García-Loro F., Martín, S., Ruipérez-Valiente, J.A., San Cristóbal, E. and Castro, M. (2020). Reviewing and analyzing peer review Inter-Rater Reliability in a MOOC platform. ***Computers & Education*****, 103894.**

https://www.sciencedirect.com/science/article/abs/pii/S0360131520300932

DOI: 10.1016/j.compedu.2020.103894

Reviewing and Analyzing Peer Review Inter-Rater Reliability in a MOOC Platform

Felix Garcia-Loro[a], Sergio Martin[a], Jose A. Ruiperez-Valiente[b], Elio San Cristobal[a], & Manuel Castro[a]

[a]Spanish University for Distance Education, of Electrical and Computer Engineering Department. Juan del Rosal 12, 28040 Madrid, Spain.

[b]Massachusetts Institute of Technology; Department of Comparative Media Studies. 77 Massachusetts Ave, Cambridge (MA) 02139, USA.

ABSTRACT

Peer assessment activities might be one of the few personalized assessment alternatives to the implementation of auto-graded activities at scale in Massive Open Online Course (MOOC) environments. However, teacher's motivation to implement peer assessment activities in their courses might go beyond the most straightforward goal (i.e., assessment), as peer assessment activities also have other side benefits, such as showing evidence and enhancing the critical thinking, comprehension or writing capabilities of students. However, one of the main drawbacks of implementing peer review activities, especially when the scoring is meant to be used as part of the summative assessment, is that it adds a high degree of uncertainty to the grades. Motivated by this issue, this paper analyses the reliability of all the peer assessment activities performed as part of the MOOC platform of the Spanish University for Distance Education (UNED) UNED-COMA. The following study has analyzed 63 peer assessment activities from the different courses in the platform, and includes a total of 27,745 validated tasks and 93,334 peer reviews. Based on the Krippendorff's alpha statistic, which measures the agreement reached between the reviewers, the results obtained clearly point out the low reliability, and therefore, the low validity of this dataset of peer reviews. We did not find that factors such as the topic of the course, number of raters or number of criteria to be evaluated had a significant effect on reliability. We compare our results with other studies, discuss about the potential implications of this low reliability for summative assessment, and provide some recommendations to maximize the benefit of implementing peer activities in online courses.

KEYWORDS

Peer assessment, MOOCs, Krippendorff's alpha, Inter-Rater Reliability (IRR), Reliability.

1. INTRODUCTION

Last century's recent changes on educational paradigms have promoted the integration of new evaluation methods that intend to advance beyond the classical knowledge assessment (summative assessment) as its only grading goal. This new mindset aims to develop evaluation methods that are more embedded within the training and learning process in what is known as formative assessment (Dochy, Segers, & Sluijsmans, 1999; Earle, 2014; Guan-Yu Lin, 2018). Formative assessment can have a significant impact on the quality of learning that students experience by practicing the required skills in advance, and by helping them to be more self-aware of their current status, but also for instructors so that they can have just-in-time feedback regarding how the class is progressing (Van der Pol, Van den Berg, Admiraal & Simons, 2008; Topping, 2017). In fact, assessment is now conceived as a central part of the learning process, of which the student has become more responsible (Dochy & McDowell, 1997; Dochy et al., 1999;

Black & Wiliam, 2009; Kilic, 2016). This new paradigm can be interpreted as trying to shift from the consolidated idea of assessment as the final goal of the learning process, to a paradigm where assessment is just one of the many tools and options (Sluijsmans, Brand-Gruwel, van Merriënboer, & Bastiaens, 2002b). Furthermore, in today's society where information is easily available and where AI is called to take over tasks that are easy to automate, higher education institutions have acknowledged the need to train students to develop more transverse skills, given that they will face a more and more uncertain future carrying out work responsibilities that might still not exist (Marton & Bowden, 1999; Boud, 2000; Susskind & Susskind, 2015).

From the very beginning, the European Higher Education Area (AHEA) has been watching over the implications of this on-going educational shift. However, it did not start talking about student-centered learning until 2009, in a meeting which took place at Leuven/Louvain-la-Neuve (EHEA, 2009). Besides, AHEA's present educational model is based on competences (de Miguel, Alfaro, Apodaca, Arias, García, & Lobato, 2005), and so the current speech is focusing now on 'competence alignment' or 'constructive alignment'. The new emphasis on student-centered learning and competences, together with the Information and Communications Technology (ICT) democracy, has facilitated the creation of new pedagogical approaches or boosted the use of underused ones, by promoting a redesign of the learning scenario (Beldarrain, 2006); some examples that have received a lot of attention include collaborative learning (Van Den Bossche, Gijselaers, Segers, & Kirschner, 2006), self-regulated learning (Boekaerts & Corno, 2005), collaborative inquiry learning (Bell, Urhahne, Schanze, & Ploetzner, 2010), competence-based learning (Benlloch-Dualde & Blanc-Clavero, 2007), personalized learning (Chen, 2008), differentiated learning (Lawrence-Brown, 2004), active learning (Gauci, Dantas, Williams & Kemm, 2009), flipped learning (Lukassen, Pedersen, Nielsen, Wahl, & Sorensen, 2014), instructional scaffolding (Quintana, Reiser, Davis, Krajcik, Fretz, Duncan et al., 2004), problem-oriented and project-based learning (Lehmann, Christensen, Du & Thrane), and so on. These approaches can be combined in order to achieve an effective metacognitive learning that can prepare better students for efficient lifelong learning (Cornford, 2002; Weinstein, Acee, & Jung, 2011; Lüftenegger, Schober, van de Schoot, Wagner, Finsterwald & Spiel, 2012). It is with the implementation of these new methodologies that evaluation has ceased being an isolated activity carried out at the end of the learning process and it is now frequently integrated more seamlessly in the learning process, and it is regarded as yet another tool for its success. According to Delgado, Borge, García-Albero & Salomón (2005), evaluation now intends to assess the quality of learning the student has developed; it is no longer based on products, but rather, on processes.

One of the tools favored by the new perspectives on educational plans has been peer assessment or peer review tasks. In this sense, Falchikov & Goldfinch consider that "peer assessment is grounded in philosophies of active learning and andragogy, and may also be seen as being a manifestation of social constructionism, often involves the joint construction of knowledge through discourse". According to Duran (2016) "the first reviews and meta-analyses on peer tutoring revealed evidence of learning by the tutor in their role of 'teacher'". Moerkerke (1996) and Dochy *et al.*, (1999) share the idea that peer assessment activities are compatible with a society of lifelong learners.

The area of learning at scale presents massive online scenarios, such as MOOCs among others, that require alternative approaches in order to implement learning and assessment approaches that target many learners at the same time . In order to provide a learning design that is sustainable and can scale to large numbers of learners, formative assessment cannot be dependent on direct feedback from teachers. Therefore, for those classes where formative assessment is a crucial part of the learning process, peer assessment turns into a tool with huge potential to solve the issue of scale. This article analyses the reliability of peer assessments developed specifically under MOOC environments. It focuses on the consistency of students as raters, by studying Inter-Rater Reliability (IRR). In addition, we aim to assess the validity of the obtained evaluations in our specific framework, taking into account our limitations. For these analyses, we have gathered the data of all the peer assessment activities carried out on UNED's MOOC platform (http://coma.uned.es/). These courses are highly diverse, being related to different knowledge areas, subjects and levels (Capdevila & Aranzadi, 2014). MOOCs have proved to be successful non-formal open learning environments (Hood, Littlejohn & Milligan, 2015), where students' motivation and self-regulation capabilities are key factors. For those reasons, MOOCs are an optimal resource for knowledge transference in our current society. Nevertheless, and in spite of the many developments on virtual tutoring, the massive nature of MOOCs limits the type of activities that can be implemented. Specifically, activities that do not scale to a high number of students (e.g., a teacher providing individualized feedback to each assignment), cannot be implemented in these environments (Suen, 2014). As many other learning activities, peer assessment generally implies receiving a score, which could potentially be used as part of the summative grade. Therefore, in this manuscript we explore the reliability and validity of scores generated through peer assessment activities, in order to evaluate whether it would be appropriate to use these scores as part of a weighted final grade. The data we analyze have been gathered based on the assessment that students performed on the activities of their peers. Both tasks, submitting an activity, and peer reviewing someone else's work, are mandatory on the platform. Consequently, our purpose was to obtain a data sample large enough to analyze the consistency of the assessments according to multiple observers in different courses and activities. For this purpose, we have collected a high number of valid submitted tasks (more than 27,000), reviews (more than 93,000) and criteria assessed (almost 334,000), conferring a solid background to the results and conclusions obtained in this analysis. Overall, the research question that has concerned us in this study is the following:

RQ: Are peer assessments reliable in a typical MOOC environment like the one provided by UNED platform?

2. *LITERATURE REVIEW*

Peer assessment can be described and implemented in many different ways. The number of studies and diversity of educational contexts suggest that peer assessment can be, practically, applied to all areas of knowledge (Topping, 1998). As an assessment approach, peer assessment has traditionally been considered valid or not, by confronting students' and teachers' grades (Stefani, 1994; Falchikov, 2000; Cho, Schunn & Wilson 2006; Sung, Chang, Chang & Yu, 2010; Jackson, 2014; Jones & Alcock, 2014; Formanek, Wenger, Buxner, Impey & Sonam, 2017), despite the fact that the core objective of peer assessment is to actually create opportunities for peers to learn from each other and to participate more in the learning process. This correction over students' evaluation has been called 'validity', while we use the term 'reliability' to determine the consistency among peer ratings (Richmond et al., 1992; Luo, Robinson & Park, 2014; Jackson, 2014).

This section is meant to frame peer assessments and, more specifically, their reliability. It does so by starting from a general point of view up to its specific impact on MOOCs.

2.1. *Definition of peer assessment*

Several authors have provided broad definitions, conceptually talking, for peer assessment. For example, Fachikov & Goldfinch (2016) highlight that, when students use them, they "judge the work of their peers". This view is similar to Reinholz (2016) although he talks about evaluating others. Orsmond, Merry & Reiling, (1996) refer to peer assessment as a learning tool and Van Zundert *et al*., (2010) focus their argument on its not necessarily bidirectional reciprocity. According to them, the goal is to "evaluate or be evaluated by peers". Topping (1998; 2009) includes the concept of learning through peer assessment in his definition: "Peer-assessment is an arrangement for learners to consider and specify the level, value, or quality of a product or performance of other equal-status learners". Van der Pol *et al*., (2008) provide a broad definition which includes every step carried out on peer assessments, described as an activity. They talk about the pre-established criteria that the student must stick to, as well as the requirements of a critical evaluation that includes feedback (formative assessment) for the evaluated student. In their words, "students engage in reflective criticism of the products of other students and provide them with feedback using previously defined criteria". De Grez Valcke & Roozen (2012) use the term 'peer assessment' on a test in which they invited students from a more advanced course to act as raters. To some extent, they might be considered peers, but this implementation misses the point where a student is rating a piece of work the student has already completed. Consequently, the cognitive process that involves personal reflection and self-criticism is lost.

On this paper, we consider as 'peers' the students of each course who are registered and active in each evaluated tasks. This implies that they all have carried out the task before engaging in the peer assessment activity. They find themselves in a position of equality towards the task and hence we can effectively consider them as peers based on the previous definitions provided by Topping (1998, 2009) and based on the idea of "other equal-status learners".

2.2. Peer assessment and its integration in MOOCs. Implications for reliability and validity

MOOCs usually implement assessment methods that do not require manual correction by the instructors, usually, these are generally known as auto-graded tools (machine-assessment): single choice and multiple-choice items are particularly common; as well as fill-in the blanks, with a number, a word or even a sentence. Other more nuanced auto-graded items include programing environments where students code their solution and the system expects an specific function output, or specific tools that can be integrated with the MOOC platform through authentication protocols such as LTI protocol (Alcarria, Bordel, Andres, & Robles, 2018; Garcia-Loro, San Cristobal, Diaz, Macho, Baizan, Blazquez et al., 2018; Mullen, Byun, Gadepally, Samsi, Reuther, & Kepner, 2017; Garcia-Loro, Sancristobal, Gil, Diaz, Castro, Albert-Gómez et al., 2016; Aleven, Sewall, Popescu, Xhakaj, Chand, Baker et al., 2015). There have also been some limited advances in auto-grading essays (Ambekar & Phatak, 2014). Auto-graded assessment instruments have high validity, but they are quite limited in what they can assess and the cognitive process of students solving them is very low, which can be especially critical in some areas of knowledge. In order to improve and support students' learning, it is essential to include feedback information that can help students understand where they are at in their learning process and their potential misconceptions.

