachers' reflection about improving their Challenge allow to deepen the results just highlighted. Most of the teachers reported reflecting on how to improve their own work

by RA. This indicates that even RA, although to a lesser extent than GA, stimulated self-reflection. Teachers reported that the peers offered useful hints and suggestions, leading them to consider improvements in their Challenge. Some teachers revised specific aspects or passages based on peer feedback.

6.3. Learning from PA: Conclusions and interpretations

The present findings, taken together, suggest that teachers learn from both roles, i.e. assessor and assessee. Moreover, the practice of GA offers considerable potential, arguably even more than could be possible through RA, for teachers. The latter result is consistent with research on students. Although, indeed, both GA and RA have strong educational benefits, recent research suggests that students learn more by GA on peers' work than by RA from peers (e.g., Cho & MacArthur, 2011; Culver, 2023; Gaynor, 2020; Li & Grion, 2019; Nicol *et al.*, 2014).

The fact that, for teachers, it seems more valuable GA than RA might depend on the quality of the assessments received, especially the feedback. As mentioned, some teachers reported that they did not receive useful feedback. However, this interpretation is not consistent with other findings. On the one hand, teachers generally considered their peers very capable of providing both qualitative feedback and quantitative scores. On the other hand, teachers' responses to the closed-ended question in the reflection phase reveal that most teachers found the feedback received from peers at least somewhat useful. The responses were as follows (total feedback = 91): not at all useful (7.7%), not very useful (19.8%), somewhat useful (45.1%), very useful (26.4%), and extremely useful (1.1%).

Considering the results of this study as a whole, the researcher interprets that teachers in the PA process seem to value GA over RA, consistent with Nicol's insights about *inner feedback*. Nicol (2019; 2021) theorised inner feedback as an internal generative process. He argues that producing - more than receiving - feedback for one's peers greatly impacts learning. When engaging in reviewing and producing feedback, each learner generates, simultaneously with the comments addressed to the authors of the work they are examining, a spontaneous and often implicit self-feedback. In these situations, learners activate processes of analysis, revision, integration, and re-elaboration of knowledge. Students spontaneously compare their peers' work with that previously produced by themselves and transfer the ideas generated by the comparison to improve their understanding of their work. GA, therefore, might strengthen what Nicol calls *inner feedback* processes and enable teachers to compare and calibrate inner and external feedback in ways that support their learning.

Finally, it should be noted that most teachers in the present study did not modify their Challenge because of the PA activity. This contrasts with findings from Nicol *et al.* (2014), where most students

reported improving their assignments and that both giving and receiving feedback were perceived as equally beneficial for this purpose. However, the study by Nicol *et al.* required an updated assignment, whereas the present study did not.

6.4. Implementing PA with teachers

Rating format. Teachers considered their peers very capable of giving equally quantitative scores and qualitative comments (items G and F: medium ESs; no statistically significant difference between items). This result is somewhat dissimilar to what is found in student literature. Research has shown that the marking component of PA causes dissatisfaction in students and that many are unfavourable to the idea of marking (Kaufman & Schunn, 2011). Students do not feel they have enough expertise to mark and/or are able to be fair or accurate (Nicol *et al.*, 2014). However, the importance of including both scores and comments in PA is emphasised in the literature (e.g., Li *et al.*, 2016; Li *et al.*, 2020). In addition, other variables that would be worth considering, for both teachers and students, are how PA is presented to teachers/students by the trainer/teacher and the quality of the guidance provided.

Assessors and assesses. Teachers would not have liked to have assessed more than 3 Challenges. This indicates that three was a good number for the PA phase (GA). However, teachers could assess more than 3 Challenges, as the researcher set up the platform to allow them to assess up to 5 other peer works. The student literature emphasises the importance of producing multiple reviews, which exposes students to multiple examples of work of varying quality, making them aware of alternative perspectives and leading to a broader range of possibilities (Nicol *et al.*, 2014). According to Sadler (2010), such exposure is essential for students to learn to recognise and generate high-quality work on their own.

