Peer overmarking and insufficient diagnosticity: The impact of the rating method for peer assessment

Accepted, version pre-print

Van Meenen, Florence[1]*

Coertjens, Liesje[1]

Van Nes, Marie-Claire[2]

Verschuren, Franck[3]

[1]Psychological Sciences Research Institute, Université catholique de Louvain, Belgium

[2]Medical Education Unit, Department of Health Care Sciences, Université catholique de Louvain, Belgium

[3]Institute of Experimental and Clinical Research, Acute medicine department, Université catholique de Louvain, Belgium

*** Corresponding author,**

Email address: florence.vanmeenen@uclouvain.be (F. Van Meenen)

Postal address: 1, Place de l'Université, B-1348 Louvain-la-Neuve, Belgium

ABSTRACT

The present study explores two rating methods for peer assessment (analytical rating using criteria and comparative judgement) in light of concurrent validity, reliability and insufficient diagnosticity (i.e. the degree to which substandard work is recognised by the peer raters). During a second-year undergraduate course, students wrote a one-page essay on an air pollutant. A first cohort (N=260) relied on analytical rating using criteria to assess their peers' essays. A total of 1297 evaluations were made, and each essay received at least four peer ratings. Results indicate a small correlation between peer and teacher marks, and three essays of substandard quality were not recognised by the group of peer raters. A second cohort (N=230) used comparative judgement. They completed 1289 comparisons, from which a rank order was calculated. Results suggest a large correlation between the university teacher marks and the peer scores and acceptable reliability of the rank order. In addition, the three essays of substandard quality were discerned as such by the group of peer raters. Although replication research is warranted, the results provide the first evidence that, when peer raters overmark and fail to identify substandard work using analytical rating with criteria, university teachers may consider changing the rating method of the peer assessment to comparative judgement.

Keywords: peer assessment, analytical rating, comparative judgement, concurrent validity, reliability

INTRODUCTION

The benefits of peer assessment - defined as 'an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status' (Topping, 1998, p. 250) - are well-established (Topping, 2009; Winstone et al., 2017). First and foremost, there are benefits for learning, both for the assessor and the assessee. Peer assessment will, for a specific course, enable the assessor and the assessee to deepen their understanding of the subject matter (Topping et al., 2009; Adachi et al., 2017) and the assessment criteria (Adachi et al., 2018). In the longer run, through multiple opportunities to practice making judgements, students develop their evaluative judgement (Tai et al., 2016), that is, 'the capacity to make decisions about the quality of work of self and others' (Tai et al., 2018, p. 471). This skill is valued in the workplace as it promotes lifelong learning (Rudy et al., 2001; Reiter et al., 2002; Schmidt, 2011; Perera et al., 2010; Papinczak et al., 2007; Tai et al., 2018; Violato and Lockyer, 2006). For the teaching staff, there may be benefits in terms of efficacy: peer assessment may time (Ali et al., 2018).

For medical practitioners, more specifically, being able to assess the work of fellow doctors is increasingly considered important for high-quality health care (CanMEDS; Frank, 2009). It is argued that a medical practitioner must 'function effectively as a mentor and teacher including contributing to the appraisal, assessment and review of colleagues' (GMC, 2009, p. 27) to safeguard or foster the quality of patient care (Chen, 2012; Hulsman et al., 2013).

This highlights the need for medical education to prepare future physicians for such peer assessment throughout their curriculum (Chen, 2012). Indeed, Rudy et al. (2001) recommend that peer assessment start in the first years of medical education to facilitate early acceptance. Because peer assessment of real-world tasks is likely difficult at this early stage in the curriculum, a first step could be structured tasks (Adachi et al., 2017; Rudy et al., 2001).

The emphasis on peer assessment in the workplace is paralleled by growing research on this topic in medical education (Speyer et al., 2011). However, this literature has raised concerns about the inter-rater reliability of peer assessment (Hulsman et al., 2013; Cottrell et al., 2006). For instance, Magzoub and colleagues (1998) detected that, depending on the criteria, up to 17.3% of the variance in marks can be attributed to the peer rater.

In addition, some studies have indicated that there is room for improvement regarding concurrent validity, an aspect of criterion-related validity representing the degree of agreement with an external criterion or standard (Cohen et al., 2008). The meta-analysis conducted by Li and colleagues (2016) shows that the correlation between student and teacher ratings is significantly lower for studies in the medical/clinical domain (Pearson correlation of .33 on average) than for studies in the social sciences/arts (Pearson correlation of .60 on average). Regardless of the domain, near-perfect correlations between peer and teacher ratings are not to be expected because of the complementary view of peers on quality (Altonji et al., 2019; Tai et al., 2016).

Nonetheless, assessors' failure to recognise substandard work is problematic. The weaker correlation found by Li et al. (2016) suggests that this is more likely in the medical/clinical domain. Indeed, multiple studies in medical education have reported peer overmarking (Hoffman et al., 2017; Heyman and Sailors, 2011; O'Brien et al., 2008; Rudy et al., 2001; Lanning et al., 2011), which is defined as peers being more generous than faculty in assigning marks (Falchikov and Goldfinch, 2000) as a results of, among others, over-politeness, fear of retaliation and friendship marking (Lanning et al., 2011; Tekian et al., 2017; Cho et al., 2006; Papinczak et al., 2007). When overmarking is strong, peer assessment marks may end up grouped on the positive end of the scale (Cottrell et al., 2006), which Emke and

colleagues (2017) have labelled 'clumping of marks'. This is not necessarily problematic: if all works meet the required standards, consistent high marks may reflect this. However, in other instances, it leads to substandard work not being recognised, which Heyman and Sailors (2011) refer to as 'insufficient diagnosticity'. In those instances, insufficient diagnosticity will result in missed opportunities for learning (Yepes-Rios et al., 2016; Scarff et al., 2019; Mak-van der Vossen, 2019; Mak-van der Vossen et al., 2018).

It remains unclear how different rating methods for peer assessment foster reliability and concurrent validity and prevent insufficient diagnosticity. There is some evidence, albeit indirect, that the rating method for peer assessment can have an impact on the concurrent validity: in their meta-analysis, Falchikov and Goldfinch (2000) compared the correlation between peer and tutor scores when peers used global judgement or analytical scoring. They found that this correlation depends on the scoring method used: it was higher when peers used global judgements. Consequently, the interplay between the rating method used in peer assessment and reliability and concurrent validity deserves more attention (Cottrell et al., 2006).

Rating methods used in peer assessment

Different rating methods can be used in peer assessment (Pollitt, 2012a; Weigle, 2002; Bacha, 2001). Any rating method should safeguard two key aspects of educational measurement: validity and reliability (Brennan, 2004; Cohen et al., 2008).

Analytical rating using criteria

The predominant method in peer assessment is analytical rating: students receive a list of criteria to use in assessing their peers' work (Li et al., 2016; Norcini, 2003). The students are asked to award a score for each criterion, and these scores are subsequently summed.

The listed criteria should adequately represent the ability or competence assessed, safeguarding the construct validity of the analytical rating (Cohen et al., 2008). Additionally, the criteria direct the raters' attention to the same predefined aspects (Jonsson and Svingby, 2007; Wald et al., 2012; Stemler, 2004), promoting the inter-rater reliability of the peer assessment (Jonsson and Svingby, 2007).

Comparative judgement

The comparative judgement method is based on the assumption that people are more reliable in comparing works than in assigning marks to a single work (Thurstone, 1927), for which the medical education literature provides convincing evidence (Cottrell et al., 2006; Yeates et al., 2013; Yeates et al., 2015). In a comparative judgement exercise, various raters independently compare several pairs

of student works on their overall quality and decide which one is the best with regard to the competence considered. This cancels out differences in rater leniency (Pollitt, 2012a). Every work is compared several times and reviewed by different raters. Based on these judgements, the performances can be ranked from worst to best on an interval scale.