Peer evaluation, besides the reliability and validity of its methodology, can provide this sort of beneficial personalized feedback to every single one of the otherwise unmanageable number of students in MOOCs. Furthermore, it is a well-aligned contribution to the current educational perspectives that locate the student in the center of the whole learning process (van Hattum-

Janssen & Lourenço, 2008; Suen, 2014). Finally, the exercise of acting as an evaluator can enact more complex cognitive processes that favor deeper learning for students (Hsia, Huang & Hwang, 2016).

With regard to the typical learning environments in MOOCs, while traditional learning contexts can assume a high similarity degree in the background of their learners, the 'Open' nature of MOOCs highly increments the diversity in learners' profiles, hence potentially breaking the equality among learners' condition. In MOOCs we find that learners have multiple backgrounds in content knowledge (especially those regarding STEM), diverse sets of skills related to writing, text comprehension, synthesis and very different intentions when enrolling in a MOOC (Alario-Hoyos, Pérez-Sanagustín, Delgado-Kloos, Parada, & Muñoz-Organero, 2014; Watson, Watson, Yu, Alamri, & Mueller, 2017). This characteristic heterogeneity in students' profiles collides even more with the assumption of equity among peers.

Feedback

Feedback is undoubtedly the core mechanism in peer assessment to become formative (Thelwall, 2000; Gipps, 2005; Miller, 2006; Nicol & Macfarlane-Dick, 2009; Ng, 2014). When correctly implemented, peer assessment involves students in both feedback roles: as evaluators, by contributing with ideas and comments to the assessed tasks, as well as evaluatees, by receiving peers' observations with constructive comments to improve their own work (Ng, 2014). This sort of assessment usually coexists with the summative ones, although it can appear on its own. Nevertheless, it is recommended that formative assessment goes alongside with the summative one (Miller, 2009; Nicol & Macfarlane-Dick, 2006; Gipps, 2005). In this sense, Ng (2014) highlights the importance of students receiving tailored feedback instead of just receiving scores. Feedback and feedforward strategies are used in critical learning (Cartney, 2010; Kilic, 2016) as well as in social learning (Guan-Yu Lin, 2018). These tools stand out in peer assessment because they help the student develop analytical thinking, critical thinking and deeper knowledge development. However, students must be well prepared and highly motivated to be capable of developing this task (Winstone, Nash, Parker, & Rowntree, 2017). On the other hand, students enrolled in MOOCs tend to be from a broad spectrum of educational backgrounds, they can have diverse levels of initial knowledge, different intended learning objectives and different self-regulated learning patterns. Such diversity in MOOC students, and, therefore, in raters, can undermine the underlying assumption of "equality" in peer assessment methodologies (Meek, Blakemore & Marks, 2016).

Assessment criteria and rubrics

Dochy *et al.*, (1999) highlight the importance of establishing clear assessment criteria: "it should be clear that students have to know the criteria clearly… criteria should include information about the area to be assessed, the aims to be pursued and the standards to be reached". In this sense, Falchikov & Goldfinch (2000) in their meta-analysis have found that the reliability and validity of peer assessment is positively correlated with the establishment of a clear assessment criteria. They also found that peer assessment tasks requiring several independent scoring dimensions were less valid than peer assessment tasks based on a global judgement. In this context, Sadler & Good (2006) as well as Meletiadou & Tsagari (2014) stated that "five or fewer criteria increase reliability". Nonetheless, studies like the one carried by Jones & Alcock (2014) based on comparative judgment (Thurstone, 1927), consider that evaluation criteria are not a necessary condition for reliable and productive peer assessment; instead, they consider that students feel stimulated as raters if they have more freedom to develop their own assessments. Furthermore, it would further promote their abilities, critical thinking and sense of responsibility.

Although traditionally teachers' and experts' grades are considered as the valid ones (Stefani, 1994; Falchikov, 2000; Cho, Schunn & Wilson 2006; Sung, Chang, Chang & Yu, 2010; Jackson, 2014; Formanek, Wenger, Buxner, Impey & Sonam, 2017), authors such as Piech, Huang, Chen, Do, Ng & Koller (2013) state that the "true mark" is not necessarily the teachers' one; they propose to distance teacher's rubric and its validity. To avoid this dichotomy in the "true grade" (teachers' vs students' grading), and also to improve validity, several authors have highlighted the benefits of training in the reviewing mechanism (Sluijsmans *et al.*, 2002b; Sadler & Good, 2006; Topping 2009; Zundert, Sluijsmans & Merriënboer, 2010; Meletiadou & Tsagari, 2014; Topping, 2017; Formanek *et al.*, 2017). Furthermore, many studies have involved students in the definition and development of the assessment criteria in order to improve assessment results and students' involvement in the activity (Orsmond, Merry & Reiling 2000; Falchikov and Goldfinch 2000; Sluijsmans et al., 2002a; Liu &Carless, 2006; Falchikov, 2013; Leenknecht, & Prins. 2018).

Different approaches to assessment criteria do not necessarily imply different points of view on whether they should be applied or not to MOOCs, as opposed to traditional learning environments. The way in which MOOCs are implemented develops new ways of student-teacher-course interaction. Several authors (van Hattum-Janssen & Lourenço, 2008; Topping, 2009) point out the relevance of student implication and participation when designing evaluation criteria for peer assessment activities. Students get more involved in the task, and a two way path of understanding the activity is created. However, this proposal cannot be applied to MOOCs: (i) the 'open' nature of MOOCs brings together students with very different backgrounds and needs, and, consequently, with very different perspectives; and (ii) another common property of these courses is students' asynchrony when following the course. Student implication and participation in the design of criteria becomes complicated due to this factor. Strict submission dates can help overcome such issue. Many authors have highlighted the important effects of deadlines on formative actions that require feedback (Ng, 2014; Black & Williams, 2009; Epstein et al., 2002; Webb, Stock & McCarthy, 1994; Kulik & Kulik, 1988; McKeachie, Pintrich, Lin, & Smith, 1986). Feedback delays can cause formative evaluations to be useless. Some studies have addressed through experimentation that immediate feedback leads to better learning than a delayed one (Kehrer, Kelly & Heffernan, 2013). In this sense, MOOCs usually take place in fast paced contexts, and hence, deadlines times are usually tight.

Number of raters

The effect of the number of raters on peer assessment has been analyzed with different results depending on the study. Falchikov & Goldfich (2000:312) hold that "singletons do not appear to be less reliable than others", however they refer to reliability by analyzing its correlation with instructor grades (validity), instead of analyzing the reliability of the raters. They also suggest that a large number of raters may cause a diffusion of responsibility in reviewing tasks. However, this may be caused due to the consequent higher number of required reviews for each student and, therefore, promote boredom in the reviewing process. The studies of Cho, Schunn, & Wilson (2006), Kilic & Cakan (2007), Xiao & Lucking (2008), Sung *et al.*, (2010) and Chang, Liang & Chen (2013) found that reliability increases by increasing the number of raters. The results obtained in the study carried out by Kulkarni, Wei, Le, Chia, Papadopoulos, Cheng *et al.*, (2013) concluded that an increasing number of raters increases accuracy (they use accuracy to express the degree of proximity to the teachers'/experts' mark). To be more specific, the improvements experimented are decreasing as the number of reviewers increases following a logarithmic trend. In the model used by Li, Xiong, Zang & Mindy (2016) for their meta-analysis, the correlation between teachers' and peers' ratings was high for assignments with more than 10 reviewers, medium for assignments with 6 to 10 reviewers, and low for 5 or less reviewers. However, the results were not statistically significant at the 95% level.

Social factors

According to Topping (2009:24), "social processes can influence and contaminate the reliability and validity of peer assessments". Social factors such as friendship, aversion, popularity, conflict avoidance and so on are present in peer assessments (Friedman, Cox, & Maher, 2008; Topping 2009). They particularly show up when peer assessment activities are carried out on face to face methodologies. Therefore, these are not a critical factor in MOOCs due to geographical distance, online anonymization, and even because of asynchrony.

Onset education often chooses to keep the assessed tasks double-blinded (Ng, 2016). This is often the case in MOOCs, where users are only identified by the nickname or just the identification number that the platform assigns to each one. However, factors such as anxiety are present at any educational scenario for both reviewer and reviewee (Topping, 2017). MOOC anonymity and distance environments diminish the assessment subjectivity caused by these social factors. However, many others social factors, such as the inevitable sympathy towards peers, the use of a foreign language, different culture, economic factors, gender, etc. (Suen, 2014; Kizilcec, Davis & Cohen, 2017; Kizilcec, Saltarelli, Reich, & Cohen, 2017) cannot be avoided nor controlled.

Haynes, Smithe, Dysthe & Ludvigsen (2012) identified another factor that affects peer assessment marginally. They tested it in six different high schools in Norway. Students perceive feedback as more or less useful depending on the manners and the terms used as well as on the classroom's atmosphere. In this way, critical feedback is taken as constructive under the appropriate circumstances and a correct choice of words. Peer evaluation promotes this sort of contexts because students are often acquainted to each other. Furthermore, Hovardas, Tsivitanidou & Zacharia (2014) hold that peer feedback entails more improvements for learners than expert feedback. Initially, this factor does not affect the reliability or validity of the assessment process as it involves the way the students perceive the feedback in the assessment.

We can conclude that social factors can also play some role in MOOC peer assessment, since "peer assessment is a multifaceted process… affected by a number of psychological and personality traits" (AlFallay, 2004:419).

2.3. Measuring reliability in peer assessments

The core aim and benefit in peer assessment is the learning that students experience during the peer assessment process, both as assessors and assessees. However, summative assessment may be considered as a possibility in some cases. Traditionally, the resulting grades from peer assessment have been considered valid or not by confronting them with teacher's/expert's ratings (Stefani, 1994; Falchikov, 2000; Tsai, Lin & Yuan, 2002; Cho, Schunn & Wilson 2006; Kilic & Cakan, 2007; Sung *et al.*, 2010; Chang, Tseng & Lou, 2012; Sung *et al.*, 2014; Li, Xiong, Zang, Kornhaber, Lyu, Chung *et al.*, 2016; Formanek *et al.*, 2017). This comparison of students' evaluation with the teachers' ratings has been referred to as 'validity', while the term 'reliability' is used to determine the consistency among peer ratings (Richmond *et al.*, 1992;

Luo, Robinson & Park, 2014; Jackson, 2014). The results obtained in terms of validity and reliability of peer assessment vary from one study to another.

Cho, Schunn & Wilson (2006) point out that both reliability and validity studies always leave aside students' point of view, in favor of the teachers'. Students and teachers perceive reliability and validity differently: "the instructor can take into account the effective reliability of ratings generated by a set of peers, whereas each student is restricted to a consideration of the reliability of individual peer ratings"; hence, students' opinion is based on the criterion that "the greater the spread of grades, the less reliable".

No matter the rater or the group of raters chosen for a specific task, Hayes & Krippendorff (2007) talk about the inherent presence of the human condition: "When relying on human observers, researchers must worry about the quality of the data". Classical test theory is based on the assumption that every grade can be understood as the sum of 'true score' (Novick, 1966; Lord and Novick, 1968), this is, "the expectation of an individual's observed score" (Zimmerman *et al*., 2005), plus the error score.

The level of agreement or consistency among the evaluations or judgments carried out by the raters or 'graders' is known as IRR (Lavrakas, 2008; Lange, 2011). Krippendorff (2011) defines reliability as "the extent to which different methods, research results or people arrive to the same interpretations or facts". However, "reliability is only a prerequisite to validity. It cannot guarantee it" (Krippendorff, 2011). Raters' consistency is the most relevant factor when studying and analyzing reliability. Through reliability, we try to figure out if raters are consistent in their judgments or assessments, without taking into account the level of agreement they reach; "The consistency of a marker is more important than whether he or she disagrees with another marker" (Brown, Bull, & Pendlebury, 1997, p.235).

Hayes and Krippedorff (2007) claim that "choosing an index of reliability is complicated by the number of indexes that have been proposed". For starters, we should reject measuring IRR by means of percentages of agreement (Hallgren, 2012) because it ignores the level of agreement, in favor of a 'correct' or 'incorrect' evaluation. Information loss is therefore severe unless the analysis is limited to dichotomic, or even nominal, variables.

