# Improving self-reflection assessment practices : comparative judgment as an alternative to rubrics

**Reference:**
Coertjens Liesje, Lesterhuis Marije, De Winter Benedicte, Goossens Maarten, De Maeyer Sven, Michels Nele.- Improving self-reflection assessment practices : comparative judgment as an alternative to rubrics
Teaching and learning in medicine - ISSN 1040-1334 - 33:5(2021), p. 525-535
Full text (Publisher's DOI): https://doi.org/10.1080/10401334.2021.1877709
To cite this reference: https://hdl.handle.net/10067/1756510151162165141

# Improving Self-Reflection Assessment Practices: Comparative Judgment as an Alternative to Rubrics

Coertjens, Liesje[a b], Lesterhuis, Marije[b] , De Winter, Benedicte Y.[c], Goossens, Maarten[b], De Maeyer, Sven[b], Michels, Nele R.M.[c]

[a] Psychological Sciences Research Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium ; [b] Department of Educational Sciences, Faculty of Social Sciences, University of Antwerp, Antwerp, Belgium; [c] Skills lab at the Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium

Contact: Liesje Coertjens Liesje.coertjens@uclouvain.be
1, Place de l'Université, B-1348 Louvain-la-Neuve, Belgium

**Improving self-reflection assessment practices: Comparative judgment as an alternative to rubrics**

## *Abstract*

*Construct.* The authors aimed to investigate the utility of the comparative judgment method for assessing students' written self-reflections.

*Background.* Medical practitioners' reflective skills are increasingly considered important and therefore included in the medical education curriculum. However, assessing students' reflective skills using rubrics does not appear to guarantee adequate inter-rater reliabilities. Recently, comparative judgment was introduced as a new method to evaluate performance assessments. This study investigates the merits and limitations of the comparative judgment method for assessing students' written self-reflections. More specifically, it examines the reliability in relation to the time spent assessing, the correlation between the scores obtained using the two methods (rubrics and comparative judgment), and, raters' perceptions of the comparative judgment method. *Approach.* Twenty-two self-reflections, that had previously been scored using a rubric, were assessed by a group of eight raters using comparative judgment. Two hundred comparisons were completed and a rank order was calculated. Raters' impressions were investigated using a focus group. *Findings.* Using comparative judgment, each self-reflection needed to be compared seven times with another self-reflection to reach a scale separation reliability of .55. The inter-rater reliability of rating (ICC, (1, k)) using rubrics was .56. The time investment required for these reliability levels in both methods was around 24 minutes. The Kendall's tau rank correlation indicated a strong correlation between the scores obtained via both methods. Raters reported that making comparisons made them evaluate the quality of self-reflections in a more nuanced way. Time investment was, however, considered heavy, especially for the first comparisons. Although raters appreciated that they did not have to assign a grade to each self-reflection, the fact that the method does not automatically lead to a grade or feedback was considered a downside. *Conclusions.* First evidence was provided for the comparative judgment method as an alternative to using rubrics for assessing students' written self-reflections. Before comparative judgment can be implemented for summative assessment, more research is needed on the time investment required to ensure no contradictory feedback is given back to students. Moreover, as the comparative judgment method requires an additional standard setting exercise to obtain grades, more research is warranted on the merits and limitations of this method when a pass/fail approach is used.

**Keywords:** comparative judgment; rubrics; reflective skills; reliability

## Introduction

Medical practitioners' reflective skills are crucial for self-regulated and lifelong learning.[1-6] They improve critical thinking, problem solving, and diagnostic skills and enhance communicative and professional behaviour.[7-13] Through positive impact on knowledge, skills, and attitudes, improved decision making in clinical practice, and better therapeutic relationships with patients, doctors' reflective skills also improve patient care.[8,14,15]

Consequently, training reflective skills during medical education is on the rise.[6,7,12,16-19] Such training hinges upon a definition of reflective skills. Though the literature offers a plethora of definitions of reflection,[17,20-22] a number of aspects are key to these definitions: active, metacognitive process, consideration of the consequences, understanding of the 'self' and consideration of future perspectives.[8,15,23,24]

To appraise reflective skills, performance assessment[25] is frequently used; students are asked to write down their reflections on their actions and/or experiences shortly after they occurred.[3,18,19,26-30] The debate about whether these written products should be assessed or not is ongoing.[18,22,31] Those against highlight the fact that summative assessment may distort the intended effects of self-reflective writing,[8,32-35] while those in favor argue that assessment directs learning and that assessment can focus attention on the evaluation of reflective skill rather than content of the writing.[3,26,36,37] This debate notwithstanding, in numerous medical education curricula, self-reflections are awarded a grade.[15,19,38,39]

Appraising Students' Written Self-reflections: Validity and Reliability

Appraising students' written self-reflections is a complex task for raters.[10] To help raters, different methods can be used.[22,40-42] Each method aims to safeguard both the validity and the reliability of the appraisal, both being key aspects of educational measurement.[43]

*Analytical rating using rubrics.* The most frequently used method for appraising students' written self-reflections is analytical rating using rubrics,[6,8,10,11,38,44] which include 'criteria for rating important dimensions of performance, as well as standards of attainment for those criteria.'[45] To achieve validity, these criteria should represent the competence under study. Moreover, to enhance inter-rater reliability, rubrics aim to ensure that all raters take into account the same, predefined aspects of performance.[8,45,46] Table 1 presents three commonly used rubrics: Reflection-on-Action,[47] REFLECT,[8,11,18] and a rubric based on the ALACT-model.[24,44]

[Insert Table 1 around here]

Although the validity of the analytical rating of self-reflections using rubrics has been studied,[8,10] inter-rater reliability has received more attention due to mixed findings (see Table 1). Michels[38] and Ottenberg et al[11] reported adequate reliabilities for the rubric based on the

ALACT-model and a modified version of the REFLECT rubric, respectively, whereas other studies could not confirm this for the REFLECT rubric.[8,10,18]

Difficulty ensuring adequate reliability when using rubrics to appraise self-reflections has been acknowledged in literature.[22,30] Similar difficulties have been encountered in assessments outside medical education that rely on subjective judgment (e.g., writing, mathematical problem-solving).[40,49-51] Despite investment in training, analytical rating using rubrics did not guarantee adequate inter-rater reliabilities for performance assessment.[50,52,53]