The comparative judgement method does not equate to norm-referenced testing (i.e. 'grading on the curve'), that is scores relative to those of other students, such as 10% pass marks (Haertel, 2006). To the best of our knowledge, in educational settings, comparative judgement has been solely used for criterion-referenced testing, which is also commonplace in competency-based medical education (Lockyer et al., 2017). In criterion-referenced testing, an extra step is needed after determining the rank order. Different options exist, such as a standard-setting exercise to determine a pass/fail boundary (Lesterhuis et al., 2016) or expert grading of two works on the scale followed by the calculation of the grades for the other works (e.g. Settembri et al., 2018).

The comparative judgement logic has been applied to medical education for tutorial evaluation in problem-based learning (PBL) settings (Regehr et al., 1996). However, results regarding the relative ranking of peers in the PBL group indicated low reliability (Reiter et al., 2002). To the best of our knowledge, comparative judgement has not been applied in medical education to evaluate students' performance on structured tasks. However, some studies have examined peer assessments in other fields of higher education, such as veterinary education (Rhind et al., 2017), teacher education (Seery et al., 2012), mathematics education (Jones and Alcock, 2014) and international business education (Bouwer et al., 2018). The results of these studies indicate large correlations (.69, Rhind et al., 2017; .60, Seery et al., 2012; .77, Jones and Alcock, 2014) between the rank orders stemming from the comparative judgement performed peer raters and the marking by experts, supporting the concurrent validity of the rank order generated by peers. Moreover, previous research has found adequate to high reliability levels, ranging from .68 (Verhavert et al., 2019) to .83 (Bouwer et al., 2018).

The present study

The present study explores two rating methods for peer assessment (analytical rating using criteria and comparative judgement) and examines the degree to which substandard work is recognised by the peers.

The following research questions guide the study:

- When applying analytical rating using criteria or comparative judgement, do peer raters detect valid differences in quality?
- If so, are these differences reliable?
- Is there evidence of insufficient diagnosticity?

OVERALL DESIGN OF THE STUDY

This study took place during a second-year undergraduate course on respiratory physiopathology in medicine. Students were instructed to write a one-page essay on an air pollutant (for the instructions, see Appendix 1) and submit an anonymous version. The first cohort of students (spring 2017) rated their peers' essays based on analytical rating using criteria. A second cohort (spring 2018) used comparative judgement. We obtained ethical approval from the ethics review committee of the Cliniques Universitaires Saint-Luc (UCLouvain, Bruxelles) (file number 2019/15OCT/445).

The peer assessment was summative for each cohort: the grades contributed 10% towards the students' overall course grades. Conditions for receiving marks were: (1) writing an essay on an air pollutant, (2) evaluating the requested number of essays and (3) providing constructive feedback on the work of peers. Due to time constraints, the professor was unable to consider the quality of students' evaluations or revised work individually.

To assess the concurrent validity of the peers' assessment, for each cohort, a random sample of 84 essays was drawn by the first author. To determine the optimal sample size (Bujang and Baharum, 2016), we relied on the mean correlation between peer and teacher ratings in medical/clinical settings described in the meta-analysis by Li et al. (2016), that is, .33. The (two times) 84 essays were corrected by the university teacher responsible for the course (last author).

The Kendall's Tau-b correlation coefficient was used to measure the association between teacher marks and peer rating. Kendall Tau-b is a nonparametric alternative to the Pearson correlation coefficient and is appropriate for ordinal variables (Puth et al., 2015; Göktas and Isci, 2011; Gibbons, 1976), such as marks based on a list of criteria (Göktas and Isci, 2011; Howell, 2006; Gibbons, 1976). To allow for comparison with previous studies on concurrent validity and interpretation of the effect size (Cohen, 1988), the Pearson correlation was estimated using the computation and the converting table presented in the literature (Walker, 2003; Gilpin, 1993; Strahan, 1982). Values exceeding .10, .30 and .50 were interpreted as small, medium and large, respectively (Cohen, 1988).

If peer raters detected valid differences in quality, the reliability of the peer raters was assessed. In addition, we explored whether there was evidence of insufficient diagnosticity by comparing the marks

of works that were deemed substandard (i.e. below 50% of the maximum possible mark) by the university teacher with those given by the peer raters.

All analyses were done using the R software, except for the two-tailed Kendall's Tau-b statistic, which was calculated using SPSS 25.

Cohort 1: Analytical rating using criteria

Method

During the spring of 2017, 260 students wrote their essays ($n_{enrolled}$ = 279) and uploaded them to EdX, an online learning platform hosting university-level courses. Students were asked to rate at least four randomly assigned essays. Nine criteria were provided, which were marked on a Likert scale as follows: 1 (completely disagree), 2 (disagree), 3 (moderately agree), 4 (agree) and 5 (completely agree) (see Appendix 2). All 260 students completed this exercise, resulting in 1,297 evaluations.

For each evaluation, the marks on the nine criteria were summed to produce a score ranging from nine to 45. Subsequently, a mark was calculated by taking the average of the sum scores provided by the (at least four) peer raters. Figure 1 depicts the distribution of these marks. All but one mark are grouped on the 11 highest scale points, implying that only the top 29.7% of the available scale is used. Moreover, 85.4% of the marks are situated at the three highest scale points (top 8.1% of the available scale). The kurtosis and skewness values were 14.83 and -2.72 respectively; using the +2/-2 cut-off (George and Mallery, 2010), they indicate a negatively skewed distribution with heavy tails.

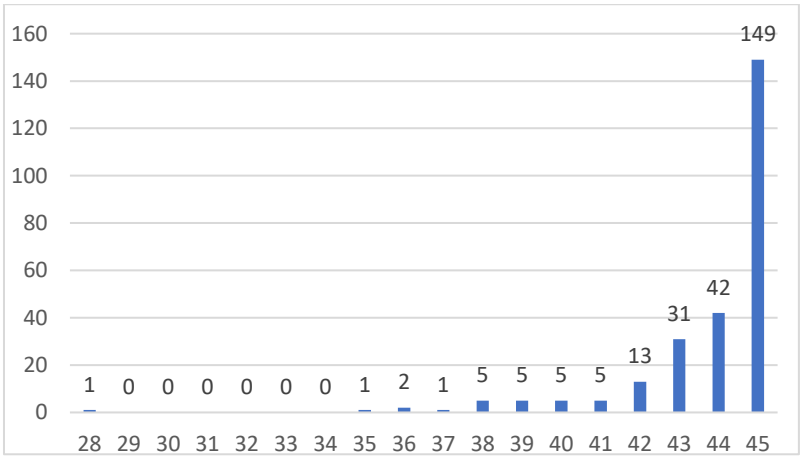

Figure 1 Overview peer assessment marks using analytical rating (N=260) (minimum score = 9)

To evaluate the concurrent validity of the peer assessment, the university teacher used the same nine criteria and Likert scale as the students.

Results

The τb rank correlation between teacher marks and peer rater marks using analytical rating is .174 (p<.05, see Figure 2), which corresponds to a Pearson's r correlation coefficient of .264. This can be interpreted as a small correlation. As evidence for concurrent validity was weak, reliability was not investigated further.

