TEACHUP

PEER VERSUS EXPERT ASSESSMENT

How to make assessment in online teacher training work

June 2020

Co-funded by the
Erasmus+ programme
of the European Union

Front cover picture: ©vegefox.com - stock.adobe.com

Publisher:
European Schoolnet
(EUN Partnership AISBL)
Rue de Trèves 61
1040 Brussels - Belgium
www.europeanschoolnet.org

Authors:
Katja Engelhardt, *European Schoolnet*
Benjamin Hertz, *European Schoolnet*
Janet Looney, *European Institute of Education and Social Policy*
Davide Azzolini, *FBK - IRVAPP*
Sonia Marzadro, *FBK - IRVAPP*
Enrico Rettore, *FBK - IRVAPP*

Editor:
Patricia Wastiau, *European Schoolnet*
Roger Blamire, *European Schoolnet*

Design and DTP:
Jessica Massini, *European Schoolnet*
Andrea Panizza, *European Schoolnet*

Published in July 2020

European Schoolnet
Transforming education in Europe

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

Policy concern, research question and key results

In scalable online courses like MOOCs, with potentially many participants, external expert assessment of course work becomes more challenging and costly. In that context, peer assessment can potentially play an important role in online teacher training, as it provides numerous potential benefits. However, peer assessment can fulfil its potential only provided that peer and expert assessment scores are quite similar (i.e. they need to demonstrate inter-rater reliability) and the qualitative feedback provided is useful.

The main research question in this report is the following: in scalable online courses, **is peer assessment a viable approach** to assess learning achievements and is it an **appropriate alternative to expert assessment?** To address this question, the TeachUP policy experimentation compared assessments of the final course assignment – a lesson plan – by external experts and by course participants in the third TeachUP online course. The assessments included numerical scores, as well as qualitative feedback. The overview below shows the four aspects of the research question and the key findings.

This report compares peer assessment and instructor/expert assessment in the context of an online course developed for the TeachUP project to address the first three aspects above. The fourth was addressed by means of a survey of teachers and student teachers participating in an assessment activity undertaken by peers and instructors/experts as part of the TeachUP online course.

In summary, peer assessment is a potentially viable approach to assess learning achievement in scalable online courses. While both peer and expert assessment was appreciated by course participants, small differences in how assessments were perceived became apparent. This might suggest that both assessment forms have their unique advantages in the eyes of participants and appear as complementary.

Aspect of the research question	Key result
1. The design of an assessment tool setting out performance categories, levels and criteria, and providing guidance on the use of evidence to support judgements	Assessments provided by three peers were generally consistent with ratings provided by experts (thus demonstrating inter-rater reliability). This underlines the importance of providing a well-designed assessment tool, such as a rubric, setting out clear standards and criteria with descriptors and exemplars of work at different performance levels.
2. The reliability of summative scores *(e.g. focusing specifically on inter-rater reliability, i.e. whether final scores given by experts and peers were consistent between and among the expert and peer raters);*	Inter-rater reliability for peer scores was high. Peer scores were however, consistently slightly higher than expert scores.
3. The quality and usefulness of qualitative feedback provided as part of the assessment process	Overall, peers' and experts' qualitative feedback was quite similar. Peers' feedback on teachers' plans was typically less detailed, included less concrete suggestions for improvement and was slightly more positive in tone than feedback provided by experts.
4. Participants' perceptions on the fairness and usefulness of assessments	Assessments by both experts and peers were largely perceived as both useful and fair, with higher agreement rates on the fairness of peer assessment.

Policy recommendations

Peer assessment in TeachUP is a viable option for assessment and may support the scalability of large online courses. Key recommendations emerging from the research are to:

1. **Boost assessment cultures**. Training for course participants on how to provide and receive feedback, in general and in online settings in particular, should be envisaged. The benefits of peer assessment both in online teacher training and with students should be further promoted. Guidance on peer learning activities should be accessible and engaging.

2. **Emphasise the learning gain of being assessor of others and of peer dialogue**. Assessing the work of others was identified as a particularly enriching learning experience. To strengthen this aspect, learning opportunities for peer assessors related to their own professional development, should be emphasised. Further, it should be possible for those assessed to respond to feedback received, thereby creating a dialogue between peers based on their concrete coursework (*on, for example, respective teaching beliefs)*.

3. **Enhance the reliability of peer assessment**. The assessments provided by three peers were generally consistent with ratings provided by experts *(demonstrating inter-rater reliability)*. This underlines the importance of providing a well-designed assessment tool, such as a rubric, setting out clear standards and criteria with descriptors and exemplars of work at different performance levels. Inter-rater reliability may be further enhanced through training on the use of the assessment tools and to ensure a shared understanding of performance levels.

4. **Develop quality frameworks for peer assessment in scalable online teacher training courses**. Given that assessments provided by three peers were generally consistent with ratings provided by experts, peer assessment can function as a reliable way to validate and certify teachers' progress when appropriate and effective assessment processes, tools, and guidance are in place. Of course, if the validation of learning/progress occurs on scalable online courses using peer assessment depends significantly on the design of the peer assessment processes, tools, and guidance provided. Therefore, to facilitate the accreditation of scalable online courses for teachers, quality frameworks for the use of peer assessment in these courses should be developed.

5. **Enhance the quality of peer feedback**. As experts' qualitative feedback was slightly more constructive and more detailed than that of peers, it would be useful to enhance the quality of peer feedback in online courses, for example, by providing examples of expert feedback, research findings about giving effective feedback, and opportunities to practise drafting feedback to fellow teachers. This could help course participants provide more constructive and detailed feedback in the peer assessment process.

6. **Select the assessment approach depending on the focus and scale of the course**. The findings suggest that both assessment forms may have their unique advantages in the eyes of participants and appear as complementary. Online course providers relying only on either peer or expert assessment could accordingly find mechanisms that allow for both types of assessment to work alongside each other. For courses with a strong focus on establishing a professional community online, peer assessment might be more suited. Equally, for courses with many participants, peer assessment might be the only feasible solution. For online courses, however, with a focus on introducing new and complex content or practices, some element of expert assessment might still be useful. While integrating expert assessment is difficult in scalable contexts, an optional and paid-for-offer of expert assessment could be offered alongside peer assessment to participants who are looking for more substantive and constructive qualitative feedback on their work.

2. ACKNOWLEDGEMENTS

This report builds on the contributions of all those involved in the TeachUP experiment. First and foremost, thanks go to the external experts who assessed the TeachUP lesson plans and TeachUP teachers and student teachers who participated in the project survey. The work of the colleagues at the Research Institute for the Evaluation of Public Policy ensured that the experiment followed a rigorous methodology and their statistical analysis and contributions allowed for a clear presentation and understanding of the results. The other 16 TeachUP partners both contributed substantially to the analysis and understanding of the results and shaped the overall direction of the project and its outcomes. Their work also ensured that all elements of the field trials were successfully implemented at national level.

3. INTRODUCTION

Peer assessment is generally recognised as an effective assessment and learning method with school students. Benefits include the development of metacognitive skills, a clearer understanding of assessment criteria, and a deeper engagement with course materials. While there is a considerable body of experience and research on peer assessment with school or higher education students, its use in teacher education and training is less investigated.

However, peer assessment in teacher education and training is becoming more important. Research confirms that teacher professional development is more effective when teachers share their expertise and experience systematically (Chong & Kong, 2012; Laurillard, 2016; Schleicher, 2016). This means that pedagogical activities such as peer assessment, where teachers engage in meaningful mutual exchanges concerning their professional practices, are becoming more prevalent. In addition, there is a need to significantly increase teacher education and training and scalable learning formats such as MOOCs offer opportunities to address this. However, due to their scale, such formats sometimes cannot use only instructor assessment and therefore make use of automated assessment and/or peer assessment to validate and certify learning. Accordingly, peer assessment in teacher education and training contexts is coming under increased scrutiny.

Peer assessment can be used in both a formative and summative way. However, for peer assessment to be effective as a formative form of assessment, teachers need to perceive the peer assessment process as a useful and fair form of assessment that offers them constructive and useful input into how to improve their work. For peer assessment to be effective as a summative form of assessment, teachers and external actors *(like employers)* need to perceive the peer assessment process as a valid and reliable way to validate and certify a teachers' learning/progress.

This report accordingly explores the peer assessment in the TeachUP policy experimentation by looking at the overarching question: **is peer assessment a viable approach to assess learning achievements in scalable online courses?** It does so by addressing the following three sub-questions:

1. Can online peer assessment processes and tools be designed to support reliable assessment by peers in an online teacher training context?
2. Is the qualitative feedback provided by peers as part of peer assessment in online teacher training contexts useful and constructive?
3. How do teachers perceive the process of peer assessment in online teacher training contexts?

The policy experimentation began with the assumption that learners are likely to prefer instructor/expert assessment to peer assessment. Experience of and expertise in conducting assessments as well as a good understanding of the object of assessment *(the work being assessed)* can help ensure that different assessors have a shared understanding of assessment criteria and what is considered good practice. Accordingly, instructors and experts acting as assessors could be considered more likely to achieve consistency of scoring between different assessors – in other words, their assessments are reliable. Furthermore, given the prevalence of instructors/experts functioning as assessors in key educational contexts such as for final exams in schools and universities, instructor/expert assessment can be considered a benchmark for other types of assessment such as peer or self-assessment. While limited evidence exists, it is highly likely that instructor/expert assessment also is the most prevalent form of summative assessment in initial teacher training and teacher professional development. The dominance of instructor/expert assessment is also likely to influence the value attributed to it by learners compared to peer or self-assessment. In other words, learners will see instructor/expert assessment as a benchmark for reliability compared to other types of assessment. This is to a certain extent confirmed by studies on student teachers' perceptions of peer assessment which show that student teachers are more critical of peer assessment, in particular regarding its validity and reliability, if they are less familiar with peer assessment (Kurtuldu & Özkan, 2019; McGarr & Clifford, 2013; Ratminingsih, Artini, & Padmadewi, 2017; Struyven, Dochy, & Janssens, 2008).

This report compares peer assessment with instructor/expert assessment in the context of an online course developed for the TeachUP project to address the first two questions above. The third question is addressed by means of a survey of teachers and student teachers participating in an assessment activity marked by peers as well as instructors/experts as part of the TeachUP online course.

4. ABOUT THE TEACHUP PROJECT

4.1. STARTING FROM A POLICY CONCERN

The TeachUP project was a major European research project (2017 - 2020) with 17 partners from 13 countries, and field trials in 10 countries. The KA3 policy experimentation was co-funded by the European Commission and coordinated by European Schoolnet. TeachUP organised a series of four online courses on formative assessment, personalised learning, collaborative learning and creativity (between October 2018 and May 2019). The first key principle of policy experimentations is that they start from a concrete policy concern. In TeachUP, this concern was that while online courses have the potential to provide flexible training opportunities for teachers at a large scale, only 36 % of teachers had already participated to an online course (OECD, 2019). Hence, the first and main research question was how to increase teachers and student teachers' participation in scalable online courses. To address this question, TeachUP tested whether a personalised support system might increase course participation. As a related question, TeachUP also investigated possible impacts on course participants' self-regulated learning online (SRLO) competence. The results for these questions are summarised in the reports Implementing personalised support in Scalable Online Courses and the TeachUP Evaluation report. The secondary research question - and focus of this report - was if peer assessment was a viable approach to assessment in online courses, and how course participants valued peer and expert assessments. This report focuses on the results with regard to this question.