*Comparative judgment.* The comparative judgment method is a recently introduced appraisal method that also aims to safeguard both validity and reliability.[40,54] Comparative judgment is based on the assumption that people are more reliable in comparing than in assigning scores to single performances.[55] Various raters independently compare several pairs of students' performances on their overall quality and decide each time which of them is the best with respect to the competence under assessment. Every performance is compared several times and seen by different raters. Based on these judgments, the performances can be ranked from the worst to the best on an interval scale. Because rank order is based on the decisions of several raters, it represents the shared consensus of what comprises a good performance,[40,56,57] thereby safeguarding validity.[57]

It is worth noting that the comparative judgment method does not equate to norm-referenced testing (i.e., "grading on the curve," scores relative to those of other students, such as 10% passes[58]). To the best of our knowledge, in educational settings, comparative judgment has been used solely for criterion-referenced testing, which also is commonplace in competency-based medical education.[59,60] To ensure criterion-referenced testing, an additional step is needed after obtaining rank order. Different options exist, such as a standard setting exercise to determine a pass/fail boundary[61] or expert grading of two performances on the scale, after which the grades for the other performances can be calculated (for example, see [62]).

Over the last 20 years, comparative judgment has been applied to assess a wide range of competencies.[63] Results on the validity of the method have been encouraging; raters took relevant aspects of performance into account while judging,[57,64,65] and rank order correlated strongly with scores from analytical rating.[66,67] Moreover, inter-rater reliability[68] estimates ranged from .80 for a chemistry task[64] to .93 for a mathematics task.[56] The comparative judgment method typically is used with raters who understand the competence description, but lack prior training. This is in line with Pollitt's hypothesis that no training is needed to obtain high reliabilities with the comparative judgment method,[54] which has been confirmed in multiple studies.[64,69]

Yet, the comparative judgment method requires multiple raters and multiple comparisons for each written product.[61] Consequently, the feasibility of the method for appraising performance may be questioned. Remarkably, however, raters' views on the comparative judgment method has received little research attention to date.

This article explores the use of comparative judgment to assess medical students' self-reflections that had previously been assessed using a rubric based on the ALACT-model. As findings regarding the inter-rater reliability of rubrics have been mixed, our first objective was to compare the reliability obtained with rubrics and with comparative judgment in relation to time spent by the rater team on assessing. The second objective was to verify whether the obtained rank order related to scores obtained using rubrics. In addition, we set out to explore raters' perception of the comparative judgment method (third objective).

## Materials and Methods

### Ethical Consent

We obtained ethical approval from the ethics review committee of the Antwerp University Hospital (file number EC UZA 17/42/465).

## Materials

As part of their portfolio assignment in the sixth year of medical school at a mid-sized Belgian University,[70] 119 students (22 or 23 years old) wrote two self-reflections related to an experience during their fulltime clerkship period. At the onset of the academic year, students received instructions for writing these self-reflections, including details on why reflective skills are important, questions that students can ask themselves to help the reflective process, and the 10 topics they could choose from (critical incident, attitude, ethics, choice of profession, the doctor's world, prevention, dealing with children and their parents, aging, psychiatry, and, primary, secondary and tertiary health care). Moreover, it was explicitly stated that the reflective process would be evaluated and not the described feelings or actions (and possible errors) that were reflected upon. During the academic year, students could opt to meet with their coach, who was a member of the portfolio team (which included the third and last authors). During such meetings, questions on the tasks including the self-reflection exercise could be discussed, and the coach provided feedback on the written self-reflections.

Students' self-reflections had previously been scored using a rubric based on the ALACT-model (see Table 1, see Appendix 1).[38,44] The course administrator estimated the average time for scoring one self-reflection to be around 12 minutes, not including the time to provide written feedback on the reflection's strengths or suggestions for improvement.

A double-rating procedure using rubrics was applied, involving two raters with complementary expertise. The first rater was one of two assessors on the communication skills team, which was responsible for teaching, among other topics, on reflective skills. These two teachers scored the self-reflections without having insight into the other portfolio tasks. The second rater was a member of the portfolio team, but not the coach. The portfolio team consisted of medical educators with broad expertise. This rater evaluated all tasks in the portfolio of a single student (including self-reflections but also the presentation of a patient case or a personal development plan, for example) as well as the portfolio's overall quality. All raters had participated in numerous, in-depth team discussions on the rubric and on the quality of self-reflections, which constituted a form of training. For full detail on the portfolio, its reliability, the double-rating procedure using rubrics, and content validity, see Michels and colleagues.[44,70]

For this study, we sampled 22 self-reflections purposefully selected to represent the full-sample variation in obtained grades, first rater, and frequency of reflection topic. For example, 'critical incident' was among the most popular reflection topics in the full sample (24.5%), and thus five out of the 22 self-reflections chosen were on this topic (22.7%). A sample of 22 self-reflections enabled us to reach 18 comparisons per self-reflection, which was needed to examine how reliability evolved with time invested by the rater team while keeping rater workload feasible. Self-reflections were anonymized prior to the comparative judgment exercise.

## Participants

Eight raters volunteered to participate in the comparative judgment of the 22 self-reflections: four of them were part of the communication team, while the other four were part of the portfolio team. All raters had at least four years of experience in rating self-reflections (see Table 2), using the rubric (see Appendix 1). Consequently, in line with the requirements of the comparative judgment, they were familiar with the competence under study.

[Insert Table 2 around here]

Procedure

The raters received an email with instructions on how to log onto D-PAC, a web-based tool that supports assessment with comparative judgment (https://comproved.com/en/).[61] Once logged in, raters could watch a short instructional video and consult the guidelines of the assignment that students had received. They also received the competence description, detailing the same dimensions as the rubric based on the ALACT model (see Appendix 1). Subsequently, each rater was presented the first pair of self-reflections of the 25 comparisons they were asked to complete. Keeping the competence description in mind, the assessors answered the question ‘Which is the better self-reflection?’ After each comparison, raters were asked to give feedback on both self-reflections by writing down strengths and recommendations for improvement (not analyzed in this study). In total, the eight raters completed 200 comparisons.