Otherwise, teachers would have liked to have received more than 3 assessments of their Challenge (large ES), thus more than those actually received. Having more assessments could address the issue of poor useful feedback and increase the chances of receiving good-quality ones. The involvement of multiple peers naturally increases the amount and diversity of feedback (Topping, 1998), leading to greater improvement in the quality of students' work compared to feedback from a single peer (Cho & MacArthur, 2010). Getting feedback from multiple peers, therefore, also increases overall learning.

6.5. Limitations and future directions

Despite its valuable contributions, this study has certain limitations that warrant acknowledgement.

Sample size. The study's small sample size of 28 in-service teachers from Italian primary and lower secondary schools may limit the generalizability of the findings to a broader population of teachers in different

educational contexts. Future studies could include larger and more diverse samples to validate and expand upon these results.

Subjective perspective. The study solely focused on teachers' perceptions of the PA experience, which individual biases and experiences might influence. Incorporating additional objective measures could provide a more comprehensive understanding of the impact of PA.

Specific context and characteristics. The study examined the teachers' experience in the context of a PA activity focused on LU design and characterised by specific features. The characteristics of the activity may act as moderating variables for teachers' perceptions. Future studies could explore PA in a broader range of contexts and activities to understand its adaptability and effectiveness across different settings and fully capture the potential variations in PA practices.

Data. Combining quantitative and qualitative data may have data integration and interpretation limitations. In addition, data was derived exclusively from a survey. Future studies could include interviews and/or focus groups to gain a more comprehensive understanding of teachers' perceptions and experience of PA.

Single instance experience. The study focused on the PA experience in a single instance during the CPD course. A longitudinal approach, tracking teachers' experiences and perceptions over time, could provide deeper insights into the long-term impact of PA on teacher perceptions, growth, and instructional practices.

In conclusion, while the study shed light on the perceptions of in-service teachers regarding PA, its limitations suggest the need for further studies to strengthen the evidence base and broaden the understanding of the effectiveness and applicability of PA in teacher CPD.

7. Conclusion

The present study provides insights into PA theory and practice in the context of teacher CPD. It highlights the value of PA involving teachers as a valuable learning experience. Teachers recognised its value both in terms of learning and improving their own work. Specifically, they perceived the importance of both processes and both roles of PA: GA and RA, assessors and assessees. Nonetheless, GA seems to be perceived as more beneficial than RA. The study also suggests implementing PA as a primarily formative evaluation vehicle for in-service teachers' work. It provides insights into the number of assignments per assessor, the number of assessments per assessee, and how to structure the PA process to enhance its formative dimension.

Finally, the researcher aimed to stimulate the adoption and the productive use of PA in teachers' teaching practices by providing concrete and direct experiences of PA, outlining fruitful strategies to apply it and enabling teachers to evaluate its potential for classroom use with students. This responds to the call of researchers advocating that teachers receive training on conducting PA with their students (e.g., Sanchez *et al.*, 2017; Sebba *et al.*, 2008).

Experiencing PA first-hand can make teachers familiar with these practices and help them acquire the necessary skills to undertake PA effectively with their students (Cheng *et al.*, 2010; Yilmaz, 2017), and thus exploit the educational benefits that make it a fully-fledged practice that fosters learning (Boud, 2000; Nicol, 2010; Sanchez *et al.*, 2017; Sebba *et al.*, 2008; Topping, 2018; 2022).

Appendix

Response	Observed frequency (f_O)	Expected frequency (f_E)	$X^2 = \frac{(f_O - f_E)^2}{f_E}$	Critical X^2 (df) (α = .05 EW)	Decision
Both*	16	7	11.57	$X^2(1) = 5.738$	Reject Ho
GA	8	7	.14	$X^2(1) = 5.738$	Accept Ho
RA	3	7	2.29	$X^2(1) = 5.738$	Accept Ho
Neither	1	7	5.14	$X^2(1) = 5.738$	Accept Ho
*Sum**	*28*	*28*	*19.14*	*$X^2(3) = 7.815$*	*Reject Ho*

* = statistically significant. EW = experimentwise = α = .05. CW = comparisonwise = corrected α = α/*k* = .05/3 = .0166. Critical X^2(*corrected* α; *df*) = Critical X^2(.0166; 1) = 5.738. H_0 = null hypothesis.