4.2. METHODOLOGY – THE EXPERIMENTAL SETUP

Policy experimentations require a rigorous research methodology to test possible solutions to policy concerns. TeachUP implemented field trials with over 4000 randomly sampled teachers and student teachers in 10 countries *(Austria, Hungary, Greece, Estonia, Malta, Lithuania, Portugal, Spain, Slovakia, Turkey)*. To answer the first research question, TeachUP compared a group receiving personalised support, which consisted of emails with personalised guidance, and an offer of support to a group not receiving it - in order to evaluate the impact of personalised support on course participation. TeachUP teachers and student teachers were invited to complete two short surveys *(Baseline/Follow-up)*, as well as shorter surveys at the end of each course. For the analysis, the project used data from the surveys and the course platform itself. To enrich this quantitative analysis, qualitative feedback from key stakeholders during three workshops 'Country Dialogue Labs' in each field trials countries was collected. Detailed reports on these aspects are available here.

To address the secondary research question concerning the reliability and usefulness of peer assessment in online courses for teachers which are the focus of this report - TeachUP compared the peer and expert assessments of 106 randomly selected course assignments from the third TeachUP course. For each of the 10 countries, one expert was appointed to assess the course assignments following the same guidelines as peers. Each assessment was based on a scoring rubric setting out criteria for performance at difference levels. The course assignments received given numerical scores and qualitative feedback. To compare the average numerical scores given by peers and experts, the average score for each single assessment was computed, as well as the average score for all peers for one lesson plan. Qualitative feedback was categorised according to their length, tone, and how constructive *(including concrete suggestions for improvement)* they were. In addition, the authors of the course assignments were asked to fill in a short survey to collect their views on both assessment forms. Finally, the experts from four countries *(HU, EE, ES, PT)* were interviewed on their opinion on the assessment process.

4.3. THE TEACHUP COURSE SERIES

The series of courses developed as part of the TeachUP project focused on four key pedagogical topics associated with the changing role of the teacher and students. The topics were identified on the basis of a survey of initial teacher education and continuous professional development organisations on topics of interest for online coursework, as well as a literature review related to the changing role of teachers for the identified topics. Each course ran for three-and-a-half weeks and provided an introduction to a specific pedagogical concept and its underlying theory, followed by examples and ideas for the practical implementation of this approach. At the end of each course, each participant developed a lesson plan, related to their own teaching context, and incorporating the ideas gathered during the course. The courses addressed four topics:

Figure 1: *The TeachUP Course Series*

Successful participants received a digital badge and certificate. Those who completed the entire course series were awarded an additional badge and a certificate acknowledging their achievement.

The courses used an instructional design allowing for scalability and based on the principles of constructivist and connectivist learning *(the former referring to learner-centred approaches, and the latter to network complexity and self-organisation)*. Accordingly, courses were structured around core content, and a number of community building dynamics were developed in order to encourage participating teachers to exchange and share their experiences and expertise with each other. The course content comprised different types of stimulus materials, including classroom observation videos, teacher and student interviews, screencasts and short practice-focused researcher presentations.

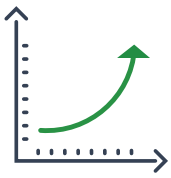
In order to fully benefit from the course, participants were encouraged to involve themselves in the course community, which built up as the courses progressed. The community was decentralised to a certain extent, with participants connecting through Facebook, the course platform forum, and/or through one or more of the of the many social networking tools used on the courses. The links between these different channels were established by dedicated course moderators who connected participants and content across the growing network of activity.

All four courses remain available in all ten TeachUP languages as open educational resources for self-study or reuse by teacher training providers on the European Schoolnet Academy.

5. THE ASSESSMENT OF PARTICIPANTS IN THE COURSES

5.1. THE PURPOSE OF PEER ASSESSMENT IN TEACHUP

Assessment is a central part of teaching and learning. However, in traditional classrooms, instructors usually assess students' work, provide feedback, and assign grades. By adopting peer assessment, students become active participants in their own learning processes (Falchikov, 2013; Kobayashi, 2020), thereby transforming the process of assessment from a passive experience for students, into an active learning experience. In line with this, the purpose of using peer assessment in the TeachUP courses was not solely to have a mechanism to assess a course participants' work. Rather peer assessment was intended to serve a range of purposes:



To support learning about learning: First and foremost, the process of giving and getting feedback helps to deepen understanding of the topic being assessed. For the assessor, the intellectual demands of reflecting, ensuring the assessment is balanced, and of formulating and delivering feedback may support the peer assessor's own learning. Reviewing others' work also motivates students to improve the quality of their own work (Nicol, Thomson, & Breslin, 2014). For the assessee, the intellectual demands of receiving, reflecting on and evaluating the usefulness of the feedback, and deciding which aspects to take on board, all lead to learning gains.

To support learning about assessment: peer assessment can also facilitate learning regarding the process of assessment itself. Pre- and in-service teachers will design and use assessments with their own *(current and/or future)* students. Making critical judgements about the performance of their peers serves as useful practice for making judgements about student performances. Additionally, when peer assessment is used to give feedback on areas for improvement, it allows teachers to appreciate the benefits of assessment not only of learning *(summative assessment)* but also for learning *(formative assessment)*. It is also worth noting that assessment competences may be valuable for teachers working collaboratively *(e.g. in a professional learning community)* where they may give and receive feedback as a collaborative process.

To assess and validate course work: in a scalable online environment which does not allow for more traditional, instructor-based assessments (such as a MOOC), peers may provide final *(summative)* assessments. While some MOOCs may support automated scoring of simple responses, for more complex assignments, such as lesson plans or essays, peer assessment can be more effective. A well-designed system includes an assessment tool *(e.g. a rubric)* setting out the areas to be assessed, clear standards, criteria and descriptors, and guidance and training on how to use them. Typically, several peers assess the same piece of work in order to support the reliability of results. Strong quality control measures also contribute to an effective system for peer assessment for validation and certification of successful course completion.

To support the establishment of an online community of professionals: peer assessment can also serve as a tool for community building within the online course, or MOOC. It offers a framework for meaningful conversations between teachers and builds connections between and among peers. In online contexts, this type of interaction might be limited, so teachers need to be encouraged to use the peer assessment process as a form of communication between two professionals.

Peer assessment therefore supports the learning of both the assessors and assessees, provides a focus for online interaction and community-building, and serves as a way to sustain and support scalable online professional learning.

5.2. HOW PEER ASSESSMENT WAS ORGANISED IN TEACHUP COURSES

The peer assessment process used as part of the TeachUP courses attempted to address the four purposes outlined above as well as other good practices considered for peer assessment such as using a rubric, generating shared understanding by means of good practice examples, and requiring multiple reviews per participant.

The focus of assessment was a **lesson plan** which the course participants produced as part of the course. The purpose of the lesson plan was to consolidate what had been learned during the course as a final course product, and which could be used in the participating teachers' and student teachers' classroom practice. An example lesson plan as a good practice example was provided to the participants to guide their work and assessments.

Once participants submitted their lesson plan in the final module of the course they were asked also to **review the lesson plans of three peers** from the course[1]. In this way, everyone served both as an assessor and an assessee.

The assessment **rubric** *(see Figures 2 and 3)*, included eight categories of effective lesson plans and four performance levels ranging from 1 to 4 with 4 representing the highest level within each category. For each performance level a description was provided that illustrated what was expected at that level within each category.

The eight categories of effective lesson plans were:

1. Classroom cultures for collaboration
2. Methods to foster students' agency
3. Effective elements of collaboration
4. Assessment of collaborative learning
5. Tools for collaborative learning
6. Alignment to learning objectives
7. Diversity of activities
8. Balance between individual and group work

The rubric was provided in two versions to the participants. One included descriptions for each of the 4 levels in each category and was to be used as a reference to complete the second version *(see figures 2 and 3)*.

The second version was in the form of a spider web *(or bullseye)* for a **visual display** of the scoring of the 8 categories *(see figure 4)*. This rubric also included space for more **qualitative feedback** in text form to the peer being assessed. This version was to be filled in by the participants as part of the peer assessment process to give scores and offer qualitative feedback to their peers. A completed example of this version, also with a good practice example for the qualitative feedback, was provided to the participants.

The whole process was not anonymous as peers were able to see each other's names when reviewing their work. The instructions explained that the peer assessment was to be seen as part of an exchange with a fellow professional and that further exchange beyond the peer assessment process itself was encouraged.

During the peer assessment process, participants had the opportunity to report an inadequate lesson plan or inadequate peer review(s). Such a report triggered an investigation by the course organisers. If the investigation corroborated the initial complaint, the reported participant would not receive the course badge and certificate. However, spot-checks showed that inadequate submissions were in fact made by participants, nevertheless, these were often not reported by participants, despite such processes being in place.

The process of submitting the assignment, assigning reviewers, and collecting assessments of the three peers was limited to one and half weeks. If a course participant did not complete all steps of the process, they did not complete the course and accordingly did not qualify for the course certification.

1 Research from Bachelet et al. (2015) suggests that in peer grading scenarios 3-4 peer grades are required to achieve a high level of "accuracy".

Area	Descriptors			
	Descriptor level 4	*Descriptor level 3*	*Descriptor level 2*	*Descriptor level 1*
Classroom culture for collaboration	The lesson includes activities that purposefully establish a positive and trustworthy class culture for collaboration, for example through teambuilding activities and thoughtful division of students into groups.	The lesson includes a few activities that establish a positive and trustworthy class culture for collaboration.	The lesson includes no activities that establish a positive and trustworthy class culture for collaboration.	The lesson includes activities which will reduce trust and positive atmosphere in the class.
Fostering student's agency	The collaborative learning activities contain elements that effectively foster students' agency (autonomy).	The collaborative learning activities contain some elements designed to foster students' agency, but these do not fully support more student autonomy.	The collaborative learning activities contain very few elements designed to foster students' agency (autonomy).	The collaborative learning activities do not contain any elements designed to foster students' agency (autonomy).
Effective elements of collaborative learning	The collaborative activities are designed in a way that helps students to learn more than they would individually, involves every student and makes every student think.	The collaborative activities are designed in a way that helps students to learn somewhat more than they would individually, it involves most of the students and evokes most students to think.	The collaborative activities are designed in a way that students learn slightly more than they would individually, it involves only part of the students and does not really evoke students to think.	The collaborative activities are designed in a way that students learn the same as they would if they worked individually, it involves only part of the students and does not evoke students to think.
Assessment of collaborative learning	Assessment of the collaborative work focusses on the process and the product and the criteria are explained and fully clear to the students prior the activities. Students are involved in shaping the criteria. A rubric or similar assessment tool is used.	Assessment of the collaborative work focusses on the process and the product. A rubric or similar assessment tool is used.	Assessment of the collaborative work focusses only on the product. A rubric or similar assessment tool is used.	Assessment of the collaborative work focusses only on the product and there is no information provided about the criteria used for the assessment.
Tools for collaborative learning (digital and non-digital)	The tools used in the lesson facilitate collaborative learning and are appropriate for the activity types used. For example, the eTwinning portal is used to get students to collaborate in a project-based context with students from another class.	The tools used in the lesson facilitate collaborative learning but they are not always fully appropriate for the activity types used. For example, a Google document is used to collect shared notes but activities offer little opportunity for students to actually take notes. Or a common sheet is distributed to the group to fill in but there is no guidance on how to work on the task collaboratively.	The tools used in the lesson do not facilitate collaborative learning and they are not fully appropriate for the activity types used.	There are hardly any, or no digital or non-digital tools used in the lesson.
Alignment to Learning Objectives	The lesson plan is well aligned with all of its learning objectives.	The lesson plan is partially aligned with its learning objectives, the majority of learning objectives are reflected in the activities.	The lesson plan is mostly unaligned with its learning objectives, there are only one or two learning objectives which are reflected in the activities.	The lesson plan is not at all aligned with its learning objectives.
Diversity of activities	The lesson plan incorporates a broad diversity of learning activity types.	The lesson plan incorporates some diversity of learning activity types.	The lesson plan incorporates little diversity of learning activity types.	The lesson plan does not incorporate diversity of learning activity types, there is mainly one type of activity.
Balance between individual and group work	The lesson plan is well balanced between whole group, small group, and individual work and it is clear why each type of work is used at that instance.	The lesson plan is mostly balanced between whole group, small group, and individual work. It is not always clear why each type of work is used at that instance.	The lesson plan is not very balanced between whole group, small group, and individual work. It is not clear why each type of work is used at that instance.	The lesson plan includes only one type of work throughout the lesson and it is not clear why this type of work is used.