The algorithm to select the self-reflections for comparative judgment had two checks. First, to select self-reflection A, the algorithm identified the self-reflections that had been compared least up to that point in the assessment and randomly drew a self-reflection from this group. This ensured that each self-reflection was selected an equal number of times. Second, to select self-reflection B, the algorithm randomly selected among those to which self-reflection A had not yet been compared, thus avoiding duplicate pairs.

The 200 comparisons were analyzed using the Bradley-Terry-Luce model,[68] which generated a rank order of the self-reflections from the weakest to the strongest self-reflection (see Figure 1). For each self-reflection, the analysis also provided a logit score for its quality, which ranged from -2.9 to 2.28. This logit score indicates the chance (more precisely, the logistic transformation of the chance) that a self-reflection will win a comparison with a self-reflection of average quality (having a logit score of 0).

When each of the raters had finished their comparisons, a focus group was organized, which was facilitated by the second and fourth authors. Six raters took part in this focus group (see Table 2). The following topics were covered: raters' experiences with the comparative judgment method, their experience with the D-PAC tool, their perception of the resulting rank order and the additional requirements needed to allow implementation of comparative judgment in their context. Given that a limited set of self-reflections were used for the comparative judgment, the resulting rank order, strengths and recommendations for improvement were not fed back to the students.

## Data Analysis

To pursue the first research objective, we calculated the one-way random ICC for single raters (ICC (1,1)) and the average measure ICC for two raters (ICC, (1,k)) using rubrics.[71] For the comparative judgment data, the scale separation reliability (SSR)[63,72] was assessed, using R. As evidenced by Verhavert and colleagues,[68] the SSR should be viewed as an interrater reliability measure: a high SSR indicates that raters agree on the relative position of the self-reflections in the rank order. The SSR was calculated each time all self-reflections had been compared one additional time (i.e., after each round). For example, the third round ends when each self-reflection has been compared three times by the set of raters.

To relate the evolution in the SSR to the time spent by the rater team, the time data per comparison -- tracked by the D-PAC tool -- were used. One comparison proved to be an outlier and was replaced by the mean time for the other comparisons, being 5 minutes 30 seconds (i.e., on average, raters needed 5 min 30 sec to read the two self-reflections and to judge which one was better, excluding the time to write feedback on both self-reflections). Subsequently, the total cumulative time per round was calculated. For example, for the seventh round, the sum was taken of all time estimates up to the moment each self-reflection had been compared seven times.

Given that both methods (rubrics and comparative judgment) set out to measure similar qualities, it was relevant to examine whether the rank orders correlate. To pursue the second research objective, Kendall's tau rank correlation (or Kendall's τ) was calculated between the scores obtained using rubrics and those obtained using comparative judgment.

To explore raters' perceptions of the comparative judgment method, we analyzed the focus group transcriptions in an iterative procedure inspired by Braun and Clark's phases of thematic analysis.[73] To enhance the quality of the coding procedure, both the first and second author were involved in the coding process. In the first phase, the two researchers individually familiarized themselves with the focus group content, performed the first coding, and decided which main aspects were important. Next, these researchers discussed the main themes regarding raters' experiences with comparative judgment. Based on the results of this discussion, the first researcher coded the transcripts anew with the aim of refining the coding and the main themes. Once again, these adjustments were discussed with the second researcher. In a last phase, the main and sub-themes were discussed with the last author. As she also participated in the focus group, this phase supported respondent validation.

## Results

### Reliability in Relation to Time Spent

With rubrics, the one-way random ICC for single raters (ICC (1,1)) and the average measure ICC for two raters (ICC, (1,k)) were .39 and .56 respectively. As indicated previously, the estimated time for scoring one self-reflection was 12 minutes. Double-rating thus required an estimated time investment of the rater team of 24 minutes per self-reflection.

Regarding the evolution in reliability for comparative judgment, a SSR of .55 was obtained when all self-reflections had appeared seven times in a pair (i.e., 7 rounds, see Figure 2). Afterwards, the SSR continues to increase rapidly up to .70, which is obtained after 10 rounds. From the 11th to the 18th comparison per self-reflection, the increase in reliability level was less pronounced: after the completion of 200 comparisons, the SSR reached a reliability of .77.

Table 3 provides an overview of reliability and time investment by the rater team. To reach the reliability level of .56 that was obtained using double rating with rubrics, 7 rounds of comparative judgment were needed. By that time, each self-reflection had been seen by 3 or 4 raters.

[Insert Table 3 around here]

The time investment per self-reflection was similar for both methods. Using double rating with rubrics, 24 minutes were necessary, while this amounted to 22 minutes using comparative judgment. To reach a reliability of .70 using comparative judgment, a time investment of 28 minutes per self-reflection was needed.

## Correlation between Rank Orders

Results indicated a strong correlation between the rank orders obtained from rubrics and from comparative judgment (Kendall's $\tau = 0.74$, $p < 0.01$, see Figure 3), suggesting that both methods appeared to rank the reflections similarly. As one of the self-reflections scored extremely low on both scoring methods (see Figure 3), we also examined the rank correlation without this self-reflection. Removing this self-reflection from the analysis, the correlation for the other 21 self-reflections was still strong (Kendall's $\tau = 0.71$, $p < 0.01$).

## Raters' Perceptions of Comparative Judgment: Themes from the Focus Group

The raters described making comparative judgments as a fruitful exercise. They indicated that by comparing, they focused more on the essential aspects of a self-reflection. Moreover, seeing a self-reflection multiple times and in comparison with different self-reflections aided in judging its quality in a more nuanced way.

> The first time you read a text, you read it as you would normally do [for rubrics rating]. But, given that you see the same text a few times, you see it in a different light. (…) The next time I saw it, I started comparing; what is better in this reflection compared to the other one? That you don't do if you would just read it [for rubric rating]. (Rater 2)

This more nuanced view was especially true if two reflections were of a similar quality. Furthermore, the raters indicated that this more nuanced opinion helped in writing detailed feedback.

> (…) indeed, you give feedback on one self-reflection. But in my view, through the comparison, you can think of better feedback. You wonder why you find the one [self-reflection] better than another. Those elements you take along to the next comparison as well. (Rater 8)

However, the time investment needed to compare self-reflections on their overall quality was perceived as burdensome, especially for the first comparisons. Once self-reflections reappeared, and especially if one could rely on notes made previously for that particular self-reflection, less time was required to make a comparison. Additionally, raters mentioned that not having to assign a grade speeded up the rating process.