Tab. 2 - *PA: Post-hoc tests.*

Compared pair	Response	Observed frequency	Expected frequency	Chi-square test result	Decision
Both vs Neither*	Both	16	8.5	$X^2 = 13.2$, $df = 1$, $p < .001$	Reject Ho
	Neither	1	8.5		
Both vs GA	Both	16	12	$X^2 = 2.67$, $df = 1$, $p = .102$	Accept Ho
	GA	8	12		
Both vs RA*	Both	16	9.5	$X^2 = 8.89$, $df = 1$, $p = .003$	Reject Ho
	RA	3	9.5		
Neither vs GA	Neither	1	4.5	$X^2 = 5.44$, $df = 1$, *exact* $p = .039$	Accept Ho
	GA	8	4.5		
Neither vs RA	Neither	1	2	$X^2 = 1$, $df = 1$, *exact* $p = .625$	Accept Ho
	RA	3	2		
GA vs RA	GA	8	5.5	$X^2 = 2.27$, $df = 1$, $p = .132$	Accept Ho
	RA	3	5.5		

The Bonferroni corrected alpha level - considered to reject H_0 - is .0083 (i.e., corrected α = .05/6 compared pairs = .0083). *p* refers to asymptotic significance. *Exact p* refers to exact significance. The *exact p* was calculated when Chi-square test assumptions were unmet.

Tab. 3 - *PA: Post-hoc pairwise comparisons.*

	RA		
GA	No	Yes	*Total*
No	0	1	1
Yes	11	16	27
Total	11	17	28

Tab. 4 - *Contingency Table GA (No; Yes) X RA (No; Yes).*

Response	Observed frequency (f_O)	Expected frequency (f_E)	$X^2 = \frac{(f_O - f_E)^2}{f_E}$	Critical X^2 (df) (α = .05 EW)	Decision
No GA-Yes RA	1	7	5.14	$X^2(1) = 5.738$	Accept Ho
Yes GA-No RA	11	7	2.29	$X^2(1) = 5.738$	Accept Ho
Yes GA-Yes RA*	16	7	11,57	$X^2(1) = 5.738$	Accept Ho
No GA-No RA*	0	7	7	$X^2(1) = 5.738$	Accept Ho
*Sum**	*11*	*28*	*26*	*$X^2(3) = 7.815$*	*Reject Ho*

EW = .05. CW = .05/3 = .0166. Critical X^2(.0166; 1) = 5.738.

Tab. 5 - *GA and/or RA: Post-hoc tests.*

Compared pair	Response	Observed frequency	Expected frequency	Chi-square test result	Decision
No GA-Yes RA vs Yes GA-No RA*	No GA-Yes RA	1	6	$X^2 = 8.33$, $df = 1$, $p = .004$	Reject Ho
	Yes GA-No RA	11	6		
No GA-Yes RA vs Yes GA-Yes RA*	No GA-Yes RA	1	8.5	$X^2 = 13.23$, $df = 1$, $p < .001$	Reject Ho
	Yes GA-Yes RA	16	8.5		
No GA-Yes RA vs No GA-No RA	No GA-Yes RA	1	.50	-	-
	No GA-No RA	0	.50		
Yes GA-No RA vs Yes GA-Yes RA	Yes GA-No RA	11	13.5	$X^2 = .926$, $df = 1$, $p = .336$	Accept Ho
	Yes GA-Yes RA	16	13.5		
Yes GA-No RA vs No GA-No RA*	Yes GA-No RA	11	5.5	$X^2 = 11$, $df = 1$, $p < .008$	Reject Ho
	No GA-No RA	0	5.5		
Yes GA-Yes RA vs No GA-No RA*	Yes GA-Yes RA	16	8	$X^2 = 16$, $df = 1$, $p < .008$	Reject Ho
	No GA-No RA	0	8		

The Bonferroni corrected alpha level - considered to reject H_0 - is .0083 (i.e., corrected α = .05/6 compared pairs = .0083).