Pearson's Correlation Coefficient (PCC), also known as the "Product Moment Correlation Coefficient" (PMCC) has been used in several studies as an interrater reliability estimator (Cho, Schunn, & Wilson, 2006; Jones, & Wheadon, 2015; Ashenafi, 2015). Particularly, it has been applied to the analysis of quantitative variables in peer assessments. However, this coefficient, besides assuming a state of normality, can only be applied if the raters are only two and if they are in charge of assessing all participants. This measure is, therefore, not applicable in our case. Some studies have chosen to overcome the limitation in the number of raters by using Fleiss' kappa (Schaer, 2012; Raman & Joachims, 2014). In this way, they have managed to include more raters, but this measure can, once again, only be either dichotomic or nominal. Cohen's kappa (Cohen, 1960), which is a non-parametric test for qualitative variables, or Scott's pi (Scott, 1955), are some of the other statistical methods that have been used for IRR measuring (Lombard, Snyder-Duch, & April, 2004; Antoine, Villaneau, & Lefeuvre, 2014; Zapf, Castell, Morawietz & Karch 2016). The most common methodology found when studying reliability in peer evaluations is the Interclass Correlation Coefficient (ICC), or other derived versions from it (Cho, Schunn, & Wilson, 2006; Xiao & Lucking, 2008; Luo, Robinson, & Park, 2014; Shieh, 2016; Formanek *et al*., 2017; Yoon, Park, Myung, Moon, & Park, 2018). Its basic advantage is that it allows high flexibility on the number of raters per test. However, within our data collection, we have 63 distinct peer assessment activities from our platform, that sum up to

27,745 submitted tasks, with three or more raters in each task distributed across different courses. Furthermore, we find differences in the number of raters within each activity due to how the peer assessment is operationalized in the MOOC platform. For all this, ICC requirements do not match the properties of our sample.

Anyhow, Shout & Fleiss (1979) presented a statistical method similar to ICC which has already been used within the MOOC context by Luo, Robinson & Park (2014). The variability in the number of raters made the authors limit their ICC study to only those tests that had five raters. We consider that subsetting the data for an ICC statistical analysis based on the number of raters, clearly undermines the robustness and trustworthiness of the reliability analysis we want to conduct.

Krippendorff's alpha statistic (Krippendorff, 1970; 2011; 2018) provides a reliability measure based on the expected and the observed disagreement. This method comes along with a very high data flexibility: it works with two or more raters, and it does not require that every rater has evaluated every test (the statistic can handle missing values). Besides, it is applicable to all sorts of data types, like ordinal, interval or binary variables. Attending to the measurement scale in our case study, the requisites that the statistic must meet are any number of raters and the existence of missing data. Therefore, we decide to use in this article Krippendorff's alpha statistic to analyze peer assessment reliability in MOOCs for the reasons already given: i) we require a statistic that can handle more than two raters, ii) we require flexibility in the number of raters for each subset, iii) we require to handle missing values, and finally iv) we require a statistic able to deal with ratio variables.

3. METHODOLOGY

3.1. Context

UNED-COMA was developed under the open platform OpenMOOC (https://github.com/OpenMOOC) and integrated within the framework of OpenupED (https://www.openuped.eu). By the date when this study was conducted, there were 23 courses, from technical topics such as basic analytical chemistry or practice-based electrical/electronics circuits, to second language learning or focused on continuous training (Capdevila & Aranzadi, 2014; García-Loro, Díaz, Tawfik, Martín, Sancristobal & Castro, 2014). The platform also hosts Small Private Online Courses (SPOCs) targeting teachers. The platform has around 140k unique students and 220k enrolments in courses that have triggered more than 25k certification badges.

The structures and the activities designed by the Educational Boards (EBs) —to-do activities, questions, answers and evaluation criteria— can be found in PostgreSQL. Answers and student activities are recorded in MongoDB. Students' data are stored in a different DDBB tables, separated from the rest of the structure. Figure 1 depicts the structure we have just described. The different Postgres tables are nested through the fields shown in the arrows in Figure 1, except for the table of users, which is independent. Each activity provided by the platform is nested in the activity table. Figure 1 exclusively presents peer evaluation activities.

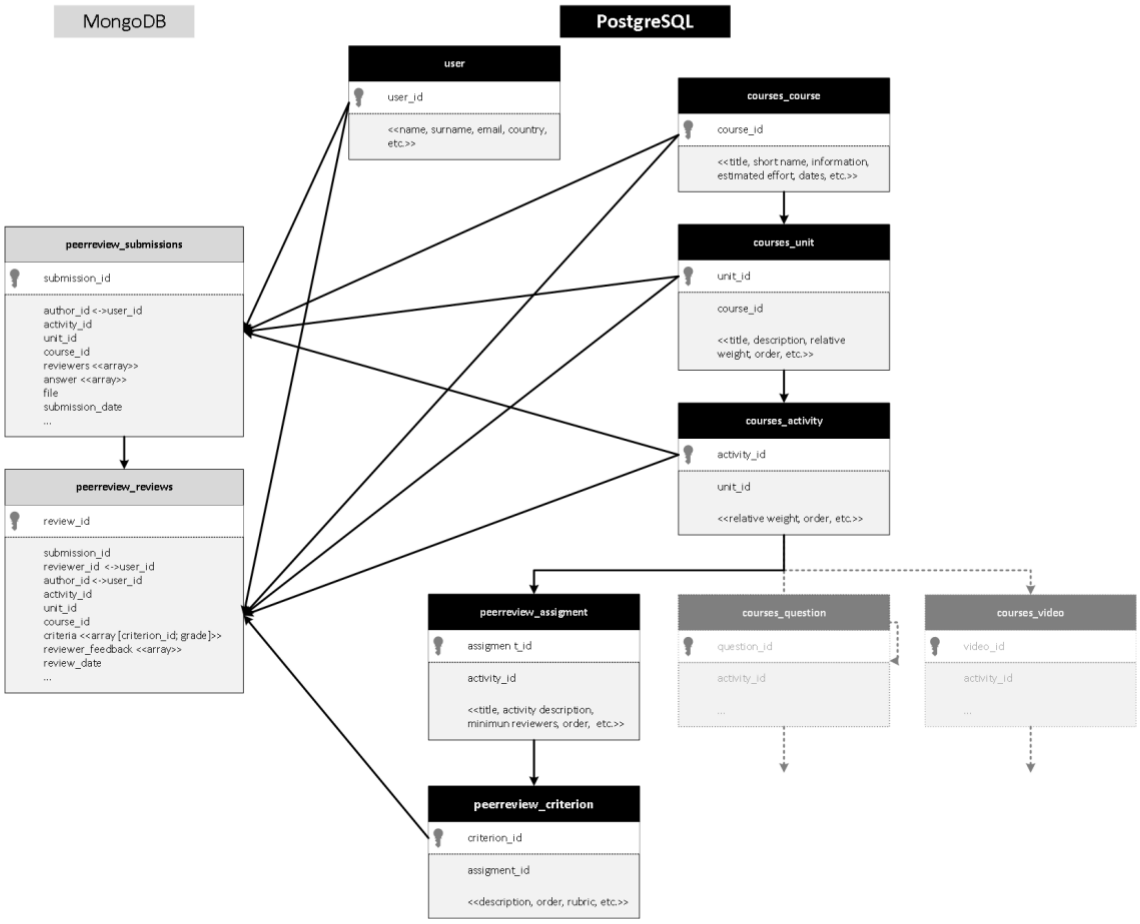


Figure 1. MongoDB and PostgreSQL joint data schema.

3.2. Peer assessment implementation

Peer evaluation activities on the platform are organized in the following two steps, which are also a requisite in order to consider the peer assessment activity as completed:

1. The student needs to upload the task developed to the platform. Strict deadlines are optional in this step.
2. The student needs to assess a minimum number of tasks from other peers. This number is fixed by the EB, and most of the times is around 3 reviews. However, they have no control on which tasks are assigned to which student since this process is automatically run by the matchmaking system of the platform.

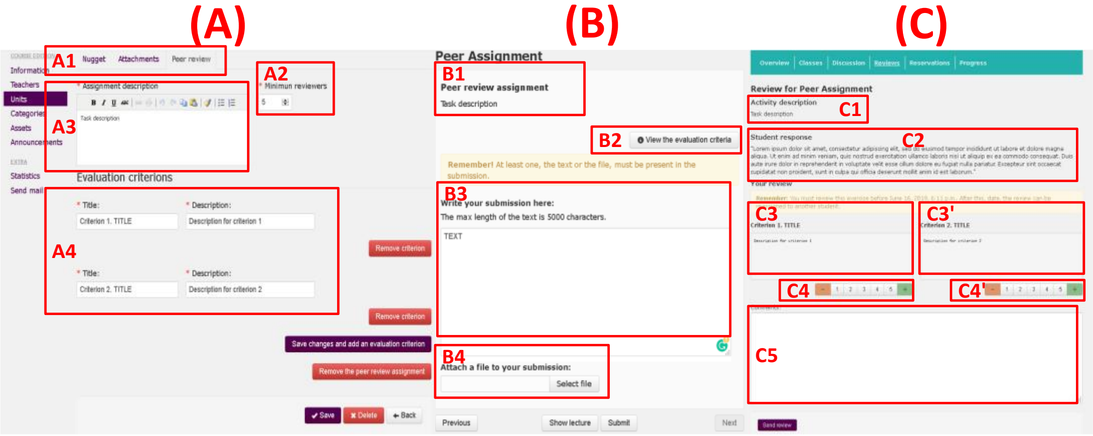
Once the student has completed both steps, the platform marks the task as completed by the student. Nevertheless, before the grading process can be finished, the students' assignment needs to be evaluated by a minimum number of students (fixed by the EB). Even if the student already completed both steps, they will need to wait until other students complete the evaluation of their own assignment.

The assessment of each task implies both a summative and a formative component. They both respond to the criteria previously set by the EB. The assessments provided to students can be classified into two types:

- Quantitative evaluation (summative assessment): The assignment is graded based on whole numbers from 1 to 5 (min and max respectively), according to evaluation criteria or rubrics, provided by the EB.

- Qualitative evaluation (formative assessment): the author of the task receives feedback written by the reviewer. It is implemented in an optional way on the platform.

The full process for a peer assessment activity is shown in Figure 2. Figure 2(A) shows the creation of a peer assessment task with the different settings that EBs may use: (A1) here the EB's may add additional contents for the activity, like a video or documents; (A2) this selection box is used to establish the minimum number of reviewers required; (A3) short description of the activity; and (A4) the definition of the criterion (title and short description) for each of the criteria to be assessed. Figure 2(B) shows the student interface to complete a peer assessment task: (B1) provides the short description provided by the EBs in (A3); meanwhile (B2) shows the criteria information provided by the EBs in (A4); (B3) and (B4) are the options provided by the platform to submit the answer, either as plain text (B3) or attaching a document (B4). Figure 2(C) shows the interface that a student sees when acting as a reviewer in a peer assessment. (C1) provides the description provided by the EB in the section (A3); (C2) is the answer provided by the student (plain text, no file attached); (C3) and (C3') are the criteria to be graded by the student, which was set up by the EB in (A4); (C4) and (C4') are the scale (1-5) to grade each criterion (in this example we have two criteria); (C5) is intended for the reviewer's written feedback.



* Assignment description

Task description

* Minimun reviewers

5

Evaluation criterions

* Title: Criterion 1. TITLE

* Description: Description for criterion 1

* Title: Criterion 2. TITLE

* Description: Description for criterion 2

Peer Assignment

Peer review assignment

Task description

Remember! At least one, the text or the file, must be present in the submission.

Write your submission here:

The max length of the text is 5000 characters.

TEXT

Attach a file to your submission:

Review for Peer Assignment

Activity description

Task description

Student response

"Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum."

Your review

Remember: You must review this exercise before June 16, 2018, 6:13 p.m.. After this, date, the review can be reassigned to another student.

Criterion 1. TITLE

Description for criterion 1

Criterion 2. TITLE

Description for criterion 2

Comments

Figure 2. Implementation of a peer activity and different stages of a peer assessment task in the platform. From left to right: (A) teacher's design of the activity; (B) student's answer; (C) peer's review.

Analyzing Figure 2 you might have deducted that all criteria have the same weigh in the grade of the task: the grade of each individual rater will be the unweighted average of the scores of each criteria proposed for the peer assessment task. The final grade will be the average of all peer raters' grades. The assessment of a certain peer activity is based, or should be based, on criteria established by the EB. The summative evaluation on the platform is mandatory, in other words, no review can be submitted unless it includes the grade. However, formative feedback is optional and raters can submit the review to the system without introducing one. Additionally, the feedback box (C5) is not particular for each criterion, but it is a global feedback, yet some EBs may choose to promote it given the bidirectional benefits we have talked about in the previous section. Since the platform does not include a detailed control of this aspect of the evaluation, we do not focus on it.