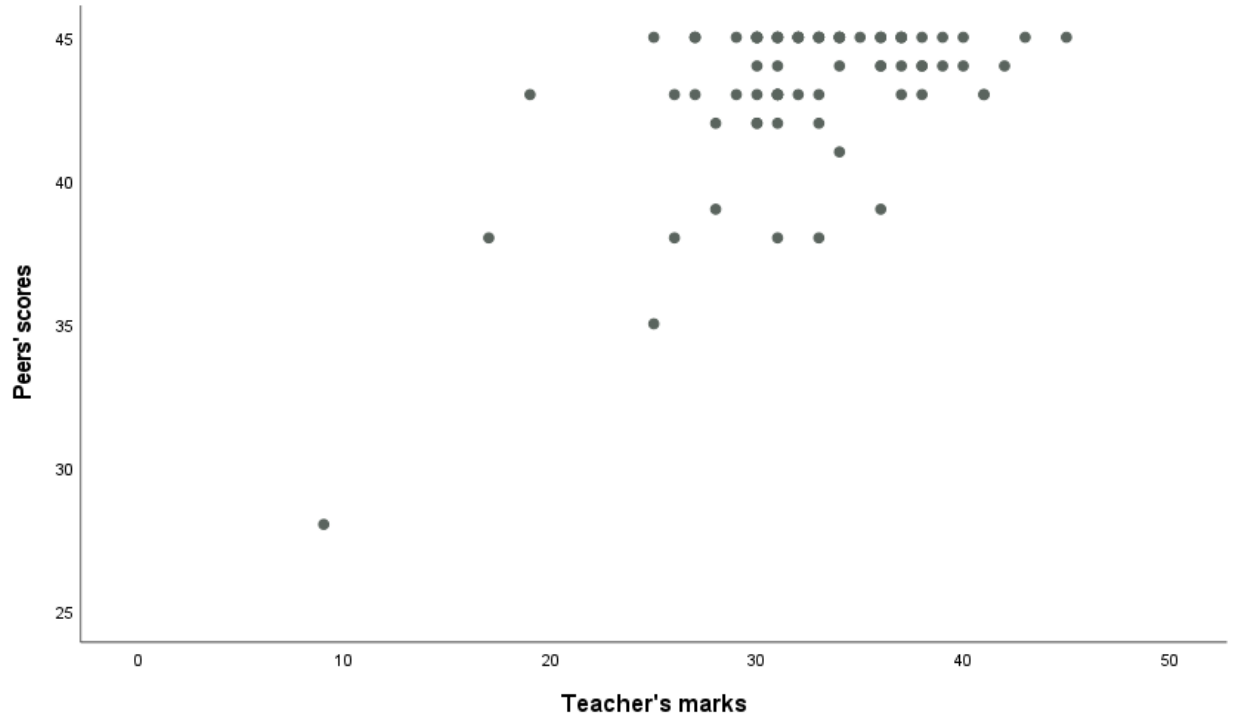

Figure 2 Scatterplot showing the relation between the marks from the university teacher and those from the peer raters using analytical rating (minimum score = 9)

The teacher marks indicate that the quality of three of the 84 works was substandard (with teacher marks of 9, 17 and 19 out of 45, respectively). None of these three essays was identified by the group of students as being of insufficient quality (with peer marks of 28, 38 and 43 out of 45 respectively).

Cohort 2: Comparative judgement

Method

During the spring of 2018, 230 students wrote their essay ($n_{enrolled}$ = 235) and uploaded it to Comproved, a web-based tool supporting an assessment with comparative judgement (comproved.com, Lesterhuis et al., 2016). To assess their peers' essays, students were invited to carefully read the competence description, that is, the nine criteria used by the students of cohort 1 (see Appendix 3). Subsequently, they were presented with a pair of randomly selected essays (excluding their own) and, based on the competence description, answered the question 'Which is the better essay?' by clicking 'Task A' or 'Task B'.

In line with recommendations from a recent meta-analysis (Verhavert et al., 2019), 12 evaluations per essay were chosen, with the aim of reaching a reliability value of .70. Hence, students were asked to complete six comparisons of two essays. A total of 222 students completed the six comparisons, while one student completed five comparisons.

The resulting 1337 judgement decisions were statistically modelled using the Bradley-Terry-Luce model (Verhavert et al., 2018; Lesterhuis et al., 2016), which generated a rank order of the essays from weakest to strongest.

To detect whether the judgement of some peer raters deviated from that of their peers', the infit measure was calculated (Lesterhuis et al., 2016; van Daal et al., 2019). It is common practice to consider an infit exceeding two standard deviations from the mean to be large and significant (Pollitt, 2012b; Jones and Alcock, 2014; Jones et al., 2015), meaning that the rater repeatedly made judgements that deviate from the final rank order (Lesterhuis et al., 2016; van Daal et al., 2019). Eight raters showed such high infit values. In line with a suggestion by Pollitt (2012b) and Lesterhuis et al. (2016), these eight raters and their 48 comparisons were omitted, which resulted in a set of 214 raters having completed 1289 comparisons. This implies that the number of evaluations per essay differed slightly: most essays were evaluated 12, 11 or 10 times (40.4%, 43.6% and 13% of all essays, respectively), but six essays (2.6%) received nine evaluations, and one essay (0.4%) was evaluated eight times (mean=11.2 evaluations per essay).

Based on the set of 1289 comparisons, the rank order and logit scores were calculated for each of the 230 essays (see Figure 3). The logit value (from -5.00 to 5.03) of each essay is represented by a dot, and a confidence interval was created.

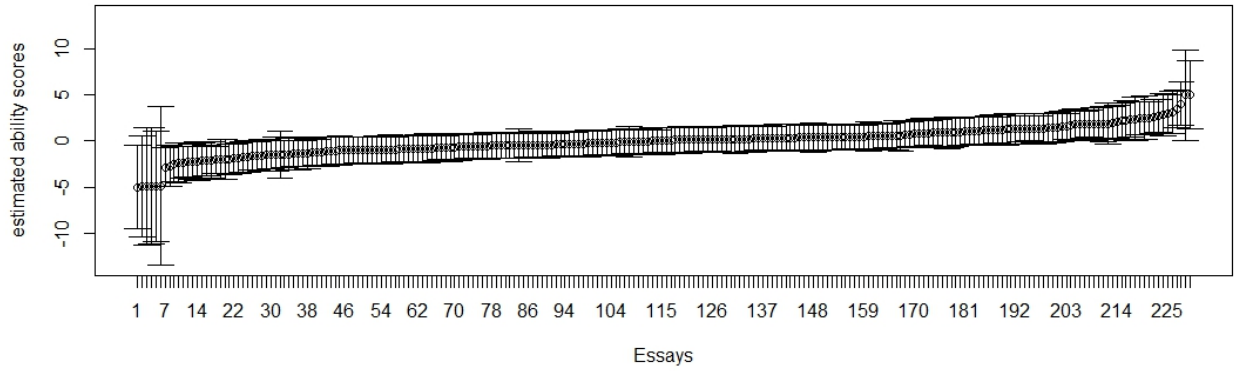

Figure 3 The rank order of the essays

To evaluate the concurrent validity of the peer assessment, the university teacher used the nine criteria (see Appendix 3) and scored each criterion as follows: 0 (disagree, grouping Likert scale items 1 and 2; see appendix 2), 1 (moderately agree) or 2 (agree, grouping Likert scale items 4 and 5). Consequently, possible marks ranged from zero to 18 points.

The Scale Separation Reliability (Bramley, 2015), which can be interpreted as inter-rater reliability (Verhavert et al., 2018), was calculated to establish the reliability of the rank order. Its value ranges from 0 to 1, and a small measurement error (i.e. a value closer to 1) implies that the raters agree on the relative position of the essays in the rank order (Verhavert et al., 2018). As the Scale Separation Reliability is a consistency measure, in line with Cho and Schunn (2018), values exceeding .60 are considered acceptable.