Tab. 6 - *GA and/or RA: Post-hoc pairwise comparisons.*

Bibliografia

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

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101.

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589-597.

Carlsson Hauff, J., & Nilsson, J. (2022). Students' experience of making and receiving peer assessment: the effect of self-assessed knowledge and trust. *Assessment & Evaluation in Higher Education*, *47*(6), 959-971. Retrieved from https://doi.org/10.1080/02602938.2021.1970713 [Accessed 05.06.24].

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

Cecchinato, G., Papa, R., & Foschi, L.C. (2019). Bringing game elements to the classroom: The role of challenge and technology. *Italian Journal of Educational Technology*, *27*(2), 158-73. Retrieved from https://doi.org/10.17471/2499-4324/1078 [Accessed 05.06.24].

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), 91-104.

Cho, Y.H., & Cho, K. (2011). Peer reviewers learn from giving comments. *Instructional Science*, *39*(5), 629-643.

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction*, *20*(4), 328-338. Retrieved from https://doi.org/10.1016/j.learninstruc.2009.08.006 [Accessed 05.06.24].

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, MI: Erlbaum.

Culver, C. (2023). Learning as a peer assessor: evaluating peer-assessment strategies. *Assessment & Evaluation in Higher Education*, *48*(5), 581-597.

Davies, P. (2000). Computerized Peer Assessment. *Innovations in Education and Training International*, *37*(4), 346-355.

Felisatti, E., Serbati, A., & Paccagnella, S. (2020). L'esperienza di Peer Review nel corso di laurea in Scienze della Formazione Primaria: esiti di una ricerca valutativa sulle percezioni degli studenti. *Italian Journal of Educational Research*, (24), 223-244. Retrieved from https://doi.org/10.7346/SIRD-012020-P223 [Accessed 05.06.24].

Foschi, L.C. (2021). Innovative aspects and evaluation methods in a teachers' continuous professional development training experience. Italian Journal of Educational Technology, 29(1), 46-64. Retrieved from https://doi.org/10.17471/2499-4324/1165 [Accessed 05.06.24].

Foschi, L.C. (2022). *Pianificare, implementare e valutare lo sviluppo professionale continuo dei docenti: Risultati di una ricerca valutativa su un'esperienza di formazione* [PhD thesis]. Padova: University of Padua. Retrieved from https://hdl.handle.net/11577/3475619 [Accessed 05.06.24].

Foschi, L.C., & Cecchinato, G. (2019a). Validity and reliability of peer-grading in in-service teacher training. *Italian Journal of Educational Research*, *Special Issue*, 177-194. Retrieved from https://ojs.pensamultimedia.it/index.php/sird/article/view/3276 [Accessed 05.06.24].

Foschi, L.C., & Cecchinato, G. (2019b). Valutazione tra pari e autovalutazione nella formazione in servizio degli insegnanti. In P. Lucisano & A. M. Notti (Eds.), *Proceedings of SIRD International Conference - Training actions and evaluation processes* (pp. 441-453). Lecce: Pensa MultiMedia. Retrieved from https://www.sird.it/wp-content/uploads/2018/07/Volume-Completo.pdf [Accessed 05.06.24].

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), 49-64. Retrieved from https://doi.org/10.17471/2499-4324/1019 [Accessed 05.06.24].

Gaynor, J.W. (2020). Peer review in the classroom: student perceptions, peer feedback quality and the role of assessment. *Assessment & Evaluation in Higher Education*, *45*(5), 758-775.

Gibbons, J. (1993). *Nonparametric Statistics*. Newbury Park, CA: Sage.

Grion, V., & Restiglian, E. (2021). Il punto di vista degli insegnanti sulla valutazione fra pari: risultati di una ricerca empirica. *RicercAzione*, *13*(1), 39-56. Retrieved from https://doi.org/10.32076/RA13109 [Accessed 05.06.24].

Grion, V., & Tino, C. (2018). Verso una "valutazione sostenibile" all'università: percezioni di efficacia dei processi di dare e ricevere feedback fra pari. *LLL, 31*(14), 38-55. Retrieved from https://doi.org/10.19241/lll.v14i31.104 [Accessed 05.06.24].