3.3. *Krippendorff's alpha*

The study described in this paper has extracted the data from all the summative evaluations from UNED-COMA platform. As we analysed in Section 3.2., Krippendorff's alpha effectively works with the data we have collected, since the number of raters is independent, it works with different data types and it can handle missing values. It also takes into account the coincidences derived from randomized answers. According to Krippendorff (2001, 2004), Krippendorff's alpha is formulated as follows:

$$\alpha = 1 - \frac{D_o}{D_e} = 1 - (n-1)\frac{\sum_c \sum_{k>c} o_{ck}\,\delta_{ck}^2}{\sum_c n_c \sum_{k>c} n_k\,\delta_{ck}^2}$$

$$\delta_{ck} = \left(\frac{c-k}{c+k}\right)$$

Where:

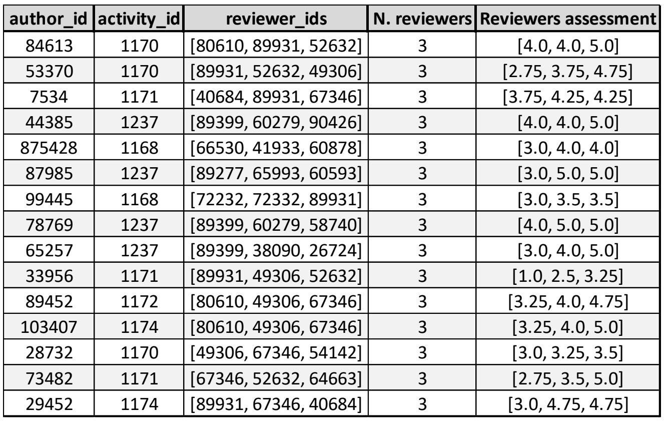
α	Krippendorff's alpha
D_o	the observed disagreement
D_e	the expected disagreement
o_{ck}, n_c, n_k and n	frequencies of values in coincidence matrix
δ_{ck}^2	difference function
c, k	elements in the difference function for the weights (row & columns)

The resulting statistical measure is a coefficient ranged from 0 to 1, where 0 is perfect disagreement and 1 is perfect agreement. The coincidence matrix is constructed from the ratings given by the reviewers. It is a square and symmetrical matrix which columns and rows are tagged with the grades assigned by raters. The coincidence matrix assigns a tabulation of the number of coincidences between values, "it visualizes the reliability of the data it tabulates" Krippendorff (2018:408). The difference function is defined according to the metric of the data in order to "weight the observed and expected coincidences of c-k pairs of values", Krippendorff (2004:232).

3.4. Data collection

Our data include a total number of 89 peer evaluation activities, of which 63 have been considered valid for this study. The main rationale behind this selection has been the validity of the activity, given that, in many cases, EBs have rejected or redesigned some activities, which have consequently become obsolete. We have determined validity based on those contents that were ratified by EBs. Another reason has been based on the size of the sample of tasks submitted; if it was too small the peer activity has not been considered. Table 1 shows one example of the, already, pre-processed raw information extracted from our DDBB, according to the methodology we have specified above, from which we have post-processed and analyzed the data.

Table 1. Extracted and post-processed information.



author_id	activity_id	reviewer_ids	N. reviewers	Reviewers assessment
84613	1170	[80610, 89931, 52632]	3	[4.0, 4.0, 5.0]
53370	1170	[89931, 52632, 49306]	3	[2.75, 3.75, 4.75]
7534	1171	[40684, 89931, 67346]	3	[3.75, 4.25, 4.25]
44385	1237	[89399, 60279, 90426]	3	[4.0, 4.0, 5.0]
875428	1168	[66530, 41933, 60878]	3	[3.0, 4.0, 4.0]
87985	1237	[89277, 65993, 60593]	3	[3.0, 5.0, 5.0]
99445	1168	[72232, 72332, 89931]	3	[3.0, 3.5, 3.5]
78769	1237	[89399, 60279, 58740]	3	[4.0, 5.0, 5.0]
65257	1237	[89399, 38090, 26724]	3	[3.0, 4.0, 5.0]
33956	1171	[89931, 49306, 52632]	3	[1.0, 2.5, 3.25]
89452	1172	[80610, 49306, 67346]	3	[3.25, 4.0, 4.75]
103407	1174	[80610, 49306, 67346]	3	[3.25, 4.0, 5.0]
28732	1170	[49306, 67346, 54142]	3	[3.0, 3.25, 3.5]
73482	1171	[67346, 52632, 64663]	3	[2.75, 3.5, 5.0]
29452	1174	[89931, 67346, 40684]	3	[3.0, 4.75, 4.75]

4. RESULTS

4.1. Distribution of the peer review assessments

We have collected globally a total number of 37,506 submitted tasks that belonged to peer evaluation activities. 9,761 tasks were discarded due to they belonged to the not validated peer activities aforementioned in section 3.4 or because they were not reviewed by at least three raters. We have thus included 27,745 valid tasks.

Regarding to the final grades, most of them span from 3.5 to 4.5 (55.81%). The most common final grade (mode) has been 4 (6.33%). 5.32% peer tasks obtained the highest grade (5); while the lowest grade (1) was only given to 43 tasks (0.155%). The average grade has been 3.859 out of 5; meanwhile, the median is 3.917. Therefore, given that the mean is lower than the median, and that they are both lower than the mode, the distribution of grades is slightly biased to the right as Figure 3 shows. Regarding to the peer reviews, we have a sample of 93,334 reviews, most of them were scored between 4 and 5 (56.74%), the mode has been 5 (24.46%), while only 2.33% of the reviews were marked with the minimum grade.

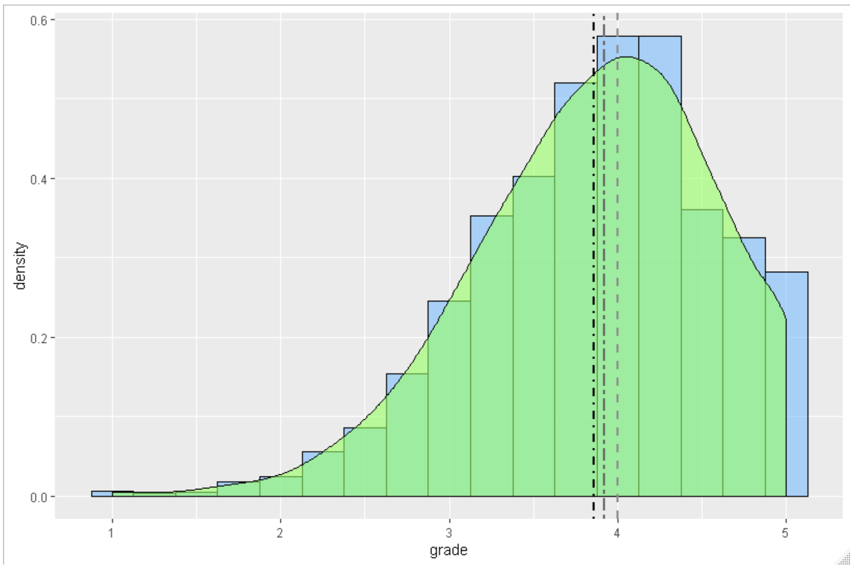


Figure 3. Distribution of the final grade over the 27,745 tasks validated; dark vertical line indicates the mean of all final grades, grey vertical line denotes the median while light-grey vertical line the mode.

Each validated task of this study involves, at least, three reviews. Taking into account that each review task has several evaluation criteria, we had to consider almost 334,000 assessed criteria to come up with the summative evaluations of each revision. All this information is contained in Figure 4 for each activity where it represents the number of submitted tasks on the *x*-axis, the average number of raters per activity on the *y*-axis, and the number of evaluation criteria for each activity.

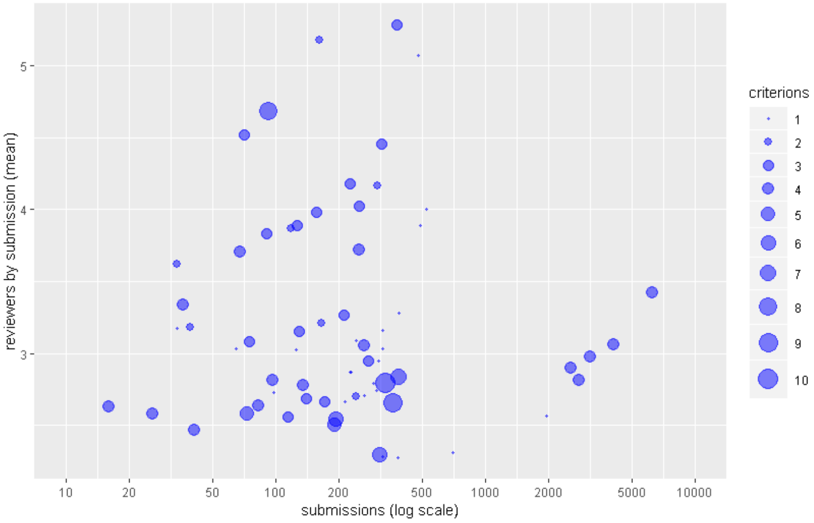


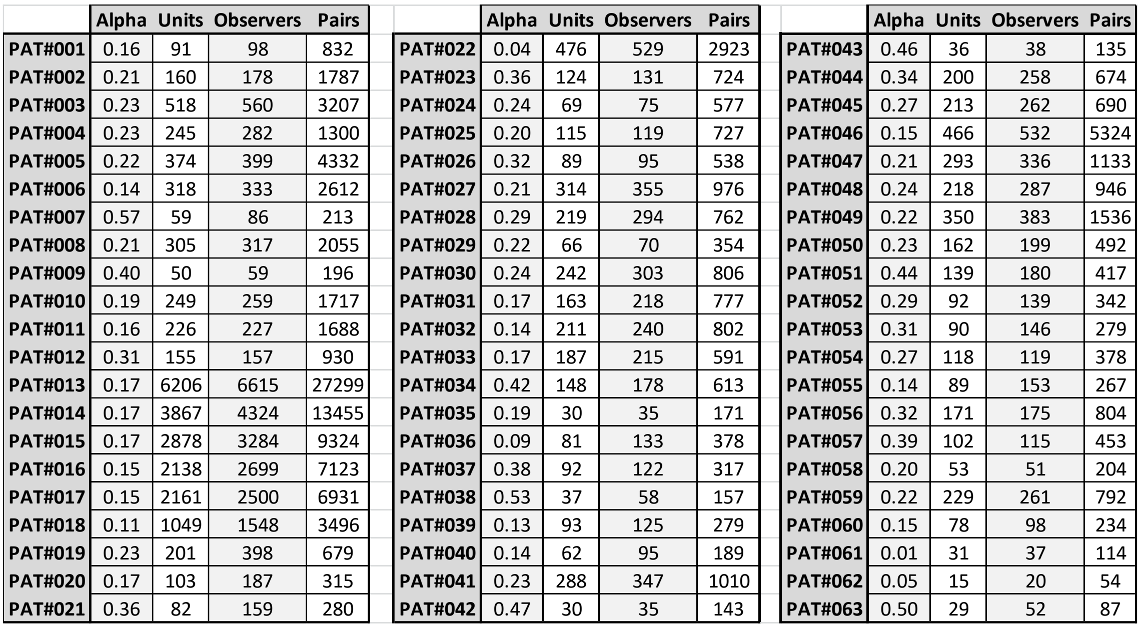
Figure 4. Scatterplot representing the available data. Each dot represents an activity with the average number of raters on the *y*-axis and the number of submitted tasks (log scale) in *x*-axis. The size of the dot codifies the number of criterions in the task.

4.2. Results of reliability based on Krippendorff's alpha

Krippendorff's alpha considers observers interchangeable with the number of pairs used. Consequently, the results are based on all the data provided by all observers, and it is not affected by their number (Hayes & Krippendorff, 2007).

The value of Krippendorff's alpha (see equation) must be found between '1', when the observed disagreement (D_o) is null, and '0' when the observed disagreement (D_o) matches the expected disagreement (D_e). According to Krippendorff (2011), as a general rule of thumb, we assume that the relevant values, or the statistically significant values for Krippendorff's alpha, should be over 0.80. However, some positive conclusions or trends can be drawn from 0.67 onwards. To this respect, Hallgren (2012) points out that these values can vary depending on research methodology and goals. Table 2 presents the Krippendorff's alpha results for the considered peer activities based on the aforementioned equation and the macro provided by Hayes & Krippendorff (2007). The box-plot representation for Krippendorff's alpha of the 63 analyzed activities in the different courses is shown in Figure 5. The mean for all 63 activities is of 0.2327; while the first and the third quartiles are on 0.1573 and 0.3092 respectively. In other words, most of the activities have a very low Krippendorff's alpha.

Table 2. Krippendorff's alpha results.