Yet, the raters underlined problems related to the collective result of the comparative judgment exercise (i.e., all comparisons together produced a single rank order of quality). Two problems were highlighted regarding this collective result. First, although it was perceived as reassuring that multiple raters contributed to the final rank order, the fact that a choice was needed (i.e., "Which is the better self-reflection?"), was perceived as potentially problematic. For two self-reflections of a similar quality, the raters questioned whether being forced to make a choice could have negatively affected the self-reflection that was not chosen. Moreover, the raters wondered whether the final rank order was impacted by the fact that some raters encountered a specific self-reflection more often than other raters in the process of comparative judgment.

Second, the final rank order did not automatically lead to a grade. This was perceived as a benefit by some raters.

> I would find it really nice if we could drop that [the grades]… From a pedagogical perspective, it would be A LOT better for our students as well. But, it is not up to us to decide on this. (Rater 3)

The comparative judgment method was seen as an opportunity to opt for a pass/fail approach. Yet, as indicated during the focus group, the current educational system requires raters to give a grade and to be as sure as possible about this grade, given its influence on students' future chances of choosing a specialty. It was explained to the raters that, to obtain such grades or a pass/fail cut-off, a standard setting exercise would be needed. The raters considered this additional time investment a clear downside of the comparative judgment method.

Third, after examining the compilation of the feedback provided per self-reflection, the raters concluded that, for some self-reflections, feedback was contradictory. The raters did not see this as troublesome or odd, given the multiple raters involved. The contradictions were even viewed as an opportunity for the rater team to further align raters' visions on the quality of a self-reflection. However, raters largely agreed that students would be confused by the contradictory feedback or would use it to plead for a higher grade. Hence, it would require another investment to make a summary of the feedbacks prior to sending them to the students.

## Discussion

This study explored the comparative judgment method as an alternative to rubric-based rating in medical education for assessing medical students' self-reflections. Regarding reliability in relation to the time spent assessing (first research objective), results indicated that the comparative judgment method required a similar time investment to reach the reliability level obtained using double rating with rubrics. This finding is in line with previous studies in the language domain.[54,66]

In line with previous research on performance assessment[40,49,50] and on self-reflections more specifically,[8,10] the inter-rater reliability for double rating of the self-reflections using rubrics was rather low, although the reliability level for an entire portfolio, including self-reflections, was adequate.[44] As the raters involved had been trained, this finding corroborates research describing that training does not suffice to reach adequate inter-rater reliabilities for performance assessments.[50,52,53] To shed more light on the relative efficiency of rubrics and comparative judgment to reach the recommended higher reliability levels (e.g., .70 or .80[45]), future research should consider including more than two rubric ratings per self-reflection in order to reach adequate reliability levels.

Regarding time investment, the time spent assessing via comparative judgment was tracked in the D-PAC tool, while the course administrator estimated the time spent assessing via rubric. To gain more insight into the time investment needed to reach a certain reliability, future research should track the time investment for both methods.

Moreover, the assessment of student work comprises other, time-consuming activities besides rating. For analytical rating, developing a rubric and training (new) raters to use this rubric is considered indispensable.[8,39,45,46] Using comparative judgment, to assign pass/fail or a grade to a given rank order, the team members need to do one (in the case of pass/fail) or multiple (in the case of grades) standard setting exercises.[61,64] A more comprehensive view of the activities per rating method requiring additional time investment is needed. Furthermore, it may be worthwhile for future research to examine the time investment from both the perspective of a single rater and the team.

Regarding the second research objective, results showed a strong correlation between the rank order obtained using rubrics and the rank order obtained using comparative judgment. This indicates that both methods likely measure similar qualities in self-reflections. This finding is in line with previous research in the language domain, where Pearson correlation coefficients of .97[67] and .85[66] were found. To estimate the relation between the rank orders stemming from both methods more accurately in future research, higher reliability levels for rubrics rating would be desirable.

Note that all raters had experience with rating self-reflections using the rubrics based on the ALACT-model and had participated in frequent discussions in the team on the rubric and on the quality of self-reflections, which constitutes a form of training. Further research should consider including two groups of raters, one with and one without previous experience in rating self-reflections. As such, it can be assessed whether raters receiving just the competence description can also rate self-reflections in a reliable way.

The focus group on raters' perceptions of the comparative judgment method (third research objective) revealed three important themes. First, raters reported that the comparative judgment method helped them to focus more on the essential aspects of a self-reflection and to judge the quality of self-reflections in a more nuanced way. This may be of interest in light of the debate on whether writing ability or storytelling impacts students' grades on written self-reflections.[22,74,75] For rubrics, primary research evidence suggests a limited impact.[76] To date, similar evidence is lacking for the more novel method of comparative judgment. There is, however, some indirect evidence: research on the validity of comparative judgment to assess essays found that raters provided few construct-irrelevant arguments to underpin their choice.[57,65] To verify raters' impressions that the comparative judgment method helps in focusing on the essential aspects of self-reflection, future research could replicate the study by Aronson and colleagues[76] and include a comparative judgment condition. Here too, it would be worthwhile to include raters with and without previous experience in rating self-reflections using rubrics; this would allow examination of whether raters who must rely on solely the competence description are more prone to assessing writing ability or storytelling instead of reflective skills.

Second, raters perceived the time investment required for comparative judgment as heavy. This may be because, for research purposes, each self-reflection was compared eighteen times (i.e., 200 comparisons). Yet, with seven rounds (i.e., 77 comparisons), a reliability level comparable to the one using rubrics was obtained. To adequately capture raters' views on the time investment, future research should consider stopping the comparative judgment exercise once a desired reliability level is reached. In addition, raters raised the fact that, for criterion-referenced testing, a standard setting exercise would be needed. This additional step was described as a clear downside of the comparative judgment method. In line with this, raters indicated that the comparative judgment method is more apt when a pass/fail approach is used,[77] which has been recommended for reflective essays,[3] as it requires only one standard.

It should be noted that recent advancements in the comparative judgment field have described alternatives for the standard setting exercise. A first option consists in grading two products of the rank order (for example by the team leader). Using this information, the grades for the other self-reflections are calculated using the logit scores from the rank order.[62] Another option, which has not been tested to our knowledge, would be to ask raters to indicate a provisional grade boundary (e.g., clearly pass/clearly fail/unsure) while they are assessing. This could inform grading afterwards. As criterion-referenced testing is recommended in competency-based medical education, future research on the feasibility of these options for the assessment of self-reflections is clearly warranted.