To assess whether there was insufficient diagnosticity, the logit scores from the rank order had to be converted to marks. Following the procedure detailed by Settembri et al. (2018), two essays were selected by the second author: one around position 33% of the rank order (logit score of -0.58) and one around 66% (logit score of 0.61). The university teacher gave these the mark of 14 and 16 out of 20, respectively. The marks for the other essays were calculated using the logit scores from the rank order and ranged from 3.4 to 19.6 out of 20. Figure 4 depicts the distribution of these marks. The kurtosis and skewness values for the distribution of scores are 2.98 and -1.38, respectively. The skewness values indicates no reason for concern, but the distribution has heavy tails.

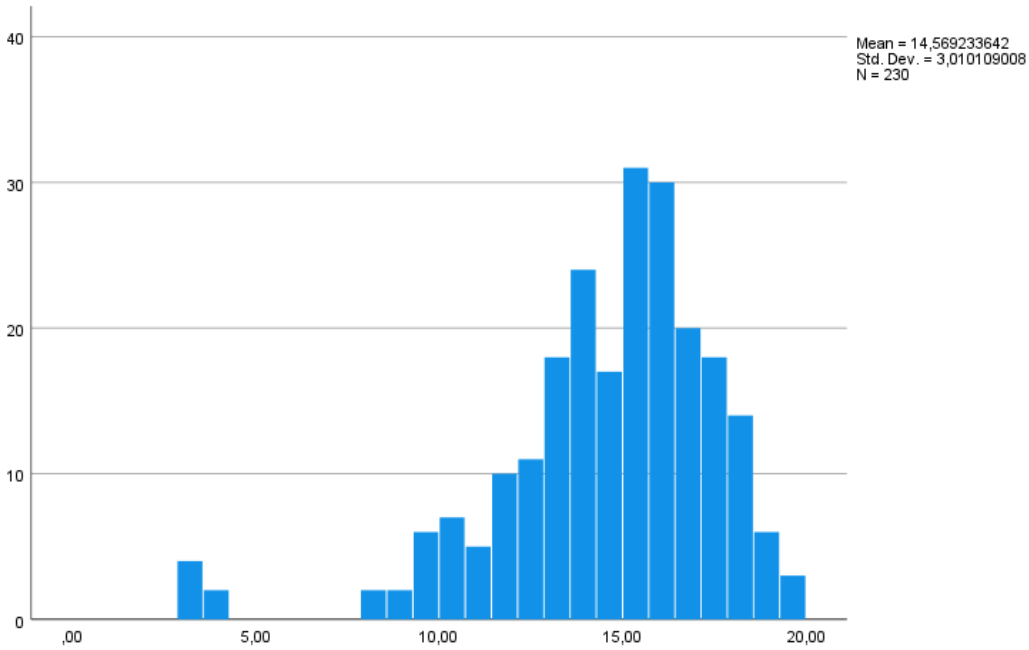

Figure 4 Distribution of marks (on 20) from peer assessment using comparative judgement

Results

The τb rank correlation between teacher marks and the logit scores is .417 (p<.001, see Figure 5), which corresponds to a Pearson's r correlation coefficient of .613. This can be interpreted as a large correlation.[1]

[1] When the comparisons performed by the eight raters showing high infit values were not excluded, the τb rank correlation between teacher marks and the logit scores was .430 (p<.001), which corresponds to a Pearson's r correlation coefficient of .625

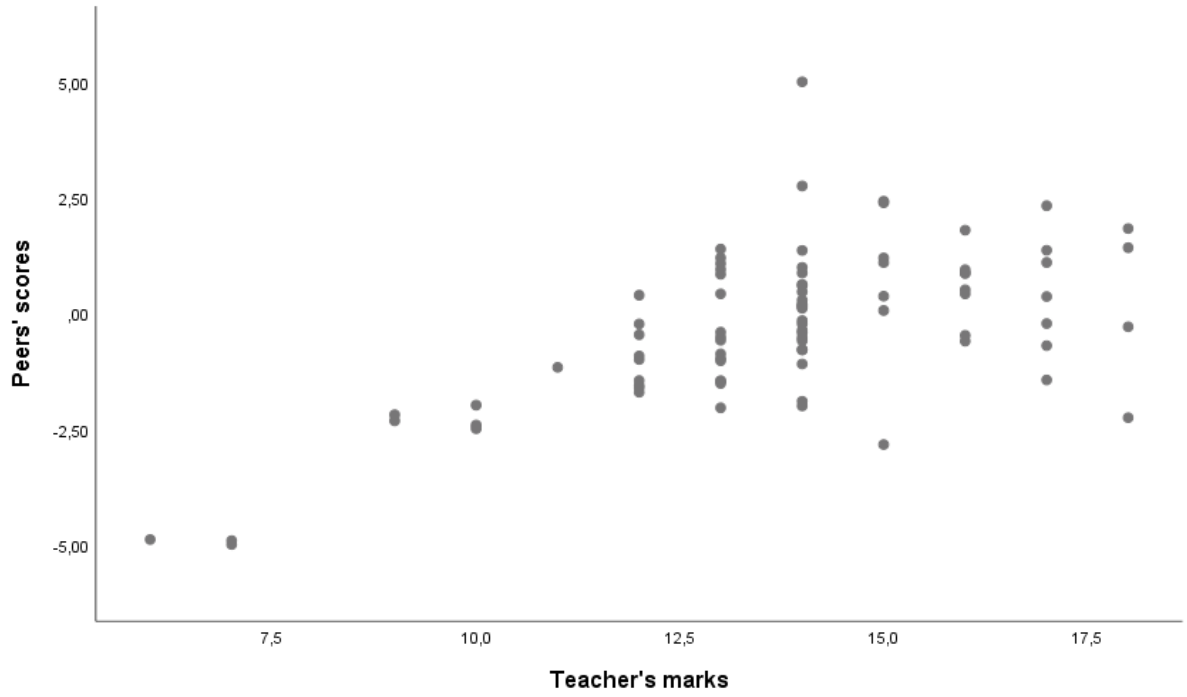

Figure 5 Scatterplot showing the relation between the marks from the university teacher (using analytical rating, minimum score = 0) and the logit scores from the rank order created by the peer raters (using comparative judgement)

To verify whether peer raters agreed on these differences, inter-rater reliability was calculated for the set of 230 essays. The Scale Separation Reliability is .63.

The teacher marks indicate that three works out of 84 were substandard (with marks of 6, 7 and 7 out of 18, respectively). These works also received insufficient marks based on the rank order created by the peer raters (with peer marks of 3.5, 3.5 and 3.4 out of 20 respectively).

DISCUSSION

This study explored two rating methods for peer assessment (analytical rating using criteria and comparative judgement) and examined the degree to which peer raters detected valid quality differences, whether peer ratings were reliable and whether there was evidence of insufficient diagnosticity.

For the first cohort of students, who used analytical rating with a list of criteria, the Pearson correlation between the teacher marks and peer marks was small (.26), which raises concerns about the

concurrent validity of the peer marks (Cohen et al., 2008). This correlation is slightly below the .33 correlation found by Li et al. (2016) in their meta-analysis for studies in medical/clinical settings.