Hammersley, M., & Atkinson, P. (1983). *Ethnography: Principles in practice*. London: 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), 387-406.

LaDonna, K.A., Taylor, T., & Lingard, L. (2018). Why open-ended survey questions are unlikely to support rigorous qualitative Insights. *Academic Medicine*, *93*(3), 347-349. Retrieved from https://doi.org/10.1097/ACM.0000000000002088 [Accessed 05.06.24].

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.

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.

Li, L., & Grion, V. (2019). The power of giving feedback and receiving feedback in peer assessment. *Education All Ireland Journal of Teaching and Learning in Higher Education*, *11*(2), 1-17. Retrieved from https://ojs.aishe.org/index.php/aishe-j/article/view/413 [Accessed 05.06.24].

Li, L., & Steckelberg, A. (2006). Perceptions of web-mediated peer assessment. *Quarterly*, *10*(2), 265-269.

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), 179-197.

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

Nicol, D. (2019). Reconceptualising feedback as an internal not an external process. *Italian Journal of Educational Research*, *Special issue*, 57-69. Retrieved from https://ojs.pensamultimedia.it/index.php/sird/article/view/3270 [Accessed 05.06.24].

Nicol, D. (2021). The power of internal feedback: exploiting natural comparison processes. *Assessment & Evaluation in Higher Education*, *46*(5), 756-778. Retrieved from https://doi.org/10.1080/02602938.2020.1823314 [Accessed 05.06.24].

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

Pallant, J. (2007). *SPSS Survival Manual*. New York, NY: McGraw Hill Open University Press.

Pearson, K. (1900). Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A*, *195*, 1-47.

Restiglian, E. (2019). Perspectives on translating Higher Education peer feedback approaches into schools. *Education All Ireland Journal of Teaching and Learning in Higher Education, 11*(2), 1-8. Retrieved from https://ojs.aishe.org/index.php/aishe-j/article/view/412 [Accessed 05.06.24].

Sadler, D.R. (1987). Specifying and Promulgating Achievement Standards. *Oxford Review of Education*, *13*(2), 191-209. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/0305498870130207 [Accessed 05.06.24].

Sadler, D.R. (2010). Beyond feedback: developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, *35*(5), 535-550.

Sanchez, C.E., Atkinson, K.M., Koenka, A.C., Moshontz, H., & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology*, *109*(8), 1049-1066.

Sebba, J., Crick, R., Yu, G., Lawson, H., Harlen, W., & Durant, K. (2008). *Systematic review of research evidence of the impact on students in secondary schools of self and peer assessment. Technical report*. London: EPPI-Centre.

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 & Evaluation in Higher Education*, *27*(5), 443-454.

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

Topping, K.J. (2009). Peer Assessment. *Theory Into Practice*, *48*(1), 20-27.

Topping, K.J. (2018). *Using peer assessment to inspire reflection and learning* (1st ed.). Routledge.

Topping, K.J. (2022). Elaborated dialogic feedback and negotiated action in peer assessment: Metacognitive benefits for assessor and assessee. In E. Postiglione (Ed.), *Fostering Inclusion in Education: Alternative Approaches to Progressive Educational Practices* (pp. 79-103). Cham, CH: Springer.

Topping, K.J. (2023). Digital peer assessment in school teacher education and development: a systematic review. *Research Papers in Education*, *38*(3), 472-498. Retrieved from https://doi.org/10.1080/02671522.2021.1961301 [Accessed 05.06.24].

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.

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), 395–409. Retrieved from https://files.eric.ed.gov/fulltext/EJ1147493.pdf [Accessed 05.06.24].

Zhan, Y., Yan, Z., Wan, Z. H., Wang, X., Zeng, Y., Yang, M., & Yang, L. (2023). Effects of online peer assessment on higher-order thinking: A meta-analysis. *British Journal of Educational Technology*, *54*(4), 817-835. Retrieved from https://doi.org/10.1111/bjet.13310 [Accessed 05.06.24].