Figure 2: *The TeachUP assessment rubric (first version)*

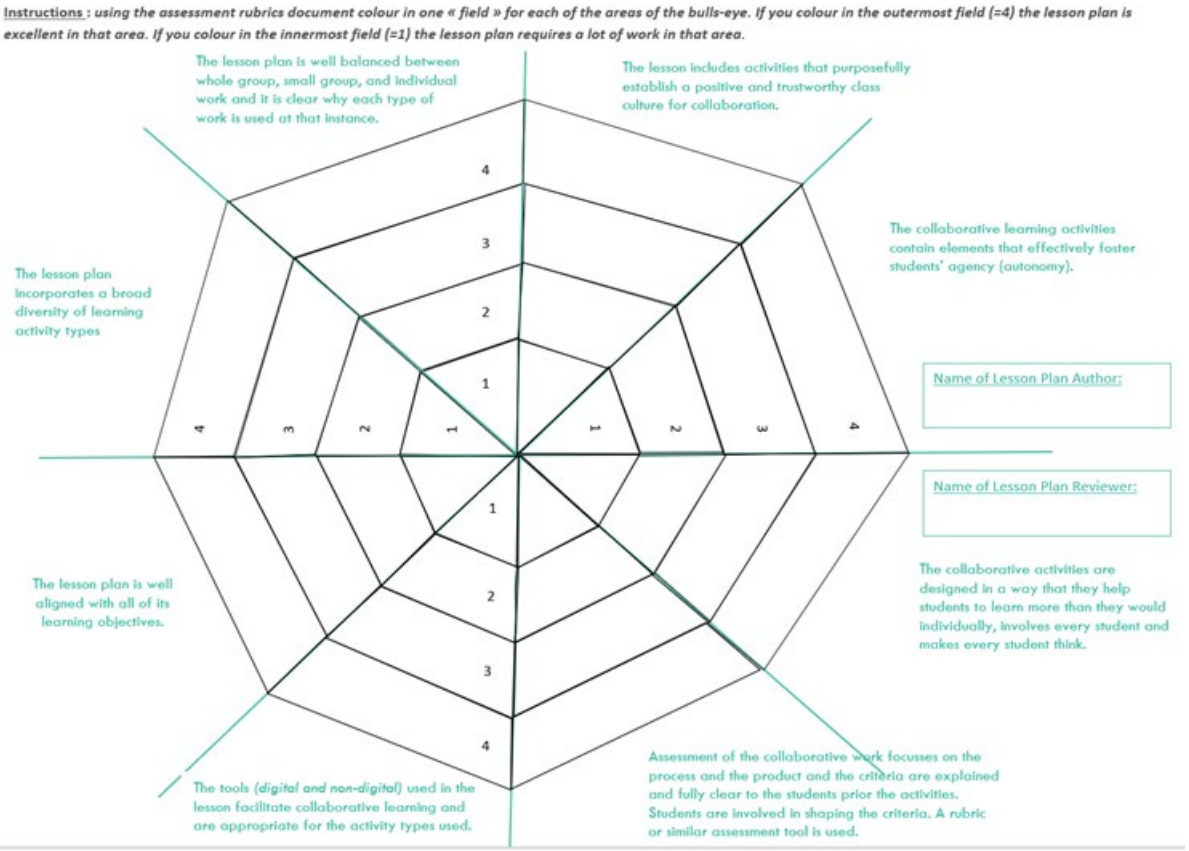


Figure 3: *The TeachUP assessment rubric (second version)*

6. EXPERT VERSUS NOVICE ASSESSORS

To understand whether peer assessment is a viable alternative approach, it is useful to understand the difference between expert and novice assessors. While peer assessors in the TeachUP courses had a large variety of teaching experience, including varied experience of assessing their students, most of them would have been new to the practice of assessing another teachers' work. In this sense, they would be regarded as novice assessors.

The seminal publication *How People Learn* (National Academy of Sciences, 2000) explores how experts differ from novices as a way to understand thinking and problem-solving. The report highlights that:

- Experts notice features and meaningful patterns of information that are not noticed by novices.
- Experts have acquired a great deal of content knowledge that is organized in ways that reflect a deep understanding of their subject matter.
- Experts' knowledge cannot be reduced to sets of isolated facts or propositions but, instead, reflects contexts of applicability: that is, the knowledge is "conditionalized" on a set of circumstances.
- Experts are able to flexibly retrieve important aspects of their knowledge with little attentional effort (National Academy of Sciences, 2000, p. 31)[1].

These qualities are central to judging the quality of learners' work and to providing meaningful and useful feedback on how to improve it.

At the same time, it is important to note that expertise in and of itself does not guarantee the quality of an assessment. For example, expertise in providing qualitative feedback is needed. Assessment researchers have found that feedback that is timely *(within two weeks)*, tied to criteria regarding expectations, includes specific suggestions for improvement, and that avoids vague phrases such as "needs work" is more effective (Boulet et al., 1990; Butler, 1988).

Validity and reliability of an assessment depend on the quality of assessment tools used, and the assessor's use of evidence to support interpretation of performance levels and scoring. While experts would be expected to identify appropriate evidence more easily *(in line with the description of expertise above)*, they may also notice different elements based on their own experience and specialisation. Individual raters, whether experts or peers, may also have a different sense as to what they judge as a "poor", "good" or "excellent" performance by a student- or professional teacher, affecting the inter-rater reliability. For experts, the context of their work and students they typically work with may also affect their views on performance levels.

For these reasons, expert and peer assessors alike benefit from well-designed assessment tools with clear standards, criteria, descriptors and exemplars of learner work at different levels. Guidance, training[2] and/or opportunities for assessment to develop shared understandings may all support quality of assessment.

1 The authors also note that while experts have thorough knowledge of their disciplines, "....this does not guarantee that they are able to teach others", and that they "...have varying levels of flexibility in their approach to new situations". National Academy of Sciences, 2000, p. 31). For TeachUP, since the experts are themselves teachers, it is safe to assume that they have avoided potential weaknesses.

7. RESEARCH QUESTIONS

The TeachUP courses were designed as scalable online courses, able to accommodate large numbers of participants. As the number of participants increases, however, external expert assessment of course work becomes more challenging and costly. Peer assessment may have numerous potential benefits *(see section 5.1)*, provided that several peers can provide an assessment *(providing both numerical scores and qualitative feedback)* on course work that is of the same quality as that of an external expert[1] . This condition is generally considered to be important because teachers and other stakeholders in education appreciate expert assessment. The assumption is that peer and expert assessment scores need to be quite similar *(i.e. they need to demonstrate inter-rater reliability)* and that the qualitative feedback provided needs to be useful.

To that end, the TeachUP research question was the following: **is peer assessment a viable approach** to assess learning achievements in scalable online courses, that provides an **appropriate alternative to expert assessment?**

To address this question, TeachUP compared assessments of the final course assignment *(a lesson plan)* by external experts and by the course participants, with attention to:

1. the design of an assessment tool setting out performance categories, levels and criteria, and providing guidance on the use of evidence to support judgements;

2. the reliability of summative scores *(e.g. focusing specifically on inter-rater reliability, i.e. whether final scores given by experts and peers were consistent between and among the expert and peer raters);*

3. the quality and usefulness of qualitative feedback provided as part of the assessment process;

4. participants' perceptions on the fairness and usefulness of assessments.

The assessments that both peers and experts provided consisted of numerical scores and a qualitative feedback.

For assessments to fulfil their purpose, they need to be both valid and reliable. This research focuses on investigating the scores' reliability. If the scores are reliable, they accurately measure course participants performance. Reliability also refers to the consistency and stability of results across student populations and assessors. The assessment scores also need to be valid[2], which means that they measure what they are intended to measure. The question whether the assessments provided by peers are valid is outside of the scope of this research.

The second element, the qualitative feedback, is also a crucial part of peer assessment, as it shapes course participants' learning. The general assumption is that a feedback that has a certain length, a positive tone and is constructive *(including concrete suggestions for improvement)* is likely to be perceived as valid and useful by the person receiving it. Research on formative feedback has found that it is more effective when it is focused on the task at hand (rather than the ego of the learner), is specific enough to support the learner to make improvements *(with more advanced learners needing less detailed guidance)*, and is timely (Looney, 2011; William, 1998). In our analysis, we included the 'positive tone' as a criterion for analysis, as the tone is likely to be even more important for online feedback, as the assessor and assessee are likely not to have

1 Kane & Lawler 1978; Topping 1998; Falchikov & Goldfinch 2000, Van Gennip, Segers, and Tillema 2009; Hodgson 2010; Tillema, Leenknecht & Segers 2011; Hoogeveen & van Gelderen 2013; Li, Xiong, Zang, Kornhaber, Lyu, Chung, et al. 2016

2 Validity refers to the degree to which assessments and evaluations measure what they are intended to measure (i.e. how well they are aligned with standards and curriculum). Other widely accepted definitions highlight the importance of the "...process of constructing and evaluating arguments for and against the identified interpretation of ...scores and their relevance to the proposed use" (AERA, APA, 2014). Validity is seen as being the "...joint responsibility of the methodologists that develop the instruments and the individuals that use them" (Yale Poorvu Center for Teaching and Learning).

established a personal rapport. Therefore, online feedback may be more readily interpreted as negative.