Third, the feedback component in a comparative judgment exercise remains a challenge. Raters reported that using comparative judgment prompted them to provide more detailed feedback. In future research, it would be worthwhile to analyze the feedback in both methods (rubrics and comparative judgment) in order to validate this viewpoint. Moreover, in line with Mortier and colleagues,[78] further research should examine the student perspective: Do they also view feedback from comparative judgment as more detailed than feedback from rubric-based rating? And, which feedback affects student learning most?

Raters perceived the multisource feedback in comparative judgment to be rich, which corroborates previous research.[60,79,80] Yet, as previously described,[48,80,81] it is likely to contain contradictions stemming from different interpretations or nuances. In line with previous research,[48,80] raters indicated that such contradictory feedback could confuse students. It remains to be examined how students view and act upon such contradictory feedback:[48] under which conditions (e.g., high stakes assessment, formative assessment) does it confuse them versus trigger learning?

Along with the directions for future research described above, four limitations of the present study need to be acknowledged. First, the eight raters volunteered to participate in this study. Future research in which raters are randomly assigned to the rubric or comparative judgment method is warranted. Second, as the present study comprised a first application of comparative judgment in medical education context, it combined a small group of raters, a limited set of written self-reflections, and a high number of rounds (i.e., 18). Subsequently, the self-reflections reappeared frequently in the set of 25 comparisons each rater was requested to make, which, according to raters, speeded up the comparisons. If this is the case, it would imply that the time investment noted in this study cannot be extrapolated to settings in which large groups of raters assess a large set of written self-reflections and fewer rounds are used. In such settings, self-reflections would re-appear less frequently and there would be more judgment time in between. Consequently, raters' views on the time investment in such settings merits further investigation. Third, as previous research indicates,[10] multiple self-reflections per student are needed to reach an adequate reliability level. It needs to be examined whether it is more reliable and time efficient to rank each self-reflection separately or whether student portfolios containing multiple self-reflections (and if so, how many) can be adequately compared by raters. Fourth, although the SSR was found to be an inter-rater reliability measure,[68] allowing for a comparison with the ICC, further statistical research preferably discerns a single measure that can be calculated in both rubrics and comparative judgment settings.

For practitioners interested in using comparative judgment, we have four recommendations, next to those formulated elsewhere.[61] First, it is recommended to do a short dry run in order to estimate the time required for a comparison. In practice, this implies for example two raters completing two comparisons each and calculating the median time per comparison. Second, the desired reliability needs to be determined. A recent meta-analysis[82] specifies that to reach an SSR of .70, 10 to 14 rounds likely would be needed. Third, it is useful to monitor reliability during assessment, as it may reach an asymptote. If so, further time investment does not provide an additional gain in reliability.[82] Fourth, to avoid contradictory feedback, it may be worthwhile to separate the judging (deciding which self-reflection is better) from the provision of feedback. Possibly, after the rank order has been established, the self-reflections could be divided among the raters to provide feedback on them.

## Conclusion

We have examined the merits and limitations of comparative judgment as an alternative to rubrics in medical education. Specifically, we explored its reliability in relation to the time spent assessing, its association to rubric ratings, and raters' perceptions of its utility and feasibility for assessing students' written self-reflections. Our results suggest that a similar time investment was required to reach the same reliability as that obtained using double rating with rubrics. Moreover, comparative judgment and rating using rubrics valued the same qualities in self-reflections. Raters emphasized the strength of the method to provide detailed and nuanced feedback. With regard to limitations, raters underlined the extra investment needed to obtain grades and the possible contradictory feedback to students. These findings emphasize the importance of future interventions using comparative judgment, particularly when raters encounter difficulties in assessing performance assessment reliably.

**Declaration of interest:** The authors report no declarations of interest.

**ORCID**

Liesje Coertjens http://orcid.org/0000-0003-1209-5089
Benedicte Y. De Winter http://orcid.org/0000-0003-0327-6304
Sven De Maeyer http://orcid.org/0000-0003-2888-1631
Nele Michels http://orcid.org/0000-0003-1971-0793

## REFERENCES

1. Mann K, Gordon J, MacLeod A. Reflection and reflective practice in health professions education: a systematic review. *Advances in Health Sciences Education.* 2007;14(4):595.
2. Ramani S. Reflections on feedback: Closing the loop. *Medical Teacher.* 2016;38(2):206-207.
3. Maloney S, Tai JH-M, Lo K, Molloy E, Ilic D. Honesty in critically reflective essays: an analysis of student practice. *Advances in Health Sciences Education.* 2013;18(4):617-626.
4. Mak-van der Vossen MC, de la Croix A, Teherani A, van Mook WNKA, Croiset G, Kusurkar RA. Developing a two-dimensional model of unprofessional behaviour profiles in medical students. *Advances in Health Sciences Education.* 2018.