	Alpha	Units	Observers	Pairs
PAT#001	0.16	91	98	832
PAT#002	0.21	160	178	1787
PAT#003	0.23	518	560	3207
PAT#004	0.23	245	282	1300
PAT#005	0.22	374	399	4332
PAT#006	0.14	318	333	2612
PAT#007	0.57	59	86	213
PAT#008	0.21	305	317	2055
PAT#009	0.40	50	59	196
PAT#010	0.19	249	259	1717
PAT#011	0.16	226	227	1688
PAT#012	0.31	155	157	930
PAT#013	0.17	6206	6615	27299
PAT#014	0.17	3867	4324	13455
PAT#015	0.17	2878	3284	9324
PAT#016	0.15	2138	2699	7123
PAT#017	0.15	2161	2500	6931
PAT#018	0.11	1049	1548	3496
PAT#019	0.23	201	398	679
PAT#020	0.17	103	187	315
PAT#021	0.36	82	159	280
PAT#022	0.04	476	529	2923
PAT#023	0.36	124	131	724
PAT#024	0.24	69	75	577
PAT#025	0.20	115	119	727
PAT#026	0.32	89	95	538
PAT#027	0.21	314	355	976
PAT#028	0.29	219	294	762
PAT#029	0.22	66	70	354
PAT#030	0.24	242	303	806
PAT#031	0.17	163	218	777
PAT#032	0.14	211	240	802
PAT#033	0.17	187	215	591
PAT#034	0.42	148	178	613
PAT#035	0.19	30	35	171
PAT#036	0.09	81	133	378
PAT#037	0.38	92	122	317
PAT#038	0.53	37	58	157
PAT#039	0.13	93	125	279
PAT#040	0.14	62	95	189
PAT#041	0.23	288	347	1010
PAT#042	0.47	30	35	143
PAT#043	0.46	36	38	135
PAT#044	0.34	200	258	674
PAT#045	0.27	213	262	690
PAT#046	0.15	466	532	5324
PAT#047	0.21	293	336	1133
PAT#048	0.24	218	287	946
PAT#049	0.22	350	383	1536
PAT#050	0.23	162	199	492
PAT#051	0.44	139	180	417
PAT#052	0.29	92	139	342
PAT#053	0.31	90	146	279
PAT#054	0.27	118	119	378
PAT#055	0.14	89	153	267
PAT#056	0.32	171	175	804
PAT#057	0.39	102	115	453
PAT#058	0.20	53	51	204
PAT#059	0.22	229	261	792
PAT#060	0.15	78	98	234
PAT#061	0.01	31	37	114
PAT#062	0.05	15	20	54
PAT#063	0.50	29	52	87

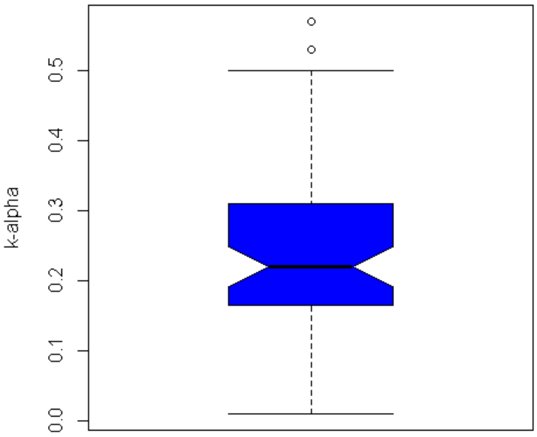


Figure 5. Boxplot of the Krippendorff's alpha values of all peer review activities in all courses.

4.3. *Factors influencing reliability*

Considering all tasks, the average standard deviation (SD) and the Pearson's Coefficient of Variation (PCV) of the Krippendorff's alpha are 0.12 and 0.5 respectively. The mean of Krippendorff's alpha for all peer review activities is 0.2327 (Figure 6). By analyzing the peer assessment tasks by course, we can draw some conclusions, e.g., in Figure 6 the reliability of the peer assessment tasks is grouped by course and arranged by its sequence order within the course. The dispersion of the reliability by course is, in general terms, much better than the global one. Considering those courses containing at least two peer assessments tasks, averaging the reliability of the tasks by course provides a better result in terms of dispersion: Only one course (C23 in Figure 6) presents worse dispersion values (SD ~ 0.16, PCV ~ 0.7), and two courses (C24 and C25) present similar dispersion values (C24:: SD ~ 0.11, PCV ~ 0.5; C25:: SD ~ 0.14, PCV ~ 0.48). Most courses (8 courses) present dispersion values for the reliability around half of the global one, both for the SD and PCV. It should be noted the case of course C20, which, with five peer assessment tasks, presents the lowest dispersion values (0.01 and 0.07 for SD and PCV respectively).

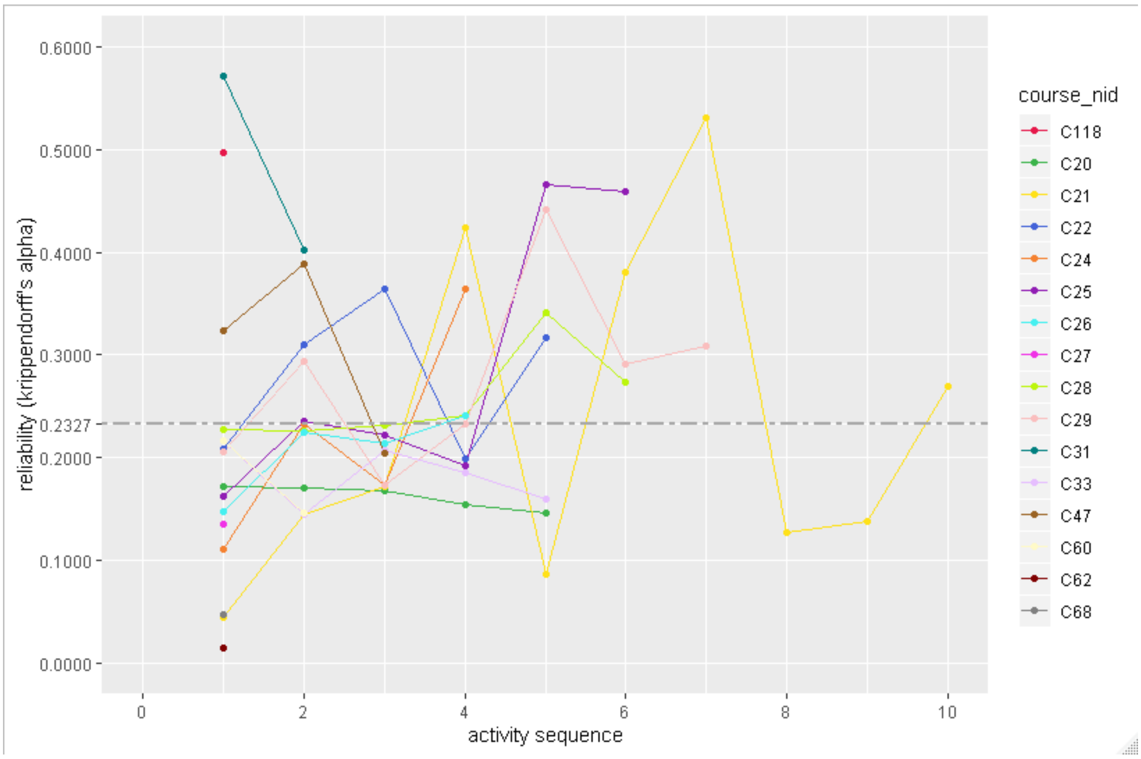


Figure 6. Evolution of the Krippendorff alpha value among the different courses and through the different tasks.

Table 3 presents the percental distribution of disagreement in the subset of raters of each task assessed. To generate this distribution, we compute the maximal distance between the grades given by each group of raters in each task and classify them in their disagreement range. Table 3 shows the dispersion between maximal and minimal grades in the subset of raters for each peer evaluation task on the platform, which is calculated without considering the number of raters in each subset of raters. Obviously the more raters, the higher the chance of disagreeing evaluations as the probability of getting larger maximal distances increases.

Table 3. Maximum distance among the subset of raters (percentages).

	n	reviews (mean)	0	(0, 0.5]	(0.5, 1]	(1, 2]	(2, 3]	(3, 4]
PAT#001	91	4,70	0,00%	4,40%	9,89%	36,26%	28,57%	20,88%
PAT#002	160	5,21	0,00%	3,25%	9,74%	37,66%	32,47%	16,88%
PAT#003	518	4,01	10,16%	0,00%	32,42%	32,42%	14,84%	10,16%
PAT#004	245	3,76	3,35%	5,86%	22,59%	36,40%	23,01%	8,79%
PAT#005	374	5,31	0,27%	0,82%	10,87%	50,27%	29,62%	8,15%
PAT#006	318	4,48	0,32%	2,88%	20,19%	51,60%	21,47%	3,53%
PAT#007	59	3,20	41,51%	0,00%	20,75%	20,75%	13,21%	3,77%
PAT#008	305	4,17	2,01%	7,36%	26,42%	46,15%	14,72%	3,34%
PAT#009	50	3,30	52,27%	0,00%	18,18%	15,91%	9,09%	4,55%
PAT#010	249	4,04	0,82%	2,47%	28,81%	48,97%	16,46%	2,47%
PAT#011	226	4,18	1,36%	3,18%	27,73%	49,09%	14,55%	4,09%
PAT#012	155	3,99	1,34%	3,36%	22,82%	46,31%	22,82%	3,36%
PAT#013	6206	3,43	0,66%	9,65%	20,85%	45,82%	19,11%	3,90%
PAT#014	3867	3,12	1,68%	13,21%	26,91%	42,19%	13,52%	2,49%
PAT#015	2878	3,06	1,78%	14,38%	26,78%	42,76%	12,36%	1,95%
PAT#016	2138	3,06	2,44%	13,13%	28,80%	42,68%	10,79%	2,16%
PAT#017	2161	3,06	2,97%	14,52%	30,58%	39,44%	10,39%	2,09%
PAT#018	1049	3,05	21,00%	0,00%	41,32%	24,07%	8,53%	5,08%
PAT#019	201	3,07	16,41%	1,54%	39,49%	29,23%	9,74%	3,59%
PAT#020	103	3,02	10,31%	0,00%	46,39%	27,84%	9,28%	6,19%
PAT#021	82	3,10	19,74%	0,00%	39,47%	17,11%	13,16%	10,53%
PAT#022	476	3,93	13,83%	0,00%	20,00%	24,47%	14,47%	27,23%
PAT#023	124	3,92	3,39%	0,85%	28,81%	39,83%	24,58%	2,54%
PAT#024	69	4,55	1,59%	0,00%	15,87%	47,62%	28,57%	6,35%
PAT#025	115	3,91	2,75%	3,67%	23,85%	40,37%	22,02%	7,34%
PAT#026	89	3,84	2,41%	3,61%	13,25%	40,96%	33,73%	6,02%
PAT#027	314	3,03	1,95%	9,42%	25,97%	44,81%	12,34%	5,52%
PAT#028	219	3,08	3,29%	15,02%	29,11%	39,44%	7,04%	6,10%
PAT#029	66	3,74	0,00%	3,33%	23,33%	38,33%	23,33%	11,67%
PAT#030	242	3,08	2,12%	15,68%	28,39%	34,32%	11,44%	8,05%
PAT#031	163	3,59	7,01%	10,19%	26,75%	33,12%	13,38%	9,55%
PAT#032	211	3,25	34,63%	0,00%	28,29%	19,02%	9,27%	8,78%

	n	reviews (mean)	0	(0, 0.5]	(0.5, 1]	(1, 2]	(2, 3]	(3, 4]
PAT#033	187	3,05	50,28%	0,00%	22,65%	15,47%	6,63%	4,97%
PAT#034	148	3,34	20,42%	9,86%	33,80%	21,13%	5,63%	9,15%
PAT#035	30	3,80	4,17%	8,33%	25,00%	33,33%	16,67%	12,50%
PAT#036	81	3,40	10,67%	2,67%	17,33%	24,00%	18,67%	26,67%
PAT#037	92	3,13	12,79%	13,95%	23,26%	31,40%	12,79%	5,81%
PAT#038	37	3,41	19,35%	6,45%	22,58%	35,48%	9,68%	6,45%
PAT#039	93	3,00	20,69%	5,75%	22,99%	32,18%	6,90%	11,49%
PAT#040	62	3,02	26,79%	0,00%	14,29%	23,21%	19,64%	16,07%
PAT#041	288	3,15	25,89%	0,00%	37,94%	22,70%	10,64%	2,84%
PAT#042	30	3,57	4,17%	12,50%	16,67%	41,67%	20,83%	4,17%
PAT#043	36	3,25	0,00%	6,67%	40,00%	40,00%	13,33%	0,00%
PAT#044	200	3,12	62,37%	0,00%	21,13%	14,95%	1,03%	0,52%
PAT#045	213	3,08	70,53%	0,00%	17,87%	7,73%	0,97%	2,90%
PAT#046	466	5,16	2,39%	0,00%	18,48%	37,17%	27,39%	14,57%
PAT#047	293	3,28	14,63%	0,00%	49,83%	27,18%	6,62%	1,74%
PAT#048	218	3,31	8,96%	0,00%	54,25%	29,25%	6,60%	0,94%
PAT#049	350	3,41	10,17%	0,00%	42,15%	36,63%	9,30%	1,74%
PAT#050	162	3,01	54,49%	12,18%	8,33%	12,82%	5,13%	7,05%
PAT#051	139	3,00	69,17%	0,00%	16,54%	12,03%	1,50%	0,75%
PAT#052	92	3,11	12,79%	16,28%	17,44%	36,05%	15,12%	2,33%
PAT#053	90	3,03	28,57%	11,90%	20,24%	26,19%	7,14%	5,95%
PAT#054	118	3,07	29,46%	0,00%	29,46%	21,43%	4,46%	15,18%
PAT#055	89	3,00	0,00%	10,84%	30,12%	48,19%	10,84%	0,00%
PAT#056	171	3,56	1,21%	7,27%	35,15%	39,39%	16,97%	0,00%
PAT#057	102	3,46	6,25%	4,17%	34,38%	30,21%	21,88%	3,13%
PAT#058	53	3,28	0,00%	4,26%	27,66%	44,68%	14,89%	8,51%
PAT#059	229	3,15	1,35%	11,21%	27,35%	40,36%	16,59%	3,14%
PAT#060	78	3,00	1,39%	8,33%	25,00%	52,78%	12,50%	0,00%
PAT#061	31	3,23	7,69%	0,00%	0,00%	38,46%	23,08%	30,77%
PAT#062	15	3,20	38,46%	0,00%	30,77%	7,69%	23,08%	0,00%
PAT#063	29	3,00	4,35%	13,04%	34,78%	43,48%	4,35%	0,00%