Further, for peer assessment to be effective as a formative form of assessment, teachers need to perceive the peer assessment process as a useful and fair form of assessment that offers them constructive and useful input into how to improve their work. In traditional face-to-face trainings, it is usually the course instructors or another expert who assesses the learners' work. They have expertise both in the course topic and are experienced as assessors. Peer assessors, on the other hand, may have different levels of expertise both on the course topic and with assessing the work of others. In peer assessment, teachers with less teaching experience or even student teachers might assess the lesson plan of a more experienced colleague. While in principle their feedback can provide a fresh and new perspective, teachers' positive perception of the peer assessments they receive cannot be taken for granted, as for many going through the process of peer assessment in online learning is still a relatively new experience.

Finally, for peer assessment to be effective as a summative form of assessment, teachers and external actors *(like employers)* need to be assured of the quality of the peer assessment process to validate and certify a teachers' learning/progress. The TeachUP research question is also relevant because accreditation for scalable online courses is becoming an increasing concern for Ministries of Education and regional authorities who have started to utilize MOOCs as a format for teacher professional development, for example in Portugal, Spain and Croatia. In order for teachers to use course certificates from MOOCs to support career progression *(e.g. salary raises)* such formal accreditation will be necessary. Accordingly, the question as to whether a peer assessment system can ensure that participants who successfully complete a course have satisfied minimum criteria becomes even more important.

8. RESEARCH METHODOLOGY

To address the two research questions concerning the reliability and usefulness of peer assessment in online courses, the peer and expert assessments of 106 randomly selected lesson plans that course participants had submitted as their final course works were compared. The 106 lesson plans were 15% of all lesson plans submitted as part of the third TeachUP course on collaborative learning *(February-March 2019)*. 66 of the lesson plans were submitted by practicing teachers and 40 by student teachers. Up to three peers had already assessed the lesson plans as part of the course.

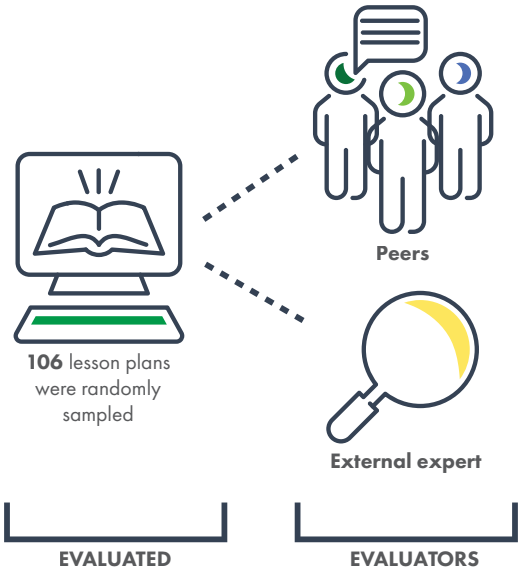


Figure 4: *The research setup*

This research only focused on the third course of the course series, as there was no reason to assume that there were considerable relevant differences in the peer assessments for the different courses. Focusing on one course only was also more feasible, given experts' availability to assess course work.

Expert assessment was organised using the same processes and tools as peer assessment. In total, ten experts assessed randomly selected lessons plans from their country, using the same instructions, assessment rubrics and templates as peers[1]. The experts were centrally contracted and paid for their tasks. The expert assessors were required to have at least five years' experience working as a qualified teacher, familiarity with teacher competences related to formative assessment, personalised learning, collaborative learning and creativity in classroom teaching *(the course areas featured on the TeachUP platform)*. They were also required to have good online communication skills, ability to learn quickly how to use new online tools, a good command of English, ability to pay close attention to detail, willingness to follow closely given instructions. Experience in teacher training was also desired but not a formal requirement. Experts were identified and selected by the TeachUP partners in each country according to the common selection criteria.

The TeachUP course participants, who also served as peer assessors, were student teachers and professional teachers. They had a varied mix of experience in formal assessment as compared to experts. However, it is possible that some teachers who acted as peer assessors had a similar or even greater level of experience as the experts did.

Both peers and experts completed the same spider web *(or bull's eye)* diagram *(the assessment rubric, with levels set out in a concentric circle)*, which included four assessment levels *(from 1 "the lesson plan requires a lot of work in that area" to 4 "the lesson plan is excellent in that area")* on eight different aspects of the lesson plan and the invitation to provide a qualitative feedback. Both peers and experts received the same instructions, including an example of what a good lesson plan looks like.

A statistical analysis was carried out to compare both the assessment scores and the qualitative feedback from experts and peers. To compare assessment scores,

a. first the average score of the eight assessment categories was computed for each single assessment,

b. then the average of scores that a single lesson plan received from several peers was calculated and

c. the overall average scores from all peer and all expert assessment were computed and compared.

1 *The external expert was assigned in each country. The Slovak external expert did not send the assessment of the one Slovak lesson plan and no Maltese expert was found to conduct the assessment.*

Peers and experts were also asked to give a qualitative feedback in writing to each lesson plan. This qualitative feedback should include concrete suggestions on what the lesson plan author could still improve. The qualitative feedbacks were categorized and compared according to their overall tone[2], length[3] , and how constructive *(including concrete suggestions for improvement)*[4] they were. The general assumption was that feedback that had a certain length, a positive tone and was constructive *(including concrete suggestions for improvement)* was likely to be perceived as useful by the person receiving it.

In addition, the authors of the randomly selected lesson plans were asked in a short survey how they valued their learning in the course in general, and the different assessment types: self-assessment as an optional element of the course, peer assessment and external expert assessment. Together with the invitation to take part in the survey, they received an expert assessment of their lesson plan. 71 Lesson plan authors[5] replied to the survey *(66 teachers, 22 student teachers)*. Moreover, online interviews about their views on the lesson plans as well as the assessment process were conducted with the expert assessors from Estonia, Hungary, Spain and Portugal.

For the general interpretation of the results, the following points might be relevant.

- First, it might be worthwhile to note that in the communication with the lesson plan authors, experts were referred to as 'external evaluators' and the experts' names were not mentioned. The peer assessment on the TeachUP courses, however, was not anonymous in order to encourage course participants to see the process as an open exchange between two professionals rather than a final summative evaluation of one's work. Keeping in mind this difference might be relevant, as it could have an impact both on the assessments themselves and how they were perceived by lesson plan authors.

- Second, while lesson plan authors were not specifically asked to indicate if they were a teacher or student teacher in their lesson plan, they were likely to do so in the background information section of their lesson plans. Hence, this information was not "double blinded", and both peers and experts were likely to know if the lesson plan they assessed had been written by a teacher or a student teacher.

- Third, experts received, together with the lesson plans, the peer assessments of those lesson plans by way of information, which may have influenced their own assessment.

- Finally, as the sample of selected lesson plans was relatively small, and not all lesson plan authors replied to the short survey, the results of this research cannot be considered to be representative for the entire TeachUP sample, nor the entire teacher population in the TeachUP countries.

---

2 1 – very negative, 2 – rather negative, 3 – rather positive, 4 - very positive

3 Length of the feedback: 1 – very short (less than 10 words), 2 – rather not detailed (between 11 and 50 words), 3 – rather detailed (between 51 and 100 words), 4 - very detailed (more than 100 words)

4 1 - feedback does not include concrete proposals for improvement, 2 – not clear / neutral, 3 – feedback contains concrete proposals for improvement

5 Austria 2, Estonia 4, Greece 7, Hungary 5, Lithuania 10, Malta 2, Portugal 7, Spain 9, Turkey 27

9. COMPARING PEER AND EXPERT ASSESSMENTS

Key results

1. Both peer and expert evaluators assessed the quality of teachers work *(i.e., the "lesson plan" that was the final course work)* as very high.
2. However, peers' scores were systematically higher than experts'. This difference is statistically significant but small.
3. Peers' feedback on teachers' plans was typically less detailed, less constructive and slightly more positive in tone than feedback provided by experts.
4. The assessments provided by peers on the same teacher's plan were generally consistent, even if there was some variability.

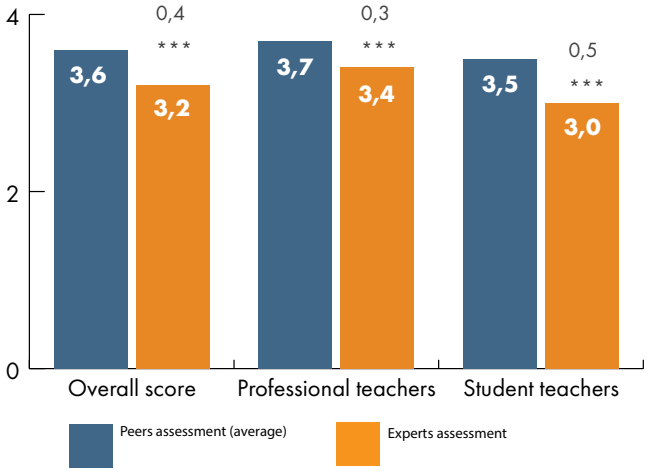


Figure 5: *Peers average score and external evaluators score, overall and by teachers' type*

Peers give higher scores than external evaluators in bold on top of the figure (as has been done for the others)

9.1. COMPARING PEER AND EXPERT ASSESSMENT SCORES

As figure 5 shows, the lesson plans were considered by both peers and experts to be very good overall. However, **peers, on average, gave higher scores than experts** *(3.6 versus 3.2, with 4 being the highest score)*. This difference is not very big but still statistically significant. Moreover, the difference was more pronounced for lesson plans from student teachers than those of professional teachers *(0.3 and 0.5 points difference respectively)*. One important first conclusion is that inter-rater reliability for peer scores is high. Peer scores were, however, consistently slightly higher than expert scores.

Note: The overall score is the mean of 8 categories scores ranging from 1 "the lesson plan requires a lot of work" to 4" the lesson plan is excellent ".

In general, the difference between scores is more pronounced in cases where experts gave lower scores, since peers systematically gave higher scores. This is also the case for student teachers' lesson plans *(as compared to those of teachers participating in the course)*. One explanation is that student teachers' lesson plans were overall slightly less good, due to student teachers' lack of teaching and assessment experience. Both peers and experts were likely to know if the lesson plan they assessed was submitted by a teacher or student teacher. Hence, in principle it is also possible that their score was influenced also by the assumption that lesson plans by student teachers are less likely to meet standards for a "good" or "excellent" course work.

The overall finding that peers provided slightly higher scores than experts seems to confirm previous research findings. For instance, Kilic and Cakan (2007) used peer assessment with pre-service primary science teachers, who taught a science topic as a team to their peers, who then assessed their performance using an assessment form provided by the teacher educator. While peer scores were considerably higher than teacher educators' scores, they still showed a significant correlation.

Based on our results, it is not possible to provide a

conclusive data-driven explanation as to why scores by peers are slightly higher. Below, four tentative possible explanations are provided, several of which might be at play, based on feedback from course participants via the survey and interviews with experts.

1. **Different understandings of course concepts** - Experts might have a better understanding of what is real mastery of collaborative learning, so what peers think is already excellent, based on their experience, is only good for experts.

2. **Feedback culture** – Possibly, course participants were not used to assessing each others' work, possibly even less so in an online environment where asessor and assessee do not know each other personally. This could lead them to be less comfortable in giving lower scores, which require more explanation and to phrase their feedback in a more positive tone *(using emoticons etc.)*. The extent to which this factor plays a role might also differ from country to country, according to the interviewed experts.