Regarding the research question on insufficient diagnosticity, none of the three essays of insufficient quality (as discerned by the university teacher) received a mark below 50% of the maximum score from the peer raters. Nevertheless, peer assessors did indicate that two out of the three essays were of lower quality than the other essays: they received clearly lower marks (28 and 38 out of 45, respectively), while 85.4% of the marks were clumped in the three highest scale points, is in line with previous findings (Emke et al., 2017; Cottrell et al., 2006; Heyman and Sailors, 2011). This may suggest that (at least some) peer raters recognised that the work was of substandard quality, but did not fail it (Hulsman et al., 2013), as was noted in previous research on student evaluators (Kakar et al., 2013; Iblher et al., 2015). Given that it is not common practice to communicate the distribution of the peer marks (in light of criterion-referenced testing), the students who wrote these essays may have (wrongfully) concluded that their work was of sufficient quality. Consequently, an opportunity for learning was missed (Yepes-Rios et al., 2016; Scarff et al., 2019; Mak-van der Vossen, 2019; Mak-van der Vossen et al., 2018). For university teachers using peer assessment with analytical rating, these findings suggest a recommendation to re-examine the lowest marked works (even if they received a passing mark) and adjust the peer marks if need be.

For the second cohort of students, who used comparative judgement, the Pearson correlation between the teacher marks and the logit scores from the rank order was large (.61). It should be noted that the results were similar when the eight outlier raters were included (.625). This result is in keeping with the .60 correlation in the meta-analysis by Li et al. (2016) with regard to studies in the social sciences/arts and corroborates research findings on comparative judgement in veterinary education (.69, Rhind et al., 2017) and teacher education (.60, Seery et al., 2012).

Importantly, however, some student raters had difficulty in discerning valid quality differences. Eight raters (3.6%) made numerous judgements contradicting the consensus among the peer raters. The literature has so far offered few insights into the reasons for this misfit: Were students distracted? Did they favour their friends' essays (Lanning et al., 2011; Cho et al., 2006; Papinczak et al., 2007)? Or did they genuinely have a different view on quality? We found only one study examining misfit, but it focused on experts (van Daal et al., 2019). More research on misfit in peer assessment is evidently needed (Pollitt, 2012a) regarding both analytical rating and comparative judgement, as it will help to unveil how this small group of students can be coached to improve their evaluative judgement (Tai et al., 2016, 2018), which is key for lifelong learning (Rudy et al., 2001; Lew and Schmidt, 2011; Perera et

al., 2010; Papinczak et al., 2007). A first step could be to examine whether these misfitting students' written feedback differs from the feedback provided by the other peer raters.

Although we aimed to reach a reliability level of .70, a Scale Separation Reliability of .63 was obtained, which is considered acceptable. This contrasts with the low reliability level found by Reiter et al. (2002), who used comparative judgement to provide tutorial evaluation in problem-based learning groups. The reliability of comparative judgement may vary with the goal of the peer assessment (Norcini, 2003), namely, providing global impressions of colleagues or evaluating the latter's performance on structured tasks. The objective of .70 reliability was not reached due to the lower number of comparisons made: eight students did not make any comparisons and an additional eight raters showed high infit values. As the number of comparisons is the key determinant of the level of reliability (Verhavert et al., 2019), future peer assessment using comparative judgement could anticipate the drop-out and infit of raters by requesting for example, two evaluations more per task than the required number (e.g. 14 evaluations when 12 are needed).

Regarding the research question on insufficient diagnosticity for the cohort using comparative judgement, the three works that received insufficient marks from the university teacher also received a mark below 50% of the maximum score based on the rank order. It should be noted that to convert the logit scores from the rank order into marks, two essays were selected (around positions 33% and 66%, respectively, following Settembri et al., 2018) and graded by the university teacher. The peer marks for the other essays (including the three works of insufficient quality according to the university teacher) were calculated using the logit scores from the rank order.

It is important to keep in mind that students undertake different tasks for each rating method. In analytical rating, students suggest a mark for each work, from which the average is then taken. In comparative judgement, students collectively create a rank order by indicating the better work of multiple pairs of work. Thus, both rating methods cannot be compared directly. When confronted with clumping of peer assessment marks at the high end of the scale, which is not considered to be an adequate reflection of the quality of the works by the teaching staff and leads to substandard work being unrecognised (Heyman and Sailors, 2011), a teacher can first use the habitual strategies to prevent peer overmarking, namely use anonymous works and multiple (at least four) peer raters (Vickerman, 2009). If these strategies do not alleviate the problem (as evidenced in the present study and prior research by Heyman and Sailors, 2011), it may be useful, as Falchikov and Goldfinch (2000) indicated, to change the rating method used for peer assessment. The results of the present study

suggest that comparative judgement could be an alternative, as it reduces problems resulting from rater leniency (Pollitt, 2012a); however, replication research is needed to confirm this.

For practitioners of peer assessment in medical education this implies that, depending on the goals of student learning and learners' characteristics, one or the other rating method may be more advantageous. Using comparative judgement, students can learn to rank works on their quality in a holistic way using the competence description. Nonetheless, they do not develop their ability to identify substandard work. Using an analytical rating approach with criteria, students assess each criterion and may identify substandard work. Consequently, analytical rating is generally considered to be a more difficult task than comparing two works (Lesterhuis et al., 2016; Pollitt, 2012b), and requires training (Gielen et al., 2011; Tekian et al., 2017).

Hence, in a medical education curriculum, the comparative judgement method may be adequate to develop students' first steps in evaluative judgement: the method can be used by learners with little experience with peer- or self-assessment (Verhavert et al., 2019) and does not require training (Lesterhuis et al., 2016), which can be difficult to organise given the large size of student groups in the first years. However, for medical professionals, to be able to identify substandard work by colleagues in the workplace (CanMEDS; Frank, 2009; GMC, 2009), to safeguard or foster the quality of patient care (Chen, 2012; Hulsman et al., 2013), the medical education curriculum should allow students to further develop their evaluative judgement in later years using analytical rating with criteria and training in the use of these criteria.

Three additional differences between the two methods should be noted. The first is the length of the works. In the present study, the peers assessed short essays. In the cohort using analytical rating, the students assessed five essays, whereas in the cohort using comparative judgement, they completed six comparisons, thus reading 12 works in total. If works are lengthy (e.g. multiple pages), the time investment for students using comparative judgement rapidly becomes significant. The number of comparisons could be reduced in such cases, but this will strongly impact the level of reliability (Verhavert et al., 2019). In other words, comparative judgement appears less apt when works are lengthy. A second aspect concerns the marks: comparative judgement does not automatically lead to marks, which are often required in university contexts. In the present study, we followed the procedure detailed by Settembri and colleagues (2018), which requires an extra step and time. If marks need to be provided to the students rapidly, comparative judgement may be less practical.

A third aspect that is ondoubtly key for practitioners is the quality of the written feedback. This requires future research, however, as the feedback in the first cohort of students was very succinct. This was

possibly due to the fact that students were asked to provide this feedback below their rating of the nine criteria (see Appendix 2): some may have missed this text field or felt that they had already provided feedback by rating the nine criteria. An alternative reason, in line with recent findings on teacher raters (Coertjens et al., 2021), is that comparative judgement fosters detailed feedback. In the present study, the prompts for feedback differed too significantly between the two cohorts (i.e. a box below the criteria versus two separate boxes on a new webpage, one for strong points and another for suggestions for improvement). Therefore, further research is needed to explore whether analytical rating and comparative judgement differ in the quality of the written feedback.

Limitations

Three limitations of the present study must be acknowledged. First, instead of randomly assigning students to the two assessment methods, two cohorts of students were used. Although the results of other assessments of the respiratory physiopathology course did not show any significant differences between the two cohorts, groups may have differed in their motivation or their judgement skills. Future research should consider the random distribution of peer assessment methods (Creswell, 2012).