We believe that additional explanation regarding the Krippendorff's alpha reliability peer assessment will be helpful to avoid misinterpreting some data points. For the results in Table 3, grades vary from 1 (lowest) to 5 (highest) in a 1 by 1 scale of whole numbers. For those tests that contain only one evaluation criterion, which is the case in over 20 activities, the lowest level of disagreement would be a distance of 1. Therefore, this is the reason why we consider the maximal distance of 1 as acceptable for an agreement percentage. Figure 7a and Figure 7b scatterplots show the relationship between Krippendorff's alpha and the percentage of tests in which the evaluation of the raters has shown a strong agreement (distance between grades below or equal to 1) and the percentage of tests in which the evaluation provided by the subset of raters has shown a strong disagreement (distance between grades bigger or equal to 3). The PCC coefficient for the Krippendorff's alpha and the percentage of peer assessment tasks with a strong agreement between the raters of each subset is low, 0.311 (p-value = 0.013). In the case of the correlation between the disagreement and the reliability, the correlation is stronger, -0.395 (p-value = 0.001).

Figure 4c and Figure 4d show the relationship between the number of criteria of the activity and the average number of raters, respectively, with the Krippendorff's alpha. In both cases, the PCC coefficient is not significant (p-values = 0.7901 and 0.2845 respectively), thus we accept the hypothesis that true correlation is equal to 0. Furthermore, the correlation is low in both cases (0.034 and -0.137 respectively).

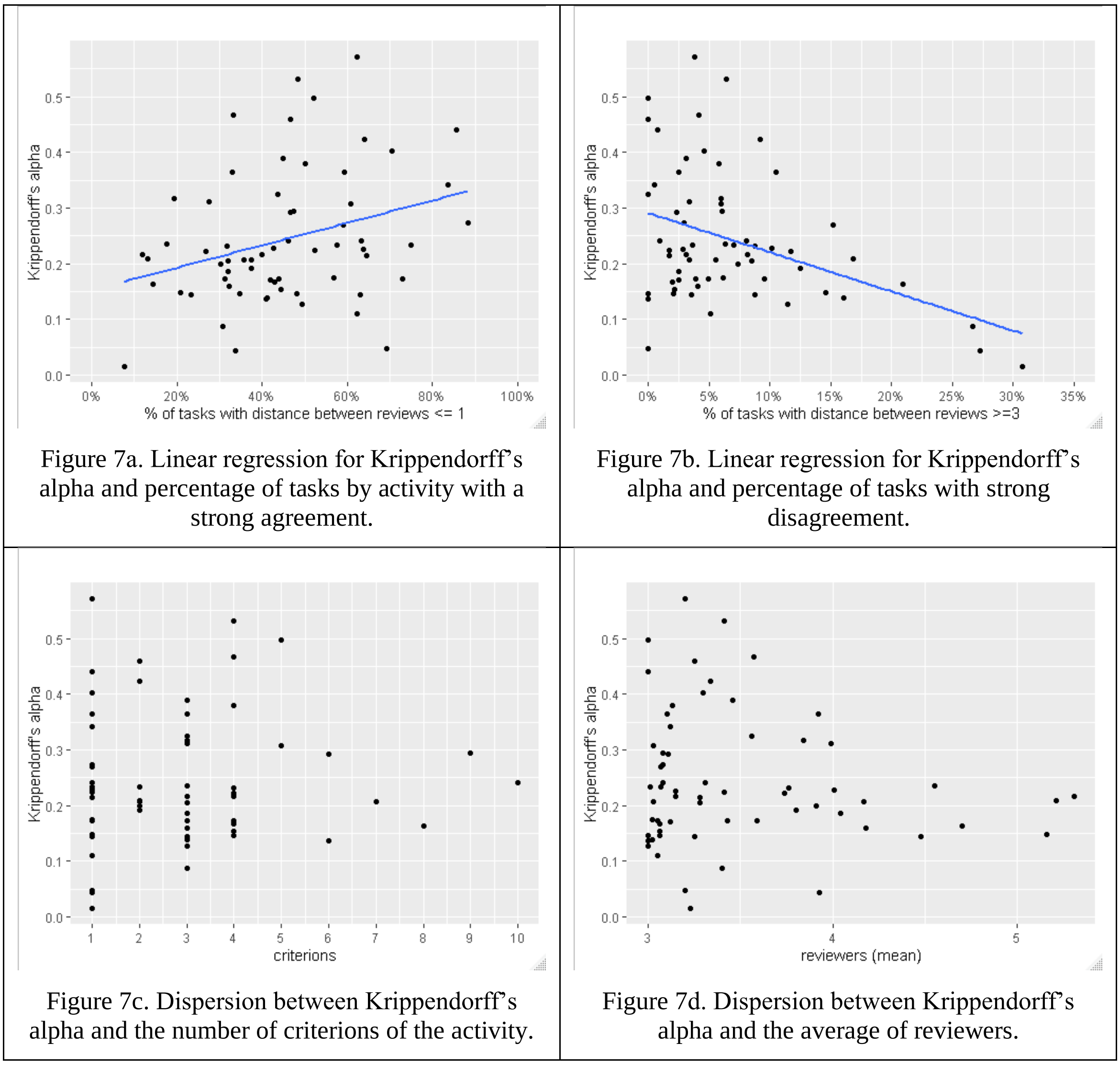


Figure 7a. Linear regression for Krippendorff's alpha and percentage of tasks by activity with a strong agreement.

Figure 7b. Linear regression for Krippendorff's alpha and percentage of tasks with strong disagreement.

Figure 7c. Dispersion between Krippendorff's alpha and the number of criterions of the activity.

Figure 7d. Dispersion between Krippendorff's alpha and the average of reviewers.

Figure 7. Scatterplots showing the reliability dispersion based on different factors.

5. *DISCUSSION*

Attending to the values obtained for Krippendorff's alpha statistic in the 63 assessed activities, and considering the recommendations offered in Krippendorff & Bock (2009: 354) and Krippendorff (2004: 241) to "rely on Krippendorff's alpha above 0.80", we find that in our peer review activity dataset there are no significant values in terms of agreement between reviewers. Therefore, none of the peer evaluation activities carried out in the different courses on the platform can be considered reliable when talking about the evaluations performed by the students.

The maximum value of Krippendorff's alpha was obtained in activity PAT#007 (0.5718). However, not even this value is enough to be used for "drawing tentative conclusions", because the value remains under the threshold value (0.667) (Krippenddorff & Bock, 2009:354; Krippenddorff, 2004:241).

Under the assumption that reliability is, although not sufficient, a necessary condition to guarantee the validity of the established evaluation methodology, with the obtained results in hand we can conclude that grades obtained by means of peer assessment in this study are not trustworthy. Jonsson & Svingby (2007) highlight that reliability is not always required for validity, because there are certain scenarios where “the basis of the assessment can be easily changed” (for example, in-classroom assessments). These scenarios are nowhere close to our case study.

Despite we cannot perform a direct comparison between our results and the ones reported in other studies due to the use of different statistics, the differences and conclusions from each separate study suggest that our study presents much lower reliability than the rest of studies that performed similar analysis in other contexts and using different metrics. The results obtained in classical learning scenarios tend to provide a solid reliability. For example the ones provided by Yoon *et al*., (2018) —with ICCs values obtained from 0.390 to 0.863; being the overall average 0.659, from 141 students, who were divided into 18 groups in 11 team-based learning classes— or the ones obtained by Salehi & Masoule (2017) —Cronbach’s alpha values from 0.709 to 0.900 for peer assessing oral production in three groups. Moreover, in other studies using MOOCs as learning scenario; for example the ICC averages measures obtained by Formanek *et al*., (2017) and Luo, Robinson & Park (2014) —0.591 for the ICC and 0.579 respectively.

Anyway, and according to our results, the fact that we did not find any peer assessment activities with Krippendorff’s alpha values even close to the recommended threshold values, drives us to think that the reason might be a systematic problem and not particularly associated with specific peer assessment activities in our case study. However, analyzing Figure 6 we can see how the mean of the Krippendorff’s alpha between courses is quite different. We do not find substantial differences after grouping courses by topic and, according to the data obtained, it does not look as if there is a significant relationship between the topic of the course and the reliability achieved. Conversely, even if they are focused on similar topics, such as C21, C31 and C32, all of them focused on TICs and its applications, which have completely different results (C21:: mean ~ 0.23, SD ~ 0.16, PCV ~ 0.7; C31:: mean ~ 0.049, SD ~ 0.12, PCV ~ 0.24; C32:: mean ~ 0.027, SD ~ 0.07, PCV ~ 0.26). Another example is C20 and C24, both dedicated to the study of foreign languages, which have relatively similar Krippendorff’s alpha value (C24 ~ 0.22; C20 ~ 0.16), but with dispersion rates quite far from each other (C24:: SD ~ 0.11, PCV ~ 0.5; C20:: SD ~ 0.01, PCV ~ 0.07). Therefore, in our case study we do not find the topic of the course as a relevant factor affecting reliability, in accordance with the conclusions obtained by Falchikov, & Goldfinch (2000).

It is noteworthy the high grades obtained in the peer assessment activities within the platform. One potential explanation regarding this aspect may be related to the involved social factors. While in MOOCs certain social aspects, described in section 2.2, are avoided due to the physical distance and anonymity, some others might still be playing a role, such as the “perception of criticism as socially uncomfortable” (Topping 2009). Students may be more generous when grading a fellow peer, if we compare grades with instructors’ ones (Marks & Jackson, 2013). Hanrahan & Isaacs (2001) pinpoint that students experience empathy with lecturers/tutors because of the large numbers of assignments, however they do not feel the same way towards their peers. In this direction, the results obtained by Formanek *et al*., (2017) do not show a global trend: “Peer graders tend to underestimate the top-scoring submissions while overestimating the lowest scoring ones”. In the meta-analysis conducted by Falchikov & Goldfinch (2000), from 22 studies (not considering atypical ones), 11 studies resulted in over-grading while 7 in under-grading, turning out a weighted mean very slightly under-grading (effect size -0.02).

Training and practicing peer assessment tasks are highlighted as requirements for students before an actual implementation in a real educational scenario (Topping, 2009). However, this training is sometimes focused on how to conduct the grading side following the recommendations of the EB, instead of on the educational component, reliability and/or validity (Kulkarni *et al*., 2013). In any case, Sluijsmans, Brand-Gruwel & Merriënboer (2002a) indicate that training promotes a more critical attitude when assessing, but that long training periods are required in order to provide tangible improvements (Sluijsmans *et al*., 2002b). Formanek *et al*., (2017) found that the performing a previous training stage in how to assess, helped to improve reliability: an average ICC of 0.591 for graders without previous training against an average ICC of 0.682 for those trained graders. If we look at the reliability of our students as they progress in each course, we hypothesize that it should improve as they are getting more experienced in conducting peer assessment. When comparing in each course the average reliability of the first half of peer assessment tasks with the average reliability of the final half of peer assessment tasks (e.g., course 25 comprises six tasks: we have compared the average reliability of tasks 1, 2 and 3 with the average reliability of tasks 4, 5 and 6; while course 29, which comprises seven tasks: we have compared the average reliability of tasks 1, 2 and 3 with the average reliability of tasks 5, 6 and 7), the next conclusions, which are in concordance with the aforementioned studies, are envisaged: Courses with more than six tasks present an improvement in the reliability. An average improvement of 54.63% when comparing the reliability of initial tasks with final tasks.