5. Ambrose LJ, Ker JS. Levels of reflective thinking and patient safety: an investigation of the mechanisms that impact on student learning in a single cohort over a 5 year curriculum. *Advances in Health Sciences Education.* 2014;19(3):297-310.
6. Alizadeh M, Mirzazadeh A, Parmelee DX, et al. Leadership Identity Development Through Reflection and Feedback in Team-Based Learning Medical Student Teams. *Teaching and Learning in Medicine.* 2018;30(1):76-83.
7. Driessen E, Tartwijk Jv, Dornan T. The self critical doctor: helping students become more reflective. *BMJ.* 2008;336(7648):827-830.
8. Wald HS, Borkan JM, Taylor JS, Anthony D, Reis SP. Fostering and Evaluating Reflective Capacity in Medical Education: Developing the REFLECT Rubric for Assessing Reflective Writing. *Academic Medicine.* 2012;87(1):41-50.
9. Kihlgren P, Spanager L, Dieckmann P. Investigating novice doctors' reflections in debriefings after simulation scenarios. *Medical Teacher.* 2015;37(5):437-443.
10. Moniz T, Arntfield S, Miller K, Lingard L, Watling C, Regehr G. Considerations in the use of reflective writing for student assessment: issues of reliability and validity. *Medical Education.* 2015;49(9):901-908.
11. Ottenberg AL, Pasalic D, Bui GT, Pawlina W. An analysis of reflective writing early in the medical curriculum: The relationship between reflective capacity and academic achievement. *Medical Teacher.* 2016;38(7):724-729.
12. Binyamin G. Growing from dilemmas: developing a professional identity through collaborative reflections on relational dilemmas. *Advances in Health Sciences Education.* 2018;23(1):43-60.
13. Humphrey-Murto S, Mihok M, Pugh D, Touchie C, Halman S, Wood TJ. Feedback in the OSCE: What Do Residents Remember? - *Teaching and Learning in Medicine.* 2016;28(1):52-60.
14. Burnett E, Phillips G, Ker JS. From theory to practice in learning about healthcare associated infections: Reliable assessment of final year medical students' ability to reflect. *Medical Teacher.* 2008;30(6):e157-e160.
15. Sandars J. The use of reflection in medical education: AMEE Guide No. 44. *Medical Teacher.* 2009;31(8):685-695.
16. Brand S, Lancaster P, Gafson I, Nolan H. Encouraging reflection: good doctor or bad doctor? *Medical Education.* 2017;51(11):1173-1174.
17. Holmes CL, Harris IB, Schwartz AJ, Regehr G. Harnessing the hidden curriculum: a four-step approach to developing and reinforcing reflective competencies in medical clinical clerkship. *Advances in Health Sciences Education.* 2015;20(5):1355-1370.
18. Hayton A, Kang I, Wong R, Loo LK. Teaching Medical Students to Reflect More Deeply. *Teaching and Learning in Medicine.* 2015;27(4):410-416.
19. Chretien KC, Chheda SG, Torre D, Papp KK. Reflective Writing in the Internal Medicine Clerkship: A National Survey of Clerkship Directors in Internal Medicine. *Teaching and Learning in Medicine.* 2012;24(1):42-48.
20. Lew MDN, Schmidt HG. Self-reflection and academic performance: is there a relationship? *Advances in Health Sciences Education.* 2011;16(4):529.
21. Poole G, Jones L, Whitfield M. Helping students reflect: lessons from cognitive psychology. *Advances in Health Sciences Education.* 2013;18(4):817-824.
22. Koole S, Dornan T, Aper L, et al. Factors confounding the assessment of reflection: a critical review. *BMC Medical Education.* 2011;11(1):104.
23. Dewey J. *How we think.* New York: BN Publishing; 1909.
24. Korthagen FAJ. Reflective Teaching and Preservice Teacher Education in the Netherlands. *Journal of Teacher Education.* 1985;36(5):11-15.
25. Kane M, Crooks T, Cohen A. Validating Measures of Performance. *Educational Measurement: Issues and Practice.* 1999;18(2):5-17.
26. Aronson L. Twelve tips for teaching reflection at all levels of medical education. *Medical Teacher.* 2011;33(3):200-205.
27. McGuire L, Lay K, Peters J. Pedagogy of Reflective Writing in Professional Education. *Journal of the Scholarship of Teaching and Learning.* 2012;9(1):93-107.

28. Levine RE, Kelly PA, Karakoc T, Haidet P. Peer Evaluation in a Clinical Clerkship: Students’ Attitudes, Experiences, and Correlations With Traditional Assessments. *Academic Psychiatry.* 2007;31(1):19-24.
29. Nothelle SK, Christmas C, Hanyok LA. First-Year Internal Medicine Residents' Reflections on Nonmedical Home Visits to High-Risk Patients. *Teaching and Learning in Medicine.* 2018;30(1):95-102.
30. Uygur J, Stuart E, De Paor M, et al. A Best Evidence in Medical Education systematic review to determine the most effective teaching methods that develop reflection in medical students: BEME Guide No. 51. *Medical Teacher.* 2019;41(1):3-16.
31. Veen M, Croix A. The swamplands of reflection: using conversation analysis to reveal the architecture of group reflection sessions. *Medical Education.* 2017;51(3):324-336.
32. Reis SP, Wald HS, Monroe AD, Borkan JM. Begin the BEGAN (The Brown Educational Guide to the Analysis of Narrative) - A framework for enhancing educational impact of faculty feedback to students' reflective writing. *Patient Education and Counseling.* 2010;80(2):253-259.
33. McNeill H, Brown JM, Shaw NJ. First year specialist trainees’ engagement with reflective practice in the e-portfolio. *Advances in Health Sciences Education.* 2010;15(4):547-558.
34. Driessen E. Do portfolios have a future? *Advances in Health Sciences Education.* 2017;22(1):221-228.
35. Branzetti J, Gisondi MA, Hopson LR, Regan L. Aiming Beyond Competent: The Application of the Taxonomy of Significant Learning to Medical Education. *Teaching and Learning in Medicine.* 2019;31(4):466-478.
36. Plack MM, Driscoll M, Marquez M, Cuppernull L, Maring J, Greenberg L. Assessing Reflective Writing on a Pediatric Clerkship by Using a Modified Bloom’s Taxonomy. *Ambulatory Pediatrics.* 2007;7(4):285-291.
37. Kember D, Leung D, McNaught C. A workshop activity to demonstrate that approaches to learning are influenced by the teaching and learning environment. *Active learning in higher education.* 2008;9(1):43-56.
38. Michels NR. *Portfolio learning and assessing at the workplace : development, reliability and validity. Doctoral thesis.* Antwerp2012.
39. Miller-Kuhlmann R, O’Sullivan PS, Aronson L. Essential steps in developing best practices to assess reflective skill: A comparison of two rubrics. *Medical Teacher.* 2016;38(1):75-81.
40. Pollitt A. Comparative judgement for assessment. *International Journal of Technology and Design Education.* 2012;22(2):157-170.
41. Weigle SC. *Assessing writing.* Cambridge: Cambridge University Press 2002.
42. Bacha N. Writing evaluation: what can analytic versus holistic essay scoring tell us? *System.* 2001;29(3):371-383.
43. Brennan RL, ed *Educational measurement.* Westport, CT: American Council on Education/Praeger; 2004; No. 4.
44. Michels NR, Driessen E, Muijtjens A, Van Gaal L, Bossaert L, De Winter B. Portfolio Assessment during Medical Internships: How to Obtain a Reliable and Feasible Assessment Procedure? *Education for Health.* 2009;22(3):313-313.
45. Jonsson A, Svingby G. The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review.* 2007;2(2):130-144.
46. Stemler SE. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation.* 2004;9(4):1-11.
47. O'Sullivan P, Aronson L, Chittenden E, Niehaus B, Learman L. Reflective ability rubric and user guide. *MedEdPORTAL.* 2010;6:8133.
48. Ginsburg S, Vleuten CP, Eva KW, Lingard L. Cracking the code: residents’ interpretations of written assessment comments. *Medical Education.* 2017;51(4):401-410.