Second, while the study relied on analytical rating using criteria, current literature suggests the use of a rubric with descriptors (Vickerman, 2009). Indeed, the criteria (see Appendix 2) may have provided too little support to students (thus undermining construct validity), and descriptors should have been formulated to represent the degree to which criterion-referenced standards were met. It must be noted, however, that peer overmarking has been observed when using rubrics as well (Lee et al., 2021; O'Brien et al., 2008). Consequently, replication research using rubric rating is warranted. In addition, more research is needed on how students interpret the analytical rating scale when overmarking. Such in-depth qualitative research on students' evaluative judgement (for example using think aloud or Cued Retrospective Reporting, Catrysse et al., 2017; Brand-Gruwel et al., 2017) when students using different rating methods could provide insight into their thought processes and motivation while assessing peers' work.

Third, comparative judgement is mostly used with raters who receive the competence description, but no additional training (McMahon and Jones, 2015; Bouwer et al., 2018; Pollitt, 2012b). For analytical rating, however, additional training (e.g. evaluating a few papers in class) has been recommended (Gielen et al., 2011; Norcini, 2003; Tekian et al., 2017) as it is believed to be beneficial for construct validity. Although research has shown that, even with such training, differences in leniency between

student raters and overmarking persist (Kakar et al., 2013; Aryadoust, 2016), the lack of training may have worsened the clumping of marks and the insufficient diagnosticity in the present study. Therefore, future research should consider three conditions: comparative judgement and analytical rating with and without training.

CONCLUSION

The present study explored two rating methods for peer assessment, namely analytical rating using criteria and comparative judgement. Results suggest that when strategies commonly used to prevent peer overmarking and insufficient diagnosticity do not alleviate the problem, university teachers may consider changing the rating method to comparative judgement. As the present study is the first to explore multiple rating methods for peer assessment in medical education with regard to concurrent validity, reliability and insufficient diagnosticity, replication research is necessary.

REFERENCES

Adachi, C., Tai, J., H., M. & Dawson, P. (2017): Academics' perceptions of the benefits and challenges of self and peer assessment in higher education. Assessment & Evaluation in Higher Education, 43(2), 294-306. doi: 10.1080/02602938.2017.1339775

Altonji, S. J., Baños, J. H., & Harada, C. N. (2019). Perceived benefits of a peer mentoring program for first-year medical students. *Teaching and Learning in Medicine, 31*(4), 445-452. doi:10.1080/10401334.2019.1574579

Aryadoust, V. (2016). Gender and Academic Major Bias in Peer Assessment of Oral Presentations Language Assessment Quarterly, 13 (1), 1-24. DOI 10.1080/15434303.2015.1133626

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? System, 29 (3), 371-383. DOI 10.1016/S0346-251X(01)00025-2

Bouwer, R., Lesterhuis, M., Bonne, P. & De Maeyer, S. (2018). Applying Criteria to Examples or Learning by Comparison: Effects on Students' Evaluative Judgment and Performance in Writing (Original Research). Frontiers in Education, 3 (86). DOI 10.3389/feduc.2018.00086

Bramley, T. (2015) 'Investigating the reliability of Adaptive Comparative Judgment'. Cambridge: Cambridge Assessment.

Brand-Gruwel, S., Kammerer, Y., van Meeuwen, L., & van Gog, T. (2017). Source evaluation of domain experts and novices during Web search. Journal of Computer Assisted Learning, 33(3), 234-251. DOI:10.1111/jcal.12162

Brennan, R. L. (Ed.) (2004). Educational measurement. (Westport, CT: American Council on Education/Praeger)

Bujang, M. A. & Baharum, N. (2016). Sample Size Guideline for Correlation Analysis. World Journal of Social Science Research, 3 (1), 37-46.
Catrysse, L., Gijbels, D., Donche, V., De Maeyer, S., Lesterhuis, M., & Van den Bossche, P. (2017). How are learning strategies reflected in the eyes? Combining results from self-reports and eye-tracking. Br J Educ Psychol, 88(1). DOI: 10.1111/bjep.12181
Chen, J. Y. (2012). Why peer evaluation by students should be part of the medical school learning environment. Medical Teacher, 34 (8), 603-606. DOI 10.3109/0142159X.2012.689031
Cho, K. & Schunn, C. D. (2018). Finding an optimal balance between agreement and performance in an online reciprocal peer evaluation system. Studies in Educational Evaluation, 56, 94-101. DOI https://doi.org/10.1016/j.stueduc.2017.12.001
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. DOI 10.1037/0022-0663.98.4.891
Coertjens, L., Lesterhuis, M., De Winter, B. Y., Goossens, M., De Maeyer, S., & Michels, N. R. M. (2021). Improving Self-Reflection Assessment Practices: Comparative Judgment as an Alternative to Rubrics. Teaching and Learning in Medicine, 33(5), 525-535. DOI: 10.1080/10401334.2021.1877709
Cohen, J. (1988). Statistical power for the behavioral sciences. (New Jersey, NJ: Lawrence Erlbauw Associates)
Cohen, L., Manion, L. & Morrison, K. (2008). Research methods in education. (London: Routledge)
Cottrell, S., Diaz, S., Cather, A. & Shumway, J. (2006). Assessing Medical Student Professionalism: An Analysis of a Peer Assessment. Medical Education Online, 11 (1), 4587. DOI 10.3402/meo.v11i.4587
Creswell, J. W. (2012). Educational Research: Planning, conducting, and evaluating quantitative and qualitative research. (Boston, USA: Pearson Education)
Emke, A. R., Cheng, S., Chen, L., Tian, D. & Dufault, C. (2017). A Novel Approach to Assessing Professionalism in Preclinical Medical Students Using Multisource Feedback Through Paired Self- and Peer Evaluations. Teaching and Learning in Medicine, 29 (4), 402-410. DOI 10.1080/10401334.2017.1306446
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. DOI 10.3102/00346543070003287
Frank, J. R. (2009) 'The CanMEDS 2005 physician competency framework. Better standards. Better physicians. Better care.'. Ottawa: The Royal College of Physicians and Surgeons of Canada.
George, D. and Mallery, P. (2010). *SPSS for Windows Step by Step: A Simple Guide and Reference 17.0 Update*. Pearson.
Gibbons, J. D. (1976). Nonparametric methods for quantitative analysis. (New York, NY: Holt, Rinehart and Winston)
Gielen, S., Dochy, F. & Onghena, P. (2011). An inventory of peer assessment diversity. Assessment & Evaluation in Higher Education, 36 (2), 137-155. DOI 10.1080/02602930903221444
Gilpin, A. R. (1993). Table for Conversion of Kendall'S Tau to Spearman'S Rho Within the Context of Measures of Magnitude of Effect for Meta-Analysis. Educational and Psychological Measurement, 53 (1), 87-92. DOI 10.1177/0013164493053001007
Ginsburg, S., Vleuten, C. P., Eva, K. W. & Lingard, L. (2017). Cracking the code: residents' interpretations of written assessment comments. Medical Education, 51 (4), 401-410. DOI doi:10.1111/medu.13158
GMC (2009) 'Tomorrow's Doctors. Outcomes and standards for undergraduate medical education'. London, UK: General Medical Council.
Göktas, A. & Isci, O. (2011). A Comparison of the Most Commonly Used Measures of Association for Doubly Ordered Square Contingency Tables via Simulation. Metodološki zvezki, 8 (1), 17-37.
Haertel, E. H. (2006). Reliability. (In R. L. Brennan (Ed.), Educational measurement. Westport, CT: American Council on Education/Praeger)