For those courses with four or five peer assessment tasks, the results present a clear difference between the reliability of the first and final halves. Perhaps new research in this direction can experiment on the impact of having an initial peer-review training as a MOOC activity in the reliability of the rest of peer-review assignments. If we recall the technical implementation of the evaluation model based on Krippendorff's alpha values, one of the underlying assumptions was the idea of equity among peer raters. As aforementioned, whereas traditional learning contexts can assume a high similarity degree in the background of their learners, the 'Open' nature of MOOCs highly increments the diversity in learners' profiles, hence potentially breaking the equity among learners' condition. In MOOCs we find that learners have multiple backgrounds in content knowledge (especially those regarding STEM), diverse sets of skills related to writing, text comprehension, synthesis and very different intentions when enrolling in a MOOC (Alario-Hoyos, Pérez-Sanagustín, Delgado-Kloos, Parada, & Muñoz-Organero, 2014; Watson, Watson, Yu, Alamri, & Mueller, 2017).

Two factors traditionally analyzed in the reliability have been the number of criterions and the number of reviewers. Figure 4c and Figure 4d show the null relationship between the number of criteria of the activity and the average number of raters, respectively, with the Krippendorff's alpha. In both cases, the Pearson's product-moment correlation was not significant, thus in our case study we do not find a relationship between these factors and reliability.

Regarding to the number of criteria or categories to be assessed by peer raters, and in contrast to what Sadler & Good (2006) and Meletiadou & Tsagari (2014) found, or the conclusions obtained by Falchikov & Goldfinch (2000), we do not find any trend in this sense. In our scenario, we found an absence of a significant correlation between the number of criteria and the reliability obtained (Figure 7c). The number of criteria for each task does not imply any correlation with the Krippendorff's alpha coefficient. In our analysis, the value of Krippendorff's alpha ranges from 0.225 to 0.275 (Figure 4c). The highest average, 0.275 is obtained with tasks requiring two criterions to be assessed, followed by 0.267 for five or more criterions. On the opposite side, the lowest average value is obtained with three criterions, 0.225.

In the case of the effect of the number of peer raters in the reliability of the assessment process, we do not find any correlation neither (Figure 7d). In our case study, we have not found any trend as the ones described in the literature review.

In Figure 7a and Figure 7b scatterplots with the relationship between Krippendorff's alpha and the percentage of tests in which the evaluation of the raters has shown a strong agreement —distance between grades below or equal to 1— and the percentage of tests in which the evaluation provided by the subset of raters has shown a strong disagreement—distance between grades bigger or equal to 3. Both scatterplots show a correlation between the percentage of agreement and reliability. However, we can see how a strong agreement or the absence of disagreement does not necessarily imply high reliability. The observable dispersion confirms that agreement among raters is mainly irrelevant from the reliability as Krippendorff (2011) predicts.

6. CONCLUSIONS AND FUTURE LINES OF RESEARCH

In the particular scenario of UNED-COMA that we have analyzed, we find that the reliability of peer evaluation activities in MOOCs is untrustworthy. Therefore, under the assumption that reliability is a necessary condition to guarantee the validity of the evaluation, peer rating might not be a very trustworthy assessment method in MOOCs, especially if implemented as a summative assessment that counts towards the certification grade. However, our analyses do not take into account the learning benefits of these kind of activities, which have been presented in our introduction. Peer-assessments have been extensively analyzed in the educational literature, finding that students engage more easily in the learning process, they develop critical thinking, etc. Therefore, beyond their reliability and validity as an evaluation method, peer assessments can still provide multiple benefits for students such as a more complex cognitive learning process or personalized feedback; for example, strategies as the one described in (Staubitz, Petrick, Bauer, Renz & Meinel, 2016) can be applied in order to motivate reviewers to enhance their feedbacks. However, for students to rigorously and fully engage in a learning activity, they often need an incentive towards the final grade. Under this case scenario, one potential pedagogical approach is to mitigate this effect by assigning a relatively low weight to these evaluations in final grades, while maintaining the rest of side transversal advantages. Based on the results obtained, we perceive the need to adapt peer assessment activities, which are traditionally carried out in (relatively) homogeneous and "quasi-controlled" environments, to massive and highly heterogeneous environments.

Future work might lead us to explore if the results of this case study replicate in the peer-assessment systems of other MOOC environments, a comparison of the Krippendorff's alpha statistic with others inter-reliability statistics, experimentation around the effect on reliability of conducting peer-review training before the actual peer-review activities, to analyze the existence and significance of any correlation between the weighting of peer assessments and the reliabilities, or a more in-depth analysis of which qualitative factors moderate the disagreement between raters, such as type of course, background of raters or if it might be more specific to the implementation of the peer evaluation activity.

Acknowledgements

This work has been co-funded by the Madrid Regional Government, through the project e-Madrid-CM (P2018/TCS-4307). The e-Madrid-CM project is also co-financed by the Structural Funds (FSE and FEDER). Authors also acknowledge the support of the e-LIVES. e-Learning InnoVative Engineering Solutions- Erasmus+ Capacity Building in Higher Education 2017 - 585938- EPP-12017-1-FR-EPPKA2-CBHE-J, IoE-EQ. Internet of Energy - Education and Qualification, Erasmus+ - Cooperation for Innovation and the Exchange of Good Practices nº 2017-1-IT01-KA202-006251 and I4EU - Key Competences for an European Model of Industry 4.0, Erasmus+ Strategic Partnership nº 2019-1-FR01-KA202-06296. As well as to the projects 2020-IEQ15, 2020-IEQ14 and 2020-IEQ13 from the Escuela Superior de Ingenieros Industriales of UNED. Ruipérez-Valiente acknowledges support from the Spanish Ministry of Economy and Competitiveness through the Juan de la Cierva Formación program (FJCI-2017-34926).

REFERENCES

Alcarria, R., Bordel, B., Andres, D.M.d., & Robles, T. (2018). Enhanced Peer Assessment in MOOC Evaluation Through Assignment and Review Analysis. *International Journal of Emerging Technologies in Learning (iJET),* 13(1), 206-219.

Alario-Hoyos, C., Pérez-Sanagustín, M., Delgado-Kloos, C., Parada, H.A., & Muñoz-Organero, M. (2014). Delving into Participants' Profiles and Use of Social Tools in MOOCs. *IEEE Transactions on Learning Technologies*, 7(3), 260-266.

Aleven, V., Sewall, J., Popescu, O., Xhakaj, F., Chand, D., Baker, R., Wang, Y., Siemens, G., Rosé, C., & Gasevic, D. (2015). The Beginning of a Beautiful Friendship? *Intelligent Tutoring Systems and MOOCs.* , 525-528.

AlFallay, I. (2004). The role of some selected psychological and personality traits of the rater in the accuracy of self- and peer-assessment. *System*, 32(3), 407-425.

Antoine, J., Villaneau, J., & Lefeuvre, A. (2014). Weighted Krippendorff's alpha is a more reliable metrics for multi- coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. *Proceedings of EACL 2014*.

Ambekar, D., & Phatak, D. B. (2014). *Evaluation of essays using incremental training for Maximizing Human-Machine agreement* (Doctoral dissertation, Indian Institute of Technology, Bombay).

Beldarrain, Y. (2006). Distance education trends: Integrating new technologies to foster student interaction and collaboration. *Distance Education, 27*(2), 139-153.

Bell, T., Urhahne, D., Schanze, S., & Ploetzner, R. (2010). Collaborative inquiry learning: Models, tools, and challenges. *International Journal of Science Education, 32*(3), 349-377.

Benlloch-Dualde, J.V., & Blanc-Clavero, S. (2007). Adapting teaching and assessment strategies to enhance competence-based learning in the framework of the European convergence process. *Proceedings - Frontiers in Education Conference, FIE*, S3B6.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability(formerly: Journal of Personnel Evaluation in Education), 21*(1), 5.

Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology*, *54*(2), 199-231.

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

Brown, G.A., Bull, J., Pendlebury, M., Bull, J., & Pendlebury, M. (1997). *Assessing Student Learning in Higher Education*. London: Routledge.

Capdevila, R., & Aranzadi, P. (2014). Los cursos online masivos y abiertos: ¿Oportunidad o amenaza para las universidades iberoamericanas?= Massive open online courses: Opportunity or threat for Iberoamerican universities? RIED: revista iberoamericana de educación a distancia, (17, n.1), 2014, 69-82.

Cartney, P. (2010). Exploring the use of peer assessment as a vehicle for closing the gap between feedback given and feedback used. *Assessment & Evaluation in Higher Education, 35*(5), 551-564.

Chang, C., Liang, C., & Chen, Y. (2013). Is learner self-assessment reliable and valid in a Web-based portfolio environment for high school students? *Computers & Education, 60*(1), 325-334.

Chang, C., Tseng, K., & Lou, S. (2012). A comparative analysis of the consistency and difference among teacher-assessment, student self-assessment and peer-assessment in a Web-based portfolio assessment environment for high school students. *Computers & Education, 58*(1), 303-320.

Chen, C.-. (2008). Intelligent web-based learning system with personalized learning path guidance. *Computers and Education, 51*(2), 787-814.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Cornford, I.R. (2002). Learning-to-learn strategies as a basis for effective lifelong learning. *International Journal of Lifelong Education, 21*(4), 357-368.

De Grez, L., Valcke, M., & Roozen, I. (2012a). How effective are self- and peer assessment of oral presentation skills compared with teachers' assessments? *Active Learning in Higher Education, 13*(2), 129-142.

De Grez, L., Valcke, M., & Roozen, I. (2012b). How effective are self-and peer assessment of oral presentation skills compared with teachers' assessments? *Active Learning in Higher Education, 13*(2), 129-142.

De Miguel Díaz, M., Alfaro Rocher, I.J., Apodaca Urquijo, P., Arias Blanco, J.M., García Jiménez, E., & Lobato Fraile, C. (2005). *Modalidades de enseñanza centradas en el desarrollo de competencias: orientaciones para promover el cambio metodológico en el espacio europeo de educación superior*: Servicio de Publicaciones. Universidad de Oviedo.

Delgado García, A.M., Borge Bravo, R., García Albero, J., Oliver Cuello, R., & Salomón Sancho, L. (2005). Competencias y diseño de la evaluación continua y final en el Espacio Europeo de Educación Superior. *Retirado em Março, 8,* 2012.

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

Dochy, Filip J. R. C., & McDowell, L. (1997). Assessment as a tool for learning. *Studies in Educational Evaluation, 23*(4), 279-298.

Duran, D. (2017). Learning-by-teaching. Evidence and implications as a pedagogical mechanism. *Innovations in Education and Teaching International, 54*(5), 476-484.

Earle, S. (2014). Formative and summative assessment of science in English primary schools: evidence from the Primary Science Quality Mark. *Research in Science & Technological Education*, *32*(2), 216-228.

Epstein, M.L., Lazarus, A.D., Calvano, T.B., Matthews, K.A., Hendel, R.A., Epstein, B.B., & Brosvic, G.M. (2002). Immediate feedback assessment technique promotes learning and corrects inaccurate first responses. *The Psychological Record, 52*(2), 187-201.

European Higher Education Area (2009). Communiqué of the Conference of European Ministers Responsible for Higher Education, Leuven and Louvain-la-Neuve, 28-29 April 2009. *European Higher Education Area (EHEA)*.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of educational research*, *70*(3), 287-322.

Nancy Falchikov (2013). *Improving assessment through student involvement: Practical solutions for aiding learning in higher and further education*. London: Taylor and Francis.

Formanek, M., Wenger, M.C., Buxner, S.R., Impey, C.D., & Sonam, T. (2017). Insights about large-scale online peer assessment from an analysis of an astronomy MOOC. *Computers & Education, 113*, 243-262.

Friedman, B.A., Cox, P.L., & Maher, L.E. (2008). An Expectancy Theory Motivation Approach to Peer Assessment. *Journal of Management Education, 32*(5), 580-612.

Garcia-Loro, F., Diaz, G., Tawfik, M., Martin, S. Sancristobal, E. & Castro, M. (2014). A practice-based MOOC for learning electronics. 2014 IEEE Global Engineering Education Conference (EDUCON), 969-974.

Garcia-Loro, F., Sancristobal, E., Gil, R., Diaz, G., Castro, M., Albert-Gómez, M., & Ribeiro-Alves, G. (2016). Electronics remote lab integration into a MOOC-Achieving practical competences into MOOCs. *EADTU 2016, The Online, Open and Flexible Higher Education Conference*, 367-379.

Garcia-Loro, F., San Cristobal, E., Diaz, G., Macho, A., Baizan, P., Blazquez, M., Castro, M., Plaza, P., Orduña, P., Auer, M., Kulesza, W., Gustavsson, I., Nilsson, K., Fidalgo, A., Alves, G., Marques, A., Hernandez-Jayo, U., Garcia-Zubia, J., Kreiter, C., Pester, A., Garcia-Hernandez, C., Tavio, R., Valtonen, K., & Lehtikangas, E. (2018). PILAR: a Federation of VISIR Remote Laboratory Systems for Educational Open Activities. 2018 *IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 134-141.