3. **Time spent on the task** – It cannot be excluded that for some course participants providing feedback with a good score required less explanation and was therefore likely to be produced quickly. This may have been an additional incentive to provide a higher score. We found some clues in favour of this interpretation by comparing the overall score of peers with the length of their feedback. Indeed, very high scores from peers were more likely to be coupled with very short feedback[1] . It can be seen as a general drawback of online courses that with online communities it is more difficult to reproduce the feeling of social pressure to perform that is at play in face-to-face trainings.

4. **Peer support and encouragement** – Peers may see an important part of their role as being the assessor to give encouragement to their fellow students. This view is also supported by peers being more likely to include phrases of direct encouragement in their qualitative feedbacks *(e.g. 'Very good job, my colleague', Good luck with your follow up!')*.

9.2. COMPARING PEER AND EXPERT ASSESSMENT SCORES ACCORDING TO ASSESSMENT CATEGORIES

As part of their assessment, experts and peers completed a spider web *(bulls' eye)* rubric, which set out criteria performance levels *(1 to 4 with 4 representing the highest level within each category)* for each of the eight different categories of the assessment rubric *(see section 5.2)*.

Figure 6 illustrates peer and expert scores in all of the eight assessment categories. For each of the eight assessment categories, peers' average scores were higher than 3.5, while expert scores ranged from an average score of 2.9 for the category 'assessment of collaborative work' to an average score of 3.4 for the category 'classroom culture for collaborative learning'. Differences between peers and experts were more pronounced and statistically significant when the object of evaluation concerned the focus of the assessment, effectiveness of designed activities, the tools *(digital and non-digital)* used in the lesson, the alignment of the lesson plan with all of its learning objectives and the balance between individual and group work. The difference is most pronounced for categories where experts gave lower average scores and is hence greatest for the category 'assessment of collaborative work' *(expert score 2.9 vs. peer score 3.5)*.

Note: stars represent significant differences at 95% confidence level.

1 *Peers who gave very short feedback gave an overall score of 3.8 while those who gave very long feedback gave an overall score of 3.6. The difference is small and still non-statistically significant also because of the low sample size.*

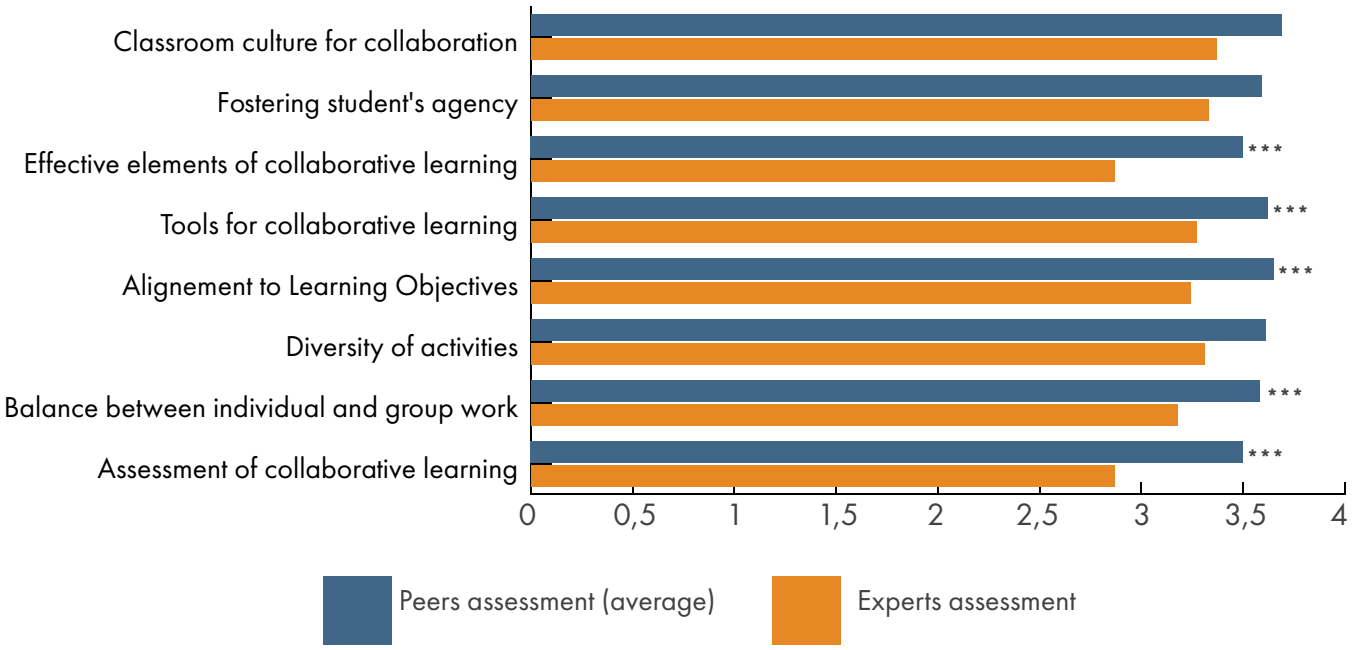


Figure 6: *Overall peers average score and expert score by categories*

One can only hypothesise as to why small but statistically significant differences were observed for some categories, but not others. In general, the difference between peer and expert scores were more pronounced for categories for which experts gave lower average scores. The same possible explanations as put forward as regards the general differences between overall scores *(section 9.1)* might be at play here. It may be that the assessment categories where the difference between scores is greater use more complex concepts with which teachers are less familiar. For these categories, experts might be more able or at ease in identifying weaknesses. It might also be the case that the assessment categories concerned are less well-defined and therefore more prone to different interpretations. Such an explanation would resonate with findings from earlier studies, e.g. by Iqbal and Mahmood (2008) where the teaching performance of pre-service teachers was observed both by supervising teachers and their peers. In this study, a greater degree of agreement between both the teacher and peer assessments was found for constructs involving tangible *(directly measurable)* characteristics.

Figure 7 also visualizes the differences between peer and expert scores for the eight assessment categories. It shows the distribution of the peer score for each possible score given by the experts to each assessment category. For instance, given an expert score equal to 1, peer average scores ranged from 1, hence exactly like experts' score, to 2.9 *(blue rectangle)*[1] . When experts scored 2 or 3, peers' scores were systematically higher in all areas even if discrepancies decreased (note the areas of the rectangles getting smaller). The expert/peer difference decreases as experts give a score equal to 4 (yellow rectangle) but at the maximum score, it is likely that variability is constrained by a "ceiling effect". To sum up, expert/peer difference decreases as expert scores increase (i.e. the height of the rectangles in the Figure is reduced).

Note: categories of assessment are: (1) Effective use of Formative Assessment Techniques; (2) Inbuilt Flexibility; (3) Feedback; (4) Goal Setting; (5) Balance between individual and group work; (6) Tools; (7) Diversity of activities; (8) Alignment to Learning Objectives.

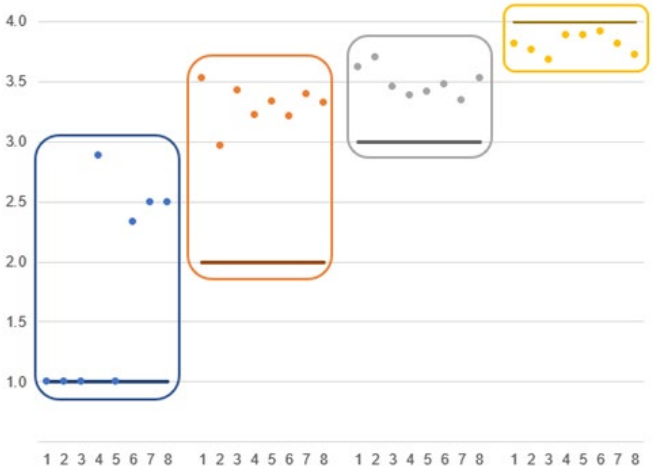


Figure 7: *Expert scores (line) and overall peers average score (dots) by the 8 categories*

1 *It is still worth remembering that scores so low are quite few.*

9.3. VARIABILITY IN PEER ASSESSMENT SCORES

While the previous sections focused on the comparison between expert and peer scores, this section looks only at the variability among peer assessors' scores. Each lesson plan received scores from up to three different peers. Having the same course work assessed by several peers is common in peer assessment and inter-rater reliability of assessors' scores helps to ensure the accurate representation of the performance level.

To understand overall peer score variability, we looked at: the differences in mean scores between lesson plans authored by different lesson plan authors and the differences between the three peer scores given for the same lesson plan. The results indicate that the overall peer score variability is primarily due to differences between the lesson plans *(70%)*, rather than from different scores to the same lesson plan by peers *(30%)*. In other words, peer assessment varies more between lesson plans (evaluated) than between peers (evaluators) of the same lesson plan. The variability between lesson plans probably simply reflects the fact that the overall quality of the lesson plans differed, as the expert scores also suggest *(see section 9.1)*. The experts interviewed *(EE, ES, HU, PT)* also confirmed that they found a certain variety in their lesson plans. However, a non-negligible part *(30%)* of the overall variance in peer assessment scores was due to different scoring by peers. This finding seems to resonate with the perception of those that received the peer assessments. 69,7% of survey respondents agreed that their lesson plan received similar scores from all peer assessments.

This may indicate a need to further refine the scoring rubric and guidance to ensure a shared understanding of the criteria for different performance levels. Alternatively, a system might be created for peers to assess a subset of the assignment and discuss any discrepancies among them before assessing the remainder of the assignment.

Overall, this is a promising result that suggests that peer assessment is an appropriate approach to online assessment. It is necessary to point out that these results are not representative, due to the relatively low number of lesson plans and survey respondents. Further investigation is needed to better understand the reasons for a certain variability of scores on the same lesson plans, and possible patterns. It could also be interesting to compare expert and peer variability for the same course work, which would allow further investigation of the question: Does the same course work receive similar scores from different experts? To that end, not only several peers but also several experts could be asked to assess the same lesson plan.

Figure 8: *Variability in peer assessment scores*

9.4. COMPARING QUALITATIVE FEEDBACK BY EXPERTS AND PEERS

In addition to the eight scores, peers and external evaluators were asked to give qualitative feedback on teachers' lesson plans. This feedback was then recorded into three indicators: the length of the feedback[1], the tone of the feedback[2] and the presence of constructive comments[3]. The general assumption was that a feedback that had a certain length, a positive tone and was constructive *(including concrete suggestions for improvement)* was likely to be perceived as useful by the person receiving it.

---

1 Length of the feedback: 1 – very short (less than 10 words), 2 – rather not detailed (between 11 and 50 words), 3 – rather detailed (between 51 and 100 words), 4 - very detailed (more than 100 words)

2 1 – very negative, 2 – rather negative, 3 – rather positive, 4 - very positive

3 1 - feedback does not include concrete proposals for improvement, 2 – not clear / neutral, 3 – feedback contains concrete proposals for improvement

Overall, peers' and experts' qualitative feedback were quite similar. Peer feedback was generally slightly more positive in tone. However, the difference was not great. This overall slightly more positive tone of peer feedback seems to be in line with the overall slightly higher scores given by peers. It might also reflect our assumption that peers may see an important part of their role as being the assessor to give encouragement to their fellow students, reflected in formulations such as *"Very good job, my colleague"*. Another difference was that expert feedback included more concrete suggestions for improvement. The biggest difference between experts and peers was, however, their length. Experts' qualitative feedback was overall longer than the one provided by peers *(see Figure 9)*. There might be several reasons for this difference. Experts may have felt more at ease in providing more detailed feedback, and generally have more experience in assessing others. As experts with more extensive knowledge and experience, they are likely to notice and interpret the quality of work more efficiently. They might have also felt more motivated or obliged to provide more detailed feedback, since they were contracted and paid for the task.