49. Bloxham S, den-Outer B, Hudson J, Price M. Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education.* 2016;41(3):466-481.

50. Bartholomew SR, Yoshikawa-Ruesch E. A Systematic Review of Research Around Adaptive Comparative Judgement (ACJ) in K-16 Education. In: Wells J, ed. *CTETE—Research Monograph Series: Council on Technology and Engineering Teacher Education*. Virginia: Council on Technology and Engineering Teacher Education. 2018;1(1):6–28.

51. Buckley J, Canty D, Seery N. An exploration into the criteria used in assessing design activities with adaptive comparative judgment in technology education. *Irish Educational Studies.* 2020:1-19. doi:10.1080/03323315.2020.1814838.

52. Rezaei AR, Lovorn M. Reliability and validity of rubrics for assessment through writing. *Assessing Writing.* 2010;15(1):18-39.

53. Weigle SC. Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing.* 1999;6(2):145-178.

54. Pollitt A. The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice.* 2012;19(3):281-300.

55. Thurstone LL. A law of comparative judgment. *Psychological review.* 1927;34(4):273.

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

57. van Daal T, Lesterhuis M, Coertjens L, Donche V, De Maeyer S. 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.* 2019;26(1):59-74.

58. Haertel EH. Reliability. In: Brennan RL, ed. *Educational measurement.* Vol 4th. Westport, CT: American Council on Education/Praeger; 2006.

59. Harris P, Bhanji F, Topps M, et al. Evolving concepts of assessment in a competency-based world. *Medical Teacher.* 2017;39(6):603-608.

60. Lockyer J, Carraccio C, Chan M-K, et al. Core principles of assessment in competency-based medical education. *Medical Teacher.* 2017;39(6):609-616.

61. Lesterhuis M, Verhavert S, Coertjens L, Donche V, De Maeyer S. Comparative judgement as a promising alternative to score competences. In: Ion G, Cano E, eds. *Innovative Practices for Higher Education Assessment and Measurement*. Hershey, PA: IGI Global; 2016.

62. Settembri P, Van Gasse R, Coertjens L, De Maeyer S. Oranges and Apples? Using Comparative Judgement for Reliable Briefing Paper Assessment in Simulation Games. In: Bursens P, Donche V, Gijbels D, Spooren P, eds. *Simulations of Decision-Making as Active Learning Tools: Design and Effects of Political Science Simulations*. Cham: Springer International Publishing; 2018:93-108.

63. Bramley T. *Investigating the reliability of Adaptive Comparative Judgment.* Cambridge: Cambridge Assessment;2015.

64. McMahon S, Jones I. A comparative judgement approach to teacher assessment. *Assessment in Education: Principles, Policy & Practice.* 2015;22(3):368-389.

65. Whitehouse C. *Testing the validity of judgements about geography essays using the Adaptive Comparative Judgement method.* Manchester: AQA Centre for Education Research and Policy; 2012.

66. Coertjens L, Lesterhuis M, Verhavert S, Van Gasse R, De Maeyer S. Judging texts with rubrics and comparative judgement: Taking into account reliability and time investment. [Teksten beoordelen met criterialijsten of via paarsgewijze vergelijking: Een afweging van betrouwbaarheid en tijdsinvestering]. *Pedagogische Studien.* 2017;94(4):283-303.

67. Heldsinger S, Humphry S. Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher.* 2010;37(2):1-19.

68. Verhavert S, De Maeyer S, Donche V, Coertjens L. Scale Separation Reliability: What Does It Mean in the Context of Comparative Judgment? *Applied Psychological Measurement.* 2018;42(6):428-445.
69. Bouwer R, Lesterhuis M, Bonne P, De Maeyer S. Applying Criteria to Examples or Learning by Comparison: Effects on Students' Evaluative Judgment and Performance in Writing. *Frontiers in Education.* 2018;3(86):1–12. doi:10.3389/feduc.2018.00086.
70. Michels NR, Avonts M, Peeraer G, et al. Content validity of workplace-based portfolios: A multi-centre study. *Medical Teacher.* 2016;38(9):936-945.
71. Gwet KL. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters.* 4th edition ed. Gaithersburg, MD: Advanced Analytics, LLC; 2014.
72. Bramley T, Vitello S. The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice.* 2018:1-16.
73. Braun V, Clarke V. Using thematic analysis in psychology. *Qualitative Research in Psychology.* 2006;3(2):77-101.
74. Braun UK, Gill AC, Teal CR, Morrison LJ. The utility of reflective writing after a palliative care experience: can we assess medical students' professionalism? *Journal of palliative medicine.* 2013;16(11):1342-1349.
75. Driessen EW, Overeem K, van Tartwijk J, van der Vleuten CP, Muijtjens AM. Validity of portfolio assessment: which qualities determine ratings? *Med Educ.* 2006;40(9):862-866.
76. Aronson L, Niehaus B, DeVries CD, Siegel JR, O'Sullivan PS. Do writing and storytelling skill influence assessment of reflective ability in medical students' written reflections? *Academic medicine : journal of the Association of American Medical Colleges.* 2010;85(10 Suppl):S29-32.
77. Jacobs JL, Samarasekera DD, Shen L, Rajendran K, Hooi SC. Encouraging an environment to nurture lifelong learning: An Asian experience. *Medical Teacher.* 2014;36(2):164-168.
78. Mortier AV, Lesterhuis M, Vlerick P, De Maeyer S. Comparative Judgment Within Online Assessment: Exploring Students Feedback Reactions. In: Ras E, Joosten-ten Brinke D, eds. *Computer Assisted Assessment. Research into E-Assessment.* Vol 571: Springer, Cham; 2015.
79. Tekian A, Watling CJ, Roberts TE, Steinert Y, Norcini J. Qualitative and quantitative feedback in the context of competency-based education. *Medical Teacher.* 2017;39(12):1245-1249.
80. ten Cate O, Regehr G. The Power of Subjectivity in the Assessment of Medical Trainees. *Academic Medicine.* 2019;94(3):333-337.
81. Gingerich A, Ramlo SE, van der Vleuten CPM, Eva KW, Regehr G. Inter-rater variability as mutual disagreement: identifying raters’ divergent points of view. *Advances in Health Sciences Education.* 2017;22(4):819-838.
82. Verhavert S, Bouwer R, Donche V, De Maeyer S. A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice.* 2019;26(5):541-562.