Heyman, J. E. & Sailors, J. J. (2011). Peer assessment of class participation: applying peer nomination to overcome rating inflation. Assessment & Evaluation in Higher Education, 36 (5), 605-618. DOI 10.1080/02602931003632365
Hoffman, L. A., Shew, R. L., Vu, T. R., Brokaw, J. J. & Frankel, R. M. (2017). The Association Between Peer and Self-Assessments and Professionalism Lapses Among Medical Students. Evaluation & the Health Professions, 40 (2), 219-243. DOI 10.1177/0163278717702191
Howell, D. C. (2006). Statistical methods for psychology. (Belmont, CA: Cengage learning)
Hulsman, R. L., Peters, J. F. & Fabriek, M. (2013). Peer-assessment of medical communication skills: The impact of students' personality, academic and social reputation on behavioural assessment. Patient Education and Counseling, 92 (3), 346-354. DOI https://doi.org/10.1016/j.pec.2013.07.004
Iblher, P., Zupanic, M., Karsten, J. & Brauer, K. (2015). May student examiners be reasonable substitute examiners for faculty in an undergraduate OSCE on medical emergencies? Medical Teacher, 37 (4), 374-378. DOI 10.3109/0142159X.2014.956056
Jones, I. & Alcock, L. (2014). Peer assessment without assessment criteria. Studies in Higher Education, 39 (10), 1774-1787. DOI 10.1080/03075079.2013.821974
Jones, I., Swan, M. & Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement (journal article). International Journal of Science and Mathematics Education, 13 (1), 151-177. DOI 10.1007/s10763-013-9497-6
Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review, 2 (2), 130-144. DOI 10.1016/j.edurev.2007.05.002
Kakar, S. P., Catalanotti, J. S., Flory, A. L., Simmens, S. J., Lewis, K. L., et al. (2013). Evaluating Oral Case Presentations Using a Checklist: How Do Senior Student-Evaluators Compare With Faculty? Academic Medicine, 88 (9), 1363-1367. DOI 10.1097/ACM.0b013e31829efed3
Lanning, S. K., Brickhouse, T. H., Gunsolley, J. C., Ranson, S. L. & Willett, R. M. (2011). Communication skills instruction: An analysis of self, peer-group, student instructors and faculty assessment. Patient Education and Counseling, 83 (2), 145-151. DOI https://doi.org/10.1016/j.pec.2010.06.024
Lee, K. Y., Hassell, D., Salleh, S.M., & Munohsamy, T.(2021). Online-based Rubric for Peer Assessment: Effectiveness and Implications. *Southeast Asia Language Teaching and Learning Journal, 2*(4). doi: https://doi.org/1035307/saltel.v4i2.76 e-ISSN: 2614-2684
Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V. & De Maeyer, S. (2016). Comparative judgement as a promising alternative to score competences. (In G. Ion & E. Cano (Eds.), Innovative Practices for Higher Education Assessment and Measurement. Hershey, PA: IGI Global)
Lew, M. D. N. & Schmidt, H. G. (2011). Self-reflection and academic performance: is there a relationship? (journal article). Advances in Health Sciences Education, 16 (4), 529. DOI 10.1007/s10459-011-9298-z
Li, H., Xiong, Y., Zang, X., L. Kornhaber, M., Lyu, Y., et al. (2016). Peer assessment in the digital age: a meta-analysis comparing peer and teacher ratings. Assessment & Evaluation in Higher Education, 41 (2), 245-264. DOI 10.1080/02602938.2014.999746
Lockyer, J., Carraccio, C., Chan, M.-K., Hart, D., Smee, S., et al. (2017). Core principles of assessment in competency-based medical education. Medical Teacher, 39 (6), 609-616. DOI 10.1080/0142159X.2017.1315082
Magzoub, M. E. M. A., Abdelhameed, A. A., Schmidt, H. G. & Dolmans, D. H. J. M. (1998). Assessing Students in Community Settings: The Role of Peer Evaluation (journal article). Advances in Health Sciences Education, 3 (1), 3. DOI 10.1023/a:1009786129941
Mak-van der Vossen, M. (2019). 'Failure to fail': the teacher's dilemma revisited. Medical Education, 53 (2), 108-110. DOI 10.1111/medu.13772
Mak-van der Vossen, M., de la Croix, A., Teherani, A., van Mook, W. N. K. A., Croiset, G., et al. (2018). Developing a two-dimensional model of unprofessional behaviour profiles in medical students (journal article). Advances in Health Sciences Education. DOI 10.1007/s10459-018-9861-y

McMahon, S. & Jones, I. (2015). A comparative judgement approach to teacher assessment. Assessment in Education: Principles, Policy & Practice, 22 (3), 368-389. DOI 10.1080/0969594X.2014.978839
Norcini, J. J. (2003). Peer assessment of competence. Medical Education, 37 (6), 539-543. DOI 10.1046/j.1365-2923.2003.01536.x
O'Brien, C. E., Franks, A. M. & Stowe, C. D. (2008). Multiple rubric-based assessments of student case presentations. American journal of pharmaceutical education, 72 (3), 58-58.
Papinczak, T., Young, L. & Groves, M. (2007). Peer assessment in problem-based learning: a qualitative study. Adv Health Sci Educ Theory Pract, 12 (2), 169-86. DOI 10.1007/s10459-005-5046-6
Perera, J., Mohamadou, G. & Kaur, S. (2010). The use of objective structured self-assessment and peer-feedback (OSSP) for learning communication skills: evaluation using a controlled trial (journal article). Advances in Health Sciences Education, 15 (2), 185-193. DOI 10.1007/s10459-009-9191-1
Pollitt, A. (2012a). Comparative judgement for assessment (journal article). International Journal of Technology and Design Education, 22 (2), 157-170. DOI 10.1007/s10798-011-9189-x
Pollitt, A. (2012b). The method of Adaptive Comparative Judgement. Assessment in Education: Principles, Policy & Practice, 19 (3), 281-300. DOI 10.1080/0969594X.2012.665354
Puth, M.-T., Neuhäuser, M. & Ruxton, G. D. (2015). Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits. Animal Behaviour, 102, 77-84. DOI https://doi.org/10.1016/j.anbehav.2015.01.010
Regehr, G., Hodges, B., Tiberius, R. & Lofchy, J. (1996). Measuring self-assessment skills: an innovative relative ranking model. Academic Medicine, 71 (10), S52-4.
Reiter, H. I., Eva, K. W., Hatala, R. M. & Norman, G. R. (2002). Self and Peer Assessment in Tutorials: Application of a Relative-ranking Model. Academic Medicine, 77 (11), 1134-1139.
Rhind, S. M., Hughes, K. J., Yool, D., Shaw, D., Kerr, W., et al. (2017). Adaptive Comparative Judgment: A Tool to Support Students' Assessment Literacy. Journal of Veterinary Medical Education, 44 (4), 686-691. DOI 10.3138/jvme.0616-113R
Rudy, D. W., Fejfar, M. C., Griffith, C. H. & Wilson, J. F. (2001). Self- and Peer Assessment in a First-Year Communication and Interviewing Course. Evaluation & the Health Professions, 24 (4), 436-445. DOI 10.1177/016327870102400405
Sargeant, J., Eva, K. W., Armson, H., Chesluk, B., Dornan, T., et al. (2011). Features of assessment learners use to make informed self-assessments of clinical performance. Medical Education, 45 (6), 636-647. DOI doi:10.1111/j.1365-2923.2010.03888.x
Scarff, C. E., Bearman, M., Chiavaroli, N. & Trumble, S. (2019). Keeping mum in clinical supervision: private thoughts and public judgements. Medical Education, 53 (2), 133-142. DOI 10.1111/medu.13728
Seery, N., Canty, D. & Phelan, P. (2012). The validity and value of peer assessment using adaptive comparative judgement in design driven practical education (journal article). International Journal of Technology and Design Education, 22 (2), 205-226. DOI 10.1007/s10798-011-9194-0
Settembri, P., Van Gasse, R., Coertjens, L. & De Maeyer, S. (2018). Oranges and Apples? Using Comparative Judgement for Reliable Briefing Paper Assessment in Simulation Games. (In P. Bursens, V. Donche, D. Gijbels & P. Spooren (Eds.), Simulations of Decision-Making as Active Learning Tools: Design and Effects of Political Science Simulations. (pp. 93-108). Cham: Springer International Publishing)
Speyer, R., Pilz, W., Van Der Kruis, J. & Brunings, J. W. (2011). Reliability and validity of student peer assessment in medical education: A systematic review. Medical Teacher, 33 (11), e572-e585. DOI 10.3109/0142159X.2011.610835
Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research & Evaluation, 9 (4), 1-11.
Strahan, R. F. (1982). Assessing Magnitude of Effect from Rank-Order Correlation Coefficients. Educational and Psychological Measurement, 42 (3), 763-765. DOI 10.1177/001316448204200306

Tai, J. H., Canny, B. J., Haines, T. P., & Molloy, E. K. (2016). The role of peer-assisted learning in building evaluative judgement: opportunities in clinical medical education. Advances in Health Sciences Education Theory Pract, 21(3), 659-676. doi: 10.1007/s10459-015-9659-0

Tai, J., Ajjawi, R., Boudt, D., Dawson, P., & Panadera, E. (2018). Developing evaluative judgement: enabling students to make decisions about the quality of work. Higher Education, 76, 467-480. DOI: 10.1007/s10734-017-0220-3

Tekian, A., Watling, C. J., Roberts, T. E., Steinert, Y. & Norcini, J. (2017). Qualitative and quantitative feedback in the context of competency-based education. Medical Teacher, 39 (12), 1245-1249. DOI 10.1080/0142159X.2017.1372564

Thurstone, L. L. (1927). A law of comparative judgment. Psychological review, 34 (4), 273.

Topping, K. (1998). Peer Assessment Between Students in Colleges and Universities. Review of Educational Research, 68 (3), 249-276. DOI 10.3102/00346543068003249

Topping, K. (2009). Peer Assessment. Theory Into Practice, 48(1), 20-27. doi: 10.1080/00405840802577569

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V. & De Maeyer, S. (2019). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. Assessment in Education: Principles, Policy & Practice, 26 (1), 59-74. DOI 10.1080/0969594X.2016.1253542

Verhavert, S., Bouwer, R., Donche, V. & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. Assessment in Education: Principles, Policy & Practice, 26 (5), 541-562. DOI 10.1080/0969594X.2019.1602027

Verhavert, S., De Maeyer, S., Donche, V. & Coertjens, L. (2018). Scale Separation Reliability: What Does It Mean in the Context of Comparative Judgment? Applied Psychological Measurement, 42 (6), 428-445. DOI 10.1177/0146621617748321

Vickerman, P. (2009). Student perspectives on formative peer assessment: an attempt to deepen learning? Assessment & Evaluation in Higher Education, 34 (2), 221-230. DOI 10.1080/02602930801955986

Violato, C. & Lockyer, J. (2006). Self and peer assessment of pediatricians, psychiatrists and medicine specialists: implications for self-directed learning. Adv Health Sci Educ Theory Pract, 11 (3), 235-44. DOI 10.1007/s10459-005-5639-0

Wald, H. S., Borkan, J. M., Taylor, J. S., Anthony, D. & Reis, S. P. (2012). Fostering and Evaluating Reflective Capacity in Medical Education: Developing the REFLECT Rubric for Assessing Reflective Writing. Academic Medicine, 87 (1), 41-50. DOI 10.1097/ACM.0b013e31823b55fa

Walker, D. A. (2003). Converting Kendall's Tau for correlational or meta-analytic analysis. Journal of modern applied statistical methods, 2 (2), 525-530. DOI 10.22237/jmasm/1067646360

Weigle, S. C. (2002). Assessing writing. (Cambridge: Cambridge University Press)

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. doi:10.1080/00461520.2016.1207538

Yeates, P., Cardell, J., Byrne, G. & Eva, K. W. (2015). Relatively speaking: contrast effects influence assessors' scores and narrative feedback. Medical Education, 49 (9), 909-919. DOI 10.1111/medu.12777

Yeates, P., O'Neill, P., Mann, K. & W Eva, K. (2013). 'You're certainly relatively competent': assessor bias due to recent experiences. Medical Education, 47 (9), 910-922. DOI 10.1111/medu.12254

Yepes-Rios, M., Dudek, N., Duboyce, R., Curtis, J., Allard, R. J., et al. (2016). The failure to fail underperforming trainees in health professions education: A BEME systematic review: BEME Guide No. 42. Medical Teacher, 38 (11), 1092-1099. DOI 10.1080/0142159X.2016.1215414

Appendix 1. Task instructions

There are numerous air pollutants. Choose one air pollutant that affects the respiratory system (e.g., particles emitted by the automobile industry, ozone that irritates the respiratory tract, sulphur-based gases or nitrogen gas, chlorine from swimming pools, smoke or dust from a sandstorm, the constituents of a cigarette, pollutants inside buildings …)
Undertake some research about this pollutant:

- What kind of pollutant is it? A gas, a particle, a dust?
- What is the source of emission? Natural, industrial?
- If it is a particle, what is its diameter?
- Where in the respiratory tract does this pollutant act?
- What is its mechanism of action related to those seen in the videos (asthma, emphysema, ...)
- Can I trust the medical literature that I have found? What makes me doubt? What makes me trust it?
- What is its effect on health?
- What advice or prevention can we offer?
- What is your personal opinion on the risks described for this air pollutant?

Write a text of 20 lines summarizing your knowledge about this pollutant.

Your 20 lines text will then be reviewed by peers who will judge it according to different criteria:

- Clarity, comprehension, and level of detail of the text
- The link with the content seen in the MOOC videos (at least 3 different links)
- The respect of the instructions provided, the authors' ability to synthesise
- The citation of the sources used
- The degree to which the text is interesting and pleasant to read

Appendix 2. Instructions for assessing for cohort 1

On the EdX platform, the following criteria were presented to the students:

1. The text provides clear and understandable information
2. The text provides detailed information
3. The text makes at least 3 links with the elements covered in MOOC videos
4. The author follows the instructions provided
5. The author did a good job in synthesizing the information
6. The author references the bibliographic sources used
7. I was interested to read this text
8. The length of the text was pleasant (not too long, not too short)
9. I was positively impressed by this text

Each of the criteria was judged on a 5-point Likert scale (1 (completely disagree), 2 (disagree), 3 (moderately agree), 4 (agree) and 5 (completely agree)).

Each student was instructed to assessed at least 4 essays.

Appendix 3. Instructions for assessing for cohort 2

On the Comproved platform, the following information was presented to the students:

- The first step in evaluating the essays is to **compare the essays in pairs**. So, read both essays taking into account the judgment criteria below:

1. The text provides clear and understandable information
2. The text provides detailed information
3. The text makes at least 3 links with the elements covered in MOOC videos
4. The author follows the instructions provided
5. The author did a good job in synthesizing the information
6. The author references the bibliographic sources used
7. I was interested to read this text
8. The length of the text was pleasant (not too long, not too short)
9. I was positively impressed by this text

- **Repeat the process 6 times.** Each student will thus receive feedback from 6 people, and the copies of the entire class will be rank ordered.