The question whether feedback that had a certain length, a positive tone and was constructive *(including concrete suggestions for improvement)* was indeed perceived as most useful by the person receiving it was beyond the scope of this research – also due to the limited amount of lesson plans and survey replies analysed.

Its first evidence suggests that the length of the qualitative feedback is not a clear predictor of it's appreciation. 66.67%[4] of those whose average qualitative feedback was rather not detailed[5] found it very useful. Perhaps surprisingly, of those that received a rather detailed feedback[6], only 50%[7] found if very useful. 31%[8] of those that received a very short average feedback[9] found it very useful. Due to the relatively low numbers of respondents, these results are not representative and would require further investigation.

Further, this research did not compare peer and expert feedback in terms of their actual effectiveness, i.e. which feedbacks triggered the biggest, and concretely which learning gains. Both questions would deserve further investigation.

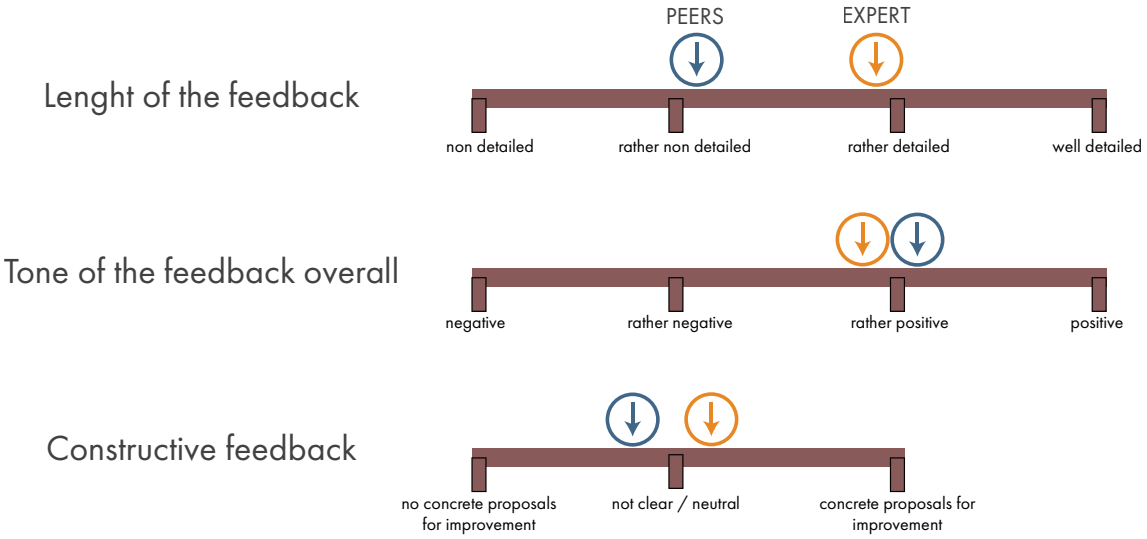


Figure 9: *Characteristics of peers' and experts' qualitative feedback on teachers' lesson plans*

4 *20 out of 30 survey respondents*

5 *between 11 and 50 words*

6 *between 51 and 100 words*

7 *10 out of 20 survey respondents*

8 *4 out of 13 survey respondents*

9 *less than 10 words*

These examples provide an idea of the range of feedback offered by peers and experts:

Example peer assessments

Dear

Your lesson plan about Australia has turned out great I've added a few suggestions below: The Answergarden activity at the onset of the lesson unit is great and should work out fine. Mind the different settings in Answergarden. A video, as an introduction to a thematic unit, can work wonders. Since you're covering the whole country, it is probably difficult to find one that is not too long. You might also want to try this activity as a flipped classroom activity and do the research activity in class, depending on the technical resources you have.

The combination of padlet/presentation/ evaluation is state oft he art, when evaluating each other, students could rely on a rubic to help them assess objectively. Here's an old but tried link to Rubistar, a rubics maker; there are probably more modern ones out there nowadays. http://rubistar.4teachers.org/

One thing you could think about is how students could profit more from each other's (learning). Overall, however, the lesson plan is very well balanced and should work fine with your students.

Austria

Example peer assessments

I don't really know what comment to write, I could read the lesson plan. It's good to be project-based and have clear and specific instructions so everyone knows what to do. There are many opportunities for students to discuss, share ideas, and I think they can get and learn a lot during these lessons, especially in the previous two hours. I also learned from it, I still have a lot of room to develop in my lesson by the time I get there to teach. And even though it was good to read a lesson plan that isn't 'too ICT', I don't think it's always necessary.

Hungary

Example expert assessment

The prepared lesson plan is of high quality and meets all the criteria for collaborative learning. It has to be admitted that many aspects are thought out. An obvious combination of individual work and group work. Group work is meaningful, because students will learn something more or new things than individually. The individuality of the child is not overshadowed, on the contrary, it allows the accumulation of individually accumulated knowledge while working with others. The lesson plan has a very prominent aspect of collaboration, learning from each other. However, there is doubt are students involved in the development of assessment criteria.

Lithuania

Example expert assessment

Expert assessment

Lesson plan sets simple and measurable learning objectives aligned with the proposed activities. Students' work patterns usually follow the pattern of assigning work to small groups but do not describe the collaborative techniques used by the groups. However, the activities of creating, evaluating and self-evaluating of group work are fully described with the simultaneous provision of the tools to be used.

Suggestions for improvement

- In activities where students record information from web pages, the recording criteria can be specified thus discouraging the simple copying and pasting of texts and images.

- Collaborative tools for recording information from various web sites could be used (eg text document sharing, shared presentation, padlet, etc.).

- A tool could be used to reduce the size of the proposed hyperlinks to make the text more readable (eg https://tiny.cc, https://www.shorturl.at etc.).

Austria

10. COURSE PARTICIPANTS' APPRECIATION OF BOTH ASSESSMENTS

Key results

1. Assessments by both experts and peers were largely perceived as both useful and fair, with higher agreement rates on the fairness of peer assessment.

2. Replies to the questions about their usefulness in general *(see section 10.1)* and their preferences for assessment in future courses *(see section 10.3)* did not reveal any clear preference for expert or peer assessment.

3. However, when asked about the usefulness of the individual assessment activities for their learning, 72.06% found expert assessment very useful and 65.22% assessing the lesson plan of their peers; compared to only 50.7% that found the assessment of their peers very useful.

The fourth research question of this report was how both the summative assessment and qualitative feedback were appreciated by the participating teachers and student teachers. Of the 103 lesson plan authors invited to fill in a short survey to that end, 73 replied. They were asked about their views on the assessments they received in online courses and how they valued their learning in the course in general.

10.1. PERCEPTION OF EXPERT AND PEER ASSESSMENT AS USEFUL AND FAIR

Figure 10 illustrates that lesson plan authors appreciated both the peer and expert assessments[1]. As regards the usefulness, agreement was 84% for both expert and peer assessment. 96% of lesson plan authors perceived the peer assessment as fair, compared to only 79% for expert assessments.

The largest majority of teachers agreed with the assessment they received

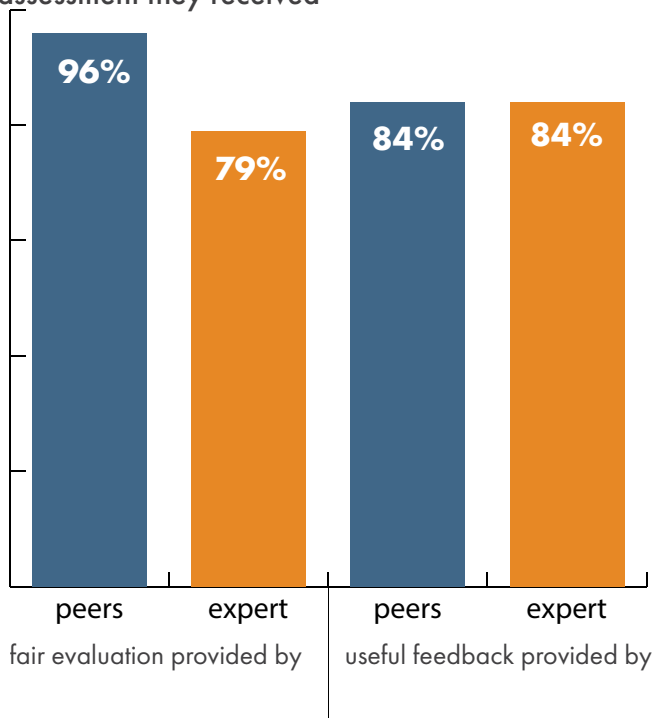


Figure 10: *Percentage of agreement with statements concerning the assessment by peers and external experts in the course*

From the results of this survey, it is not possible to say why the peer assessment was perceived as fair by more lesson plan authors than the expert assessment. However, it is plausible that the slightly lower scores and less positive feedback of expert assessment, compared to peer assessment, might have played a role. Another factor that may have played a role is that the Turkish external evaluator had formulated their feedback in a way that may have been perceived as less useful. Only 68%[2] of lesson plan authors found their assessment by the external evaluator fair,

1 *The assessment by the external evaluator was a fair evaluation of my work. The assessment by the external evaluator provided useful feedback (detailed enough/constructive). Answer options: 'Disagree', 'more disagree than agree', 'neither agree nor disagree', 'agree'). – the same question was asked for peer assessment.*

2 *17 survey respondents*

compared to 84%[3] of lesson plan authors from the other TeachUP countries. There may also have been an issue of culture and of levels of trust. Unlike the peer assessment, the expert assessment was anonymous. Finally, possibly also the direct comparison between the peer and expert assessments of their lesson plan may have played a role, as lesson plan authors received both. This finding is interesting, however, not conclusive and would require further investigation.

10.2. PERCEPTION OF THE USEFULNESS OF INDIVIDUAL ASSESSMENT ACTIVITIES

Figure 11 shows as how useful lesson plan authors rated each of the activities for their own learning *(from 'not all useful' to 'very useful')*[4]. All activities were perceived by all lesson plan authors as at least 'somewhat useful'.

Almost all respondents *(96% and 95%)* found writing their own lesson plan and their lesson plan being assessed by an expert as at least 'overall useful' for their learning. Most respondents also found assessing the lesson plans of their peers *(91.3%)*, self-assessing their own lesson plan *(89.6%)*, and having their lesson plan assessed by peers *(88.4%)* at least 'overall useful'.