Gauci, S.A., Dantas, A.M., Williams, D.A., & Kemm, R.E. (2009). Promoting student-centered active learning in lectures with a personal response system. *American Journal of Physiology - Advances in Physiology Education, 33*(1), 60-71.

Gipps, C.V. (2005). What is the role for ICT-based assessment in universities? *Studies in Higher Education*, 30(2), 171-180.

Guan-Yu Lin (2018). Anonymous versus identified peer assessment via a Facebook-based learning application: Effects on quality of peer feedback, perceived learning, perceived fairness, and attitude toward the system. *Computers & Education; Anonymous versus identified peer assessment via a Facebook-based learning application: Effects on quality of peer feedback, perceived learning, perceived fairness, and attitude toward the system, 116,* 81-92.

Hallgren, K.A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in quantitative methods for psychology, 8*(1), 23-34.

Havnes, A., Smith, K., Dysthe, O., & Ludvigsen, K. (2012). Formative assessment and feedback: Making learning visible. *Studies in Educational Evaluation, 38*(1), 21-27.

Hayes, A.F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures, 1*(1), 77-89.

Hood, N., Littlejohn, A., & Milligan, C. (2015). Context counts: How learners' contexts influence learning in a MOOC. *Computers & Education, 91*, 83-91.

Hsia, L.-H., Huang, I., & Hwang, G.-J. (2016). Effects of different online peer-feedback approaches on students’ performance skills, motivation and self-efficacy in a dance course. *Computers & Education, 96,* 55–71.

Jackson, L. (2014). Validity and rater reliability of peer and self assessments for urban middle school students.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*(10), 1774-1787.

Jones, I., & Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation, 47*, 93-101.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review, 2*(2), 130-144.

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

Hovardas, T., Tsivitanidou, O.E., & Zacharia, Z.C. (2014). Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students. *Computers & Education*, 71, 133-152.

Kehrer, P., Kelly, K., & Heffernan, N. (2013). Does immediate feedback while doing homework improve learning? Paper presented at the *FLAIRS 2013 - Proceedings of the 26th International Florida Artificial Intelligence Research Society Conference,* pp. 542-545.

Kilic, D. (2016). An Examination of Using Self-, Peer-, and Teacher-Assessment in Higher Education: A Case Study in Teacher Education. *Higher Education Studies, 6*, 136.

Kilic, G.B., & Cakan, M. (2007). Peer Assessment of Elementary Science Teaching Skills. *Journal of Science Teacher Education, 18*(1), 91-107.

Kizilcec, R.F., Davis, G.M., & Cohen, G.L. (2017). Towards equal opportunities in MOOCs: affirmation reduces gender & social-class achievement gaps in China. *ACM conference on learning@ scale*, 121-130.

Kizilcec, R.F., Saltarelli, A.J., Reich, J., & Cohen, G.L. (2017). Closing global achievement gaps in MOOCs. *Science*, 355(6322), 251-252.

Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology*. (4th ed.). Thousand Oaks, CA: Sage Publications, Inc.

Krippendorff, K. (2011). Agreement and information in the reliability of coding. *Communication Methods and Measures, 5*(2), 93-112.

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology*. (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement, 30*(1), 61-70.

Krippendorff, K., & Bock, M.A. (2009). *The content analysis reader*: Sage.

Kulik, J.A., & Kulik, C.C. (1988). Timing of feedback and verbal learning. *Review of educational research*, *58*(1), 79-97.

Kulkarni, C., Wei, K.P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., & Klemmer, S.R. (2013). Peer and Self Assessment in Massive Online Classes. *ACM Trans.Comput.-Hum.Interact.*, 20(6), 33:31.

Lange, R.T. (2011). Inter-rater Reliability. In: J.S. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of Clinical Neuropsychology* (pp. 1348). New York, NY: Springer New York.

Lavrakas, P. (2008). Encyclopedia of Survey Research Methods.

Lawrence-Brown, D. (2004). Differentiated Instruction: Inclusive Strategies For Standards-Based Learning That Benefit The Whole Class. *American Secondary Education, 32*(3), 34-62.

Lehmann, M., Christensen, P., Du, X., & Thrane, M. (2008). Problem-oriented and project-based learning (POPBL) as an innovative learning strategy for sustainable development in engineering education. *European Journal of Engineering Education, 33*(3), 283-295.

Leenknecht, M. & Prins, F. (2018). Formative peer assessment in primary school: the effects of involving pupils in setting assessment criteria on their appraisal and feedback style. *European Journal of Psychology of Education, 33*(1), 101-116.

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

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

Lombard, M. (2005). Practical resources for assessing and reporting intercorder reliability in content analysis research projects. *http://www.temple.edu/sct/mmc/reliability/*.

Lüftenegger, M., Schober, B., van de Schoot, R., Wagner, P., Finsterwald, M., & Spiel, C. (2012). Lifelong learning as a goal – Do autonomy and self-regulation in school result in well prepared pupils? *Learning and Instruction*, *22*(1), 27-36.

Lukassen, N.B., Pedersen, A., Nielsen, A., Wahl, C., & Sorensen, E.K. (2014). Digital education with IT: How to create motivational and inclusive education in blended learning environments using flipped learning - a study in nurse education. *Proceedings of the European Conference on e-Learning, ECEL, 2014-January*, 305-312.

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

Marks, L., & Jackson, M. (2013). Student Experience of Peer Assessment on an MSc Programme. *Bioscience Education, 21*(1), 20-28.

Marton, F., & Bowden, J. (1999). The University of Learning:: Beyond Quality and Competence. *Education + Training, 41*(5), ii.

McKeachie, W.J. (1987). Teaching and Learning in the College Classroom. A Review of the Research Literature (1986) and November 1987 Supplement.

McKeachie, W.J., Pintrich, P.R., Lin, Y., & Smith, D. (1986). Teaching and learning in the college classroom. *Ann Arbor, MI: University of Michigan*.

Meletiadou, E., & Tsagari, D. (2014). An Exploration of the Reliability and Validity of Peer Assessment of Writing in Secondary Education. *Major Trends in Theoretical and Applied Linguistics 3* (pp. 235-250): Sciendo Migration.

Miller, T. (2009). Formative computer-based assessment in higher education: The effectiveness of feedback in supporting student learning. *Assessment & Evaluation in Higher Education, 34*(2), 181-192.

Moerkerke, G. (1996). Assessment for flexible learning. Utrecht: Lemma.

Mullen, J., Byun, C., Gadepally, V., Samsi, S., Reuther, A., & Kepner, J. (2017). Learning by doing, High Performance Computing education in the MOOC era. *Journal of Parallel and Distributed Computing*, 105, 105-115.

Ng, E. (2014). Using a mixed research method to evaluate the effectiveness of formative assessment in supporting student teachers' wiki authoring. *Computers & Education, 73*, 141-148.

Ng, E. (2016). Fostering pre-service teachers’ self-regulated learning through self- and peer assessment of wiki projects. *Computers & Education, 98*, 180–191.

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

Orsmond, P., Merry, S., & Reiling, K. (1996). The Importance of Marking Criteria in the Use of Peer Assessment. *Assessment & Evaluation in Higher Education, 21*(3), 239-250.

Orsmond, P., Merry, S., & Reiling, K. (2000). The Use of Student Derived Marking Criteria in Peer and Self-assessment. *Assessment & Evaluation in Higher Education, 25*(1), 23-38.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (2013). Tuned models of peer assessment in MOOCs. *arXiv preprint arXiv:1307.2579*.

Quintana, C., Reiser, B.J., Davis, E.A., Krajcik, J., Fretz, E., Duncan, R.G., Kyza, E., Edelson, D., & Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. *Journal of the Learning Sciences, 13*(3), 337-386.

Raman, K., & Joachims, T. (2014). Methods for Ordinal Peer Grading. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1037–1046.

Reinholz, D. (2016). The assessment cycle: a model for learning through peer assessment. *Assessment & Evaluation in Higher Education, 41*(2), 301-315.

Richmond, S., Shaw, W.C., O'brien, K.D., Buchanan, I.B., Jones, R., Stephens, C.D., Roberts, C.T., & Andrews, M. (1992). The development of the PAR Index (Peer Assessment Rating): reliability and validity. *The European Journal of Orthodontics, 14*(2), 125-139.

Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment,* 11(1), 1-31. doi:10.1207/s15326977ea1101_1

Salehi, M. & Masoule, Z. (2017). An investigation of the reliability and validity of peer, self-, and teacher assessment. *Southern African Linguistics and Applied Language Studies, 35*(1), 1–15.

Schaer, P. (2012). Better than Their Reputation? On the Reliability of Relevance Assessments with Students. , 124-135.

Scott, W.A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public opinion quarterly, 19*, 321-325.

Shieh, G. (2016). Exact power and sample size calculations for the two one-sided tests of equivalence. *PLoS ONE, 11*(9).

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological bulletin, 86*(2), 420-428.

Sluijsmans, D., Brand-Gruwel, S., & van Merriënboer, J. (2002a). Peer Assessment Training in Teacher Education: Effects on performance and perceptions. *Assessment & Evaluation in Higher Education, 27*(5), 443-454.

Sluijsmans, D., Brand-Gruwel, S., van Merriënboer, J. & Bastiaens, T. (2002b). The training of peer assessment skills to promote the development of reflection skills in teacher education. *Studies in Educational Evaluation, 29*(1), 23-42.

Staubitz, T., Petrick, D., Bauer, M., Renz, J., & Meinel, C. (2016). Improving the Peer Assessment Experience on MOOC Platforms. *Proceedings of the Third (2016) ACM Conference on learning @ scale,* 389-398.

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

Suen, H.K. (2014). Peer assessment for massive open online courses (MOOCs). *The International Review of Research in Open and Distributed Learning, 15*(3).

Sung, Y., Chang, K., Chang, T., & Yu, W. (2010). How many heads are better than one? The reliability and validity of teenagers' self-and peer assessments. *Journal of adolescence, 33*(1), 135-145.

Susskind, R.E., & Susskind, D. (2015). *The future of the professions: How technology will transform the work of human experts*. Oxford University Press, USA.

Thelwall, M. (2000). Computer-based assessment: a versatile educational tool. *Computers & Education, 34(1),* 37-49.Thurstone, L.L. (1927). A law of comparative judgment. *Psychological review, 34*(4), 273-286.

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

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

Topping, K.J. (2017). Peer Assessment: Learning by Judging and Discussing the Work of Other Learners. *Interdisciplinary Education and Psychology*, *1*(1), 1-17. [7]. https://doi.org/10.31532/InterdiscipEducPsychol.1.1.00

Tsai, C., Lin, S.S.J., & Yuan, S. (2002). Developing science activities through a networked peer assessment system. *Computers & Education, 38*(1), 241-252.

Van der Pol, J., Van den Berg, B., Admiraal, W.F., & Simons, P.R. (2008). The nature, reception, and use of online peer feedback in higher education. *Computers & Education, 51*(4), 1804-1817.

Van Den Bossche, P., Gijselaers, W.H., Segers, M., & Kirschner, P.A. (2006). Social and cognitive factors driving teamwork in collaborative learning environments: Team learning beliefs and behaviors. *Small Group Research, 37*(5), 490-521.

Van Zundert, M., Sluijsmans, D., & Van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and instruction, 20*(4), 270-279.

Watson, S.L., Watson, W.R., Yu, J.H., Alamri, H., & Mueller, C. (2017). Learner profiles of attitudinal learning in a MOOC: An explanatory sequential mixed methods study. *Computers & Education*, 114, 274-285.

Webb, J.M., Stock, W.A., & McCarthy, M.T. (1994). The effects of feedback timing on learning facts: The role of response confidence. *Contemporary educational psychology, 19*(3), 251-265.

Weinstein, C.E., Acee, T.W., & Jung, J. (2011). Self-regulation and learning strategies. *New Directions for Teaching and Learning, 2011*(126), 45-53.

Winstone, N.E., Nash, R.A., Parker, M., & Rowntree, J. (2017). Supporting Learners' Agentic Engagement With Feedback: A Systematic Review and a Taxonomy of Recipience Processes. *Educational Psychologist, 52*(1), 17-37.

Xiao, Y., & Lucking, R. (2008). The impact of two types of peer assessment on students' performance and satisfaction within a Wiki environment. *The Internet and Higher Education, 11*(3), 186-193.

Yoon, H.B., Park, W.B., Myung, S., Moon, S.H., & Park, J. (2018). Validity and reliability assessment of a peer evaluation method in team-based learning classes. *Korean journal of medical education*, *30*(1), 23-29.

Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, *16*(1), 93.