TABLE 1. Overview of used rubrics to assess students' reflective skills

TABLE 2. Overview of raters' characteristics

TABLE 3. Reliability level in relation to the time investment by the rater team

FIG. 1. The rank order of the self-reflections

FIG. 2. The evolution of the reliability level per round of comparisons

FIG. 3. Relation between the rank orders stemming from rubrics and from comparative judgment

TABLE 1

Overview of used rubrics to assess students' reflective skills

| | Reflection-on-Action | REFLECT | Rubric based on ALACT-model |
|---|---|---|---|
| Characteristics | holistic score on a 6-point scale | analytical rubric, detailing four levels for each dimension | set of 3 dimensions, analytical but with an overall 8-point grade |
| Dimensions | One dimension: reflection performance[47] | 5 dimensions: writing spectrum; presence; description of conflict or disorienting dilemma; attending to emotions; analysis and meaning making.<br>Optional minor criterion: attention to assignment (when relevant)[8] | 3 dimensions: relevance/choice of the topic; the ALACT model (action; looking back; awareness of essential aspects; creating alternative methods; trial); and personal point of view (see Appendix 1) |
| Training required to reach reliability above .80[39] | 2 hours | 6 hours | no research evidence available |

| | | | |
|---|---|---|---|
| Evidence on the inter-rater reliability | no research evidence available | Wald et al[8]: ICC for the single measures from .376 to .748<br>Moniz et al[10]: ICC for the single measures of .457<br>Hayton et al[18]: Kappa from .27 to .38<br>Ottenberg et al[11] (modified version of REFLECT): ICC of .68 | Michels[38]: ICC for the single measures from .60 to .66; ICC for the average measure from .75 to .80 |
| References | [39,47] | [8,10,11,39] | [38,44] |

TABLE 2

Overview of raters' characteristics

| Rater | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Years of experience with analytical rating of written self-reflections | 12 | 12 | 5 | 9 | 9 | 5 | 4 | 10 |
| Number of reflections previously rated (using rubrics) from the sample of 22 | 0 | 0 | 3 | 2 | 11 | 0 | 0 | 0 |
| Participation in focus group | Yes | Yes | Yes | Yes | No | No | Yes | Yes |

TABLE 3

Reliability level in relation to the time investment by the rater team

| | Rubrics | Comparative Judgment 7 rounds | Comparative Judgment 10 rounds |
|---|---|---|---|
| Reliability level | .56 | .55 | .70 |
| Number of different raters | 2 | 3 or 4 | between 4 and 6 |
| Total time spent rating per self-reflection* | 24 min | 22 min | 28 min |
| Total time spent rating the 22 self-reflections* | 8 h 50 min | 8 h | 10 h 15 min |

* The time estimates do not comprise the time for writing feedback

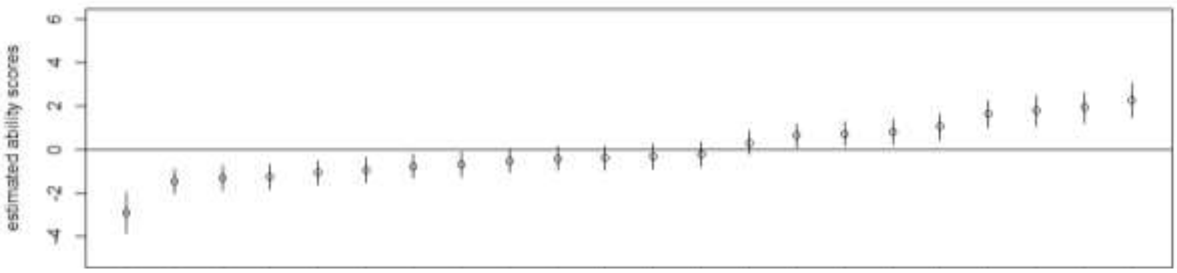

FIG. 1. The rank order of the self-reflections

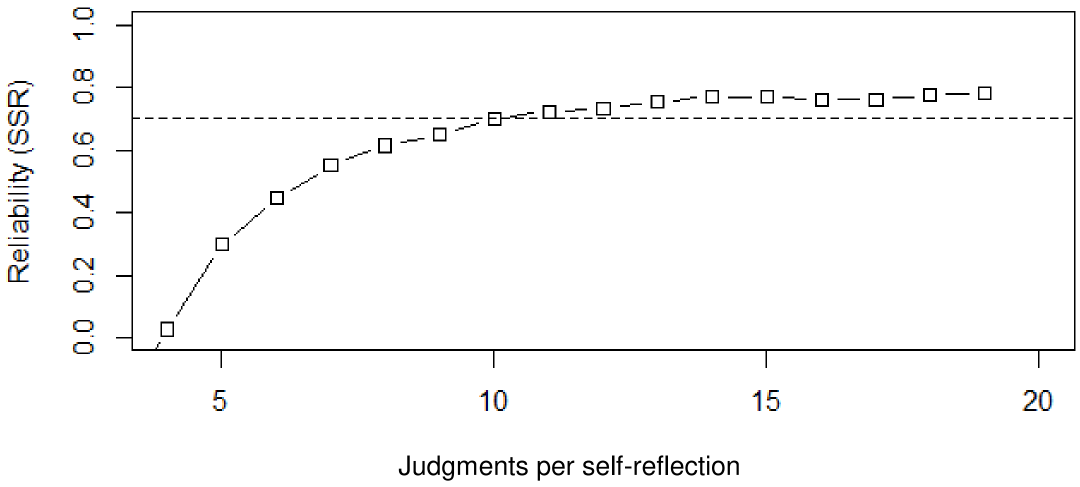

FIG. 2. The evolution of the reliability level per round of comparisons