When focusing only on the share of participants that rated an activity as 'very useful', some differences in the appreciation of the different activities become apparent. The expert assessment of their lesson plan and writing their own lesson plan was perceived as 'very useful' by *7* out of 10 respondents. Quite interestingly, both assessing the lesson plan of their peers *(65.22%)* and self-assessing their own lesson plan *(55.22%)* were perceived as 'very useful' – more so than having their lesson plan assessed by peers *(50.7%)*. Perhaps, more people found expert assessment very useful because the qualitative feedback experts provided was longer and slightly more constructive. The findings also highlight the important learning experience that being the assessor of another lesson plan can provide, with almost two out of three peers regarding it as 'very useful' *(see section 5.1)*. This finding confirms previous research that highlights that both the assessor and the assessee role are important in peer assessment (Li, Liu and Zhou (2012). The additional learning experience provided through assessment of others' course work is a unique advantage of peer assessment.

The majority of the individual assessment activities were perceived by all lesson plan authors as overall useful or very useful

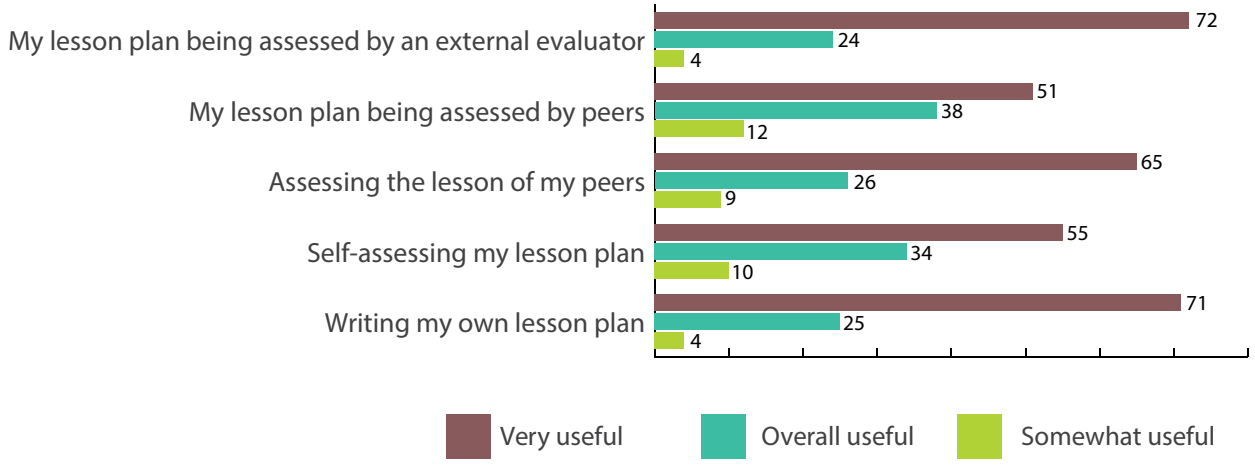


Figure 11: *Percentage of teachers finding the received overall assessment as useful for their learning*

3 *38 survey respondents*

4 *Please rate for each of the activities below, how useful they were for your own learning. If you did not do the self-assessment of your Lesson Plan, simply select 'not applicable'. Answer options: 'Writing my own Lesson Plan', ;Self-assessing my Lesson Plan (optional activity)', 'Assessing the Lesson Plans of my peers', 'My Lesson Plan being assessed by peers (overall)', 'My Lesson Plan being assessed by an external evaluator' ('not at all useful', 'somewhat useful', 'overall useful', 'very useful', 'I cannot say', 'not applicable').*

10.3. PERCEPTION OF USEFULNESS OF EXPERT AND PEER ASSESSMENT FOR FUTURE COURSES

Fig. 12 illustrates that the majority of 71 lesson plan authors that filled in the survey did not express a clear preference which assessment[1] form – expert or peer - they would find more useful for their learning in future courses *(77%)*. 11% found peer assessment more useful, and 12% found expert assessment more useful. The average score of those that preferred peer assessment was with 3.5 slightly lower than for those with a preference for expert assessment *(3.8)* and those with no preference *(3.7)*.

The majority of respondents did have a clear preference for either assessment form

Figure 12: *Preferred assessment approach for future online courses*

Overall, while peer and expert assessment were both largely appreciated by lesson plan authors, small differences in how both assessments were perceived became apparent. This might suggest that perhaps both assessment forms may have their unique advantages in the eyes of participants and appear as complementary.

> "Both evaluation methods (peer-to-peer and external evaluator) have struck me as very useful. Peer evaluation because it allows to compare the own didactic unit with that of other teachers, and the evaluation of the external evaluator because it is a much more specific evaluation than the previous one and, therefore, even more useful." Course participant, Spain

Course participants overall highly appreciated the TeachUP online course on collaborative learning, with a mean score of 4.4 *(with 1 being the lowest and 5 being the highest)*. There was no evidence that the peer and expert assessment scores had an impact on course participants' overall appreciation of the courses.

1 *For other online courses that you might take in the future, which assessment method (by peers vs. by an external evaluator) would be more useful for your learning? Answer options: 'The peer assessment (including giving and receiving feedback)', 'The external evaluator assessment', 'Both assessments are equally useful', 'I cannot say'*

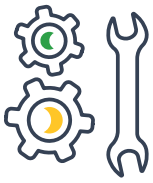
11. RECOMMENDATIONS

Peer assessment in TeachUP appears as a viable option for assessment, and may support the scalability of large online courses. Key recommendations emerging from the research are to:

Boost assessment cultures. Training for course participants on how to provide and receive feedback, in general and in online settings in particular, should be envisaged. The benefits of peer assessment both in online teacher professional development and with students should be further promoted. Guidance on peer learning activities should be accessible and engaging.

Emphasise the learning gain of being assessor of others and of peer dialogue. Assessing the work of others was identified as a particularly enriching learning experience. To strengthen this aspect, learning opportunities for peer assessors related to their own professional development, should be emphasised. Further, it should be possible for those assessed to respond to feedback received, thereby creating a dialogue between peers based on their concrete coursework *(on, for example, respective teaching beliefs)*. Ideally, this dialogue could also include their actual implementation of new practices they learned about during the course with their students. To that end, the online course would need to allow for sufficient time to try out new practices in the classroom and provide feedback opportunities for participants. Guidance on how to provide effective feedback *(specificity of suggestions, tone, and so on)*, would also be valuable.

Enhance the reliability of peer assessment. The assessments provided by three peers were generally consistent with ratings provided by experts *(demonstrating inter-rater reliability)*. This underlines the importance of providing a well-designed assessment tool, such as a rubric, setting out clear standards and criteria with descriptors and exemplars of work at different performance levels. Inter-rater reliability may be further enhanced through training on the use of the assessment tools and to ensure a shared understanding of performance levels. Other approaches may include: opportunities for peers to assess a subset of the assignment and discuss any discrepancies before assessing the remainder of the assignment: and/or combining peer assessment with expert assessment for a small random selection of the course assignments. Moreover, further research to better understand significant variability would be useful in ensuring the quality of the assessment tool, identifying needs for further training, and helping to reinforce the peer assessment as an appropriate approach for large-scale online learning.



Develop quality frameworks for peer assessment in scalable online teacher training courses. Given that assessments provided by three peers were generally consistent with ratings provided by experts, peer assessment can function as a reliable way to validate and certify teachers' progress when appropriate and effective assessment processes, tools, and guidance are in place. While further investigation is necessary to corroborate and understand better the results, the fact that many scalable online courses for teachers use peer assessment for this validation is not sufficient reason in itself to withhold accreditation of these courses. Of course, if the validation of learning/progress occurs on scalable online courses using peer assessment depends significantly on the design of the peer assessment processes, tools, and guidance provided. Therefore, and to facilitate the accreditation of scalable online courses for teachers, quality frameworks for the use of peer assessment in these courses should be developed. The use of such framework would provide accrediting organisations a tool to better evaluate the use of peer assessment in scalable online courses and accordingly facilitate the process of accreditation.

Enhance the quality of peer feedback. As experts' qualitative feedback was slightly more constructive and more detailed than that of peers, it would be useful to enhance the quality of peer feedback in online courses, for example, by providing examples of expert feedback, research findings about giving effective feedback, and opportunities to practise drafting feedback to fellow teachers. This could help course participants provide more constructive and detailed feedback in the peer assessment process.

Select the assessment approach depending on the focus and scale of the course. The findings suggest that both assessment forms may have their unique advantages in the eyes of participants and appear as complementary. Online course providers relying only on either peer or expert assessment could accordingly find mechanisms that allow for both types of assessment to work alongside each other. Peer assessment in the TeachUP context did not only serve the purpose of validating the learning of participants, but was also designed to facilitate community building and professional exchange, as well as help participants learn about assessment processes for use in their own classroom. Both of these purposes are more difficult to address with expert assessment. For online courses, however, with a focus on introducing new and complex content or practices, some element of expert assessment might still be useful. While integrating expert assessment is difficult in scalable contexts, an optional and paid-for-offer of expert assessment could be offered alongside peer assessment to participants who are looking for more substantive and constructive qualitative feedback on their work. Once a stronger culture of peer assessment is cultivated in the teaching profession, such additional expert assessment might not be needed anymore.

12. CONCLUSION

The TeachUP policy experimentation investigated whether peer assessment could provide an appropriate alternative to expert assessment. The assessments included numerical scores, as well as qualitative feedback. The results are encouraging in three ways.

First, inter-rater reliability for peer scores is high. Peer scores were, however, consistently slightly higher than expert scores. Second, peers' and experts' qualitative feedback was similar. Third, assessments by both experts and peers were perceived as both useful and fair, with higher agreement rates on the fairness of peer assessment.

In sum, peer assessment is a potentially viable approach to assess learning achievements in scalable online courses. While peer and expert assessment were both appreciated by course participants, small differences in how both assessments were perceived became apparent. This might suggest that perhaps both assessment forms may have their unique advantages in the eyes of participants and appear as complementary. Online course providers relying only on either peer or expert assessment could accordingly find mechanisms that allow for both types of assessment to work alongside each other.

Nevertheless, more research is needed. For assessments to fulfil their purpose, they need to be both valid and reliable. This research focused on the aspect of scores' reliability. Whether the assessment provided by peers was valid is a separate issue outside the scope of this research and would necessitate further investigation.

Further research might also be needed on aspects such as how to further strengthen the purpose of peer assessment being a useful learning experience for course participants, and in particular, how to further improve peer assessment mechanisms, in view of possible accreditation of online courses.

Moreover, it would also be of interest to compare peer and expert feedback in terms of their effectiveness for short-and longer-term learning gains. This may indicate a need to further refine the scoring rubric and guidance to ensure a shared understanding of the criteria for different performance levels. Alternatively, a system might be created for peers to assess a subset of the assignment and discuss any discrepancies among them before assessing the remainder of the assignment.

The TeachUP findings support policy moves to increase the provision of online teacher professional development with evidence that peer assessment can be both viable and cost-effective and they provide evidence on how to design peer assessment in online courses.

13. GLOSSARY

Connectivism: a learning theory which posits that in a digital age where knowledge is stored and readily available, learning is less about the acquisition of knowledge and more about the process of creating connections to people and content and being able to use and navigate these connections to access the right knowledge when needed.

Digital badge: a digital icon (= symbol or picture) or title that shows you have achieved something in an educational context. In the TeachUP project a digital badge was awarded upon successful completion of a course.

Digital certificate: a digital certificate showing that you have completed or achieved something in an educational context. In the TeachUP project a digital certificate was awarded upon successful completion of a course.

European Schoolnet Academy: platform offering MOOCs for school teachers and other school practitioners. It is run by European Schoolnet, the network of 34 European Ministries of Education, based in Brussels.

Expert assessment: assessment conducted by individuals with extensive knowledge and skills in the area being assessed

Inter-rater reliability: whether final scores are consistent between and among the different assessors (or raters)

KA3: Key Action 3 provides grants for a wide variety of actions aimed at stimulating innovative policy development, policy dialogue and implementation, and the exchange of knowledge in the fields of education, training and youth, under the Erasmus+ programme of the European Commission. Two main instruments are managed through specific calls for proposals: Initiatives for policy innovation giving support to forward-looking cooperation projects on policy developments, and European policy experimentations led by high level organisations and public authorities to stimulate innovative policies and prepare their implementation.

MOOCs: Massive Open Online Courses are free online courses available for anyone to enroll. Accordingly, they are designed for scalability and can accommodate large numbers of participants.

Peer assessment: In peer assessment, a collaborative learning technique, students evaluate their peers' work and have their work evaluated by peers. Often used as a learning tool, peer assessment gives students feedback on the quality of their work, often with ideas and strategies for improvement. At the same time, evaluating peers' work can enhance the evaluators' own learning and self-confidence. Peer involvement personalizes the learning experience, potentially motivating continued learning. When used in grading, peer assessment can give the instructor needed information on student performance. Especially for large online classes, it may allow inclusion of assignments where students' creative work could not be graded reliably through automation or efficiently by teaching staff.

Policy experimentation: initiative that helps ministries and government departments test new ways to solve policy problems within a limited scale, and within a set timeframe

Reliability: refers to the consistency and stability of results across student populations or across assessors. Reliability also means that assessments accurately measures student performance and knowledge.

Scalable online learning environments: The term "scalable online learning environments" refers to any environment that is designed in such a way that there is no practical, technical, or other limit to the number of learners in the environment. While such environment has the potential to accommodate "massive" numbers of learners, it does not necessarily do so. Massive open online courses or MOOCs would be considered a typical example of such an environment – even though the use of the term "massive" could be misleading in this context as numbers of learners are not necessarily high in numbers. Another example would be mobile learning applications like Edupills or Babbel or social media environments like a Facebook Group.

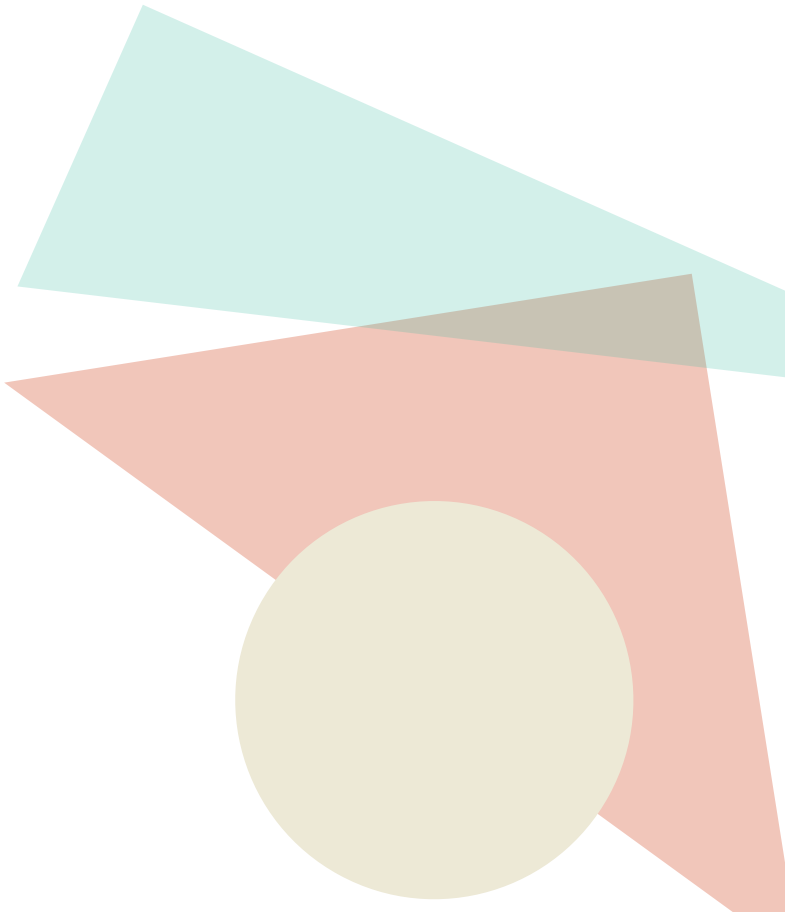
Scalable online courses: As "scalable online learning environments" but just referring to courses.

Social constructivism: a learning theory which posits that learners actively attempt to create meaning from experience and that the process of creating meaning is a social process shaped by interactions with others.

TeachUP teachers: Student teachers and professional teachers who signed up to TeachUP and enrolled in at least one course.

Teacher: either student or professional teacher.

Validity: refers to the degree to which assessments and evaluations measure what they are intended to measure (i.e. how well they are aligned with standards and curriculum).



14. BIBLIOGRAPHY

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. *(1985, 1999, 2014). Standards for educational and psychological testing. Washington, DC.*

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. *(1966, 1974). Standards for educational and psychological tests and manuals. Washington, DC.*

Annis, L. F. (1983). The processes and effects of peer tutoring. Human Learning: *Journal of Practical Research & Applications.*

Bachelet, R., Zongo, D., & Bourelle, A. (2015). Does peer grading work? How to implement and improve it? Comparing instructor and peer assessment in MOOC GdP. Paper delivered at *European MOOCs Stakeholders Summit 2015, May 2015, Mons, Belgium.*

Black, P. and D. Wiliam (1998). Assessment and Classroom Learning, Assessment in Education: Principles, *Policy and Practice, CARFAX, Oxfordshire, Vol. 5, No. 1, pp. 7-74.*

Boulet, M., G. Simard and D. Demelo (1990). Formative Evaluation Effects on Learning Music, *Journal of Educational Research, Vol. 84, pp 119-125.*

Butler, R, (1998). Enhancing and Undermining Intrinsic Motivation: The Effects of Task-involving and Ego-involving Evaluation on Interest and Performance, *British Journal of Educational Psychology, Vol. 58, pp. 1-14.*

Caldwell, C., C.G. Thorton and L.M. Gruys (2003), Ten Classic Assessment Center Errors: Challenges to Selection Validity, *Public Personnel Management, Vol. 32, pp. 73-88.*

Chong, W. H., & Kong, C. A. (2012). Teacher collaborative learning and teacher self-efficacy: The case of lesson study. *Journal of Experimental Education, 80(3), 263–283.*

Falchikov, N. (2013). Improving assessment through student involvement: Practical solutions for aiding learning in higher and further education. In *Improving Assessment through Student Involvement: Practical Solutions for Aiding Learning in Higher and Further Education.*

Iqbal, Z., & Mahmood, N. (2008). Compatibility of peer assessment and teacher assessment in observational situations: An emerging assessment tool in higher education. *Bulletin of Education and Research, 30(2), 61-77.*

Kilic, G. B., & Cakan, M. (2007). Peer assessment of elementary science teaching skills. *Journal of Science Teacher Education, 18(1), 91-107.*

Kobayashi, M. (2020). Does anonymity matter? Examining quality of online peer assessment and students' attitudes. *Australian Journal of Educational Technology.*

Kurtuldu, E., & Özkan, Y. (2019). Pre-service language teachers' reflection on peer assessment in *Micro teaching of a methodology course. 276–284.*

Laurillard, D. (2016). The educational problem that MOOCs could solve: Professional development for teachers of disadvantaged students. *Research in Learning Technology, 24(1), 29369.*

Li, L., Liu, L. X., & Zhou, Y. (2012). Give and take: A re-analysis of assessor and assessee's roles in technology-facilitated peer assessment. *British Journal of Educational Technology, 43(3), 376–384.*

Looney, J.W. (2011), Integrating Formative and Summative Assessments: Progress toward a Seamless System?, *OECD Education Working Papers, No. 58, OECD, Paris.* doi: 10.1787/5kghx3kbl734-en

Looney, J. (2015). D2.1 Literature Review. *Online self-assessmeMENTEP+D2.1+Literature+review+on+online+self+assessment.pdf/ee6120af-6849-4cb8-b464-b434df8efba0*

Looney, J.W. (2011), Integrating Formative and Summative Assessments: Progress toward a Seamless System?, *OECD Education Working Papers, No. 58, OECD, Paris.* doi: 10.1787/5kghx3kbl734-en

McGarr, O., & Clifford, A. M. (2013). Just enough to make you take it seriously: Exploring students' attitudes towards peer assessment. *Higher Education.*

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

OECD. (2019). TALIS 2018 Results (Volume I). *Paris: OECD.*

Ratminingsih, N. M., Artini, L. P., & Padmadewi, N. N. (2017). Incorporating self and peer assessment in reflective teaching practices. *International Journal of Instruction.*

Schleicher, A. (2016). Teaching Excellence through Professional Learning and Policy Reform.

Struyven, K., Dochy, F., & Janssens, S. (2008). The effects of hands-on experience on students' preferences for assessment methods. *Journal of Teacher Education.*

PROJECT COORDINATOR

European Schoolnet
Transforming education in Europe

europeanschoolnet.org
Belgium

RESEARCH ORGANISATION

IRVAPP
RESEARCH INSTITUTE FOR THE EVALUATION OF PUBLIC POLICIES
FONDAZIONE BRUNO KESSLER

irvapp.fbk.eu
Italy

PARTNERS

BMB Bundesministerium für Bildung
bmbwf.gv.at
Austria

HITSA Information Technology Foundation for Education
hitsa.ee
Estonia

CTI "DIOPHANTUS"
cti.gr
Greece

EDUCATIONAL AUTHORITY
oktatas.hu
Hungary

VYTAUTO DIDŽIOJO UNIVERSITETAS
vdu.lt
Lithuania

NACIONALINĖ ŠVIETIMO AGENTŪRA
nsa.smm.lt
Lithuania

MINISTRY FOR EDUCATION AND EMPLOYMENT
education.gov.mt
Malta

CFAE centro-oeste
cfaecentro-oeste.pt
Portugal

REPÚBLICA PORTUGUESA | EDUCAÇÃO
direção-geral educação
dge.mec.pt
Portugal

Universidade do Minho
uminho.pt
Portugal

mpc METODICKO-PEDAGOGICKÉ CENTRUM
mpc-edu.sk
Slovakia

UNIVERZITA ŽILINSKÁ V ŽILINE
uniza.sk
Slovakia

GOBIERNO DE ESPAÑA | MINISTERIO DE EDUCACIÓN Y FORMACIÓN PROFESIONAL
intef INSTITUTO NACIONAL DE TECNOLOGÍAS EDUCATIVAS Y DE FORMACIÓN DEL PROFESORADO
intef.es
Spain

yegitek.meb.gov.tr
Turkey

gtc SCOTLAND
gtcs.org.uk
United Kingdom

teachup.eun.org

Co-funded by the Erasmus+ Programme of the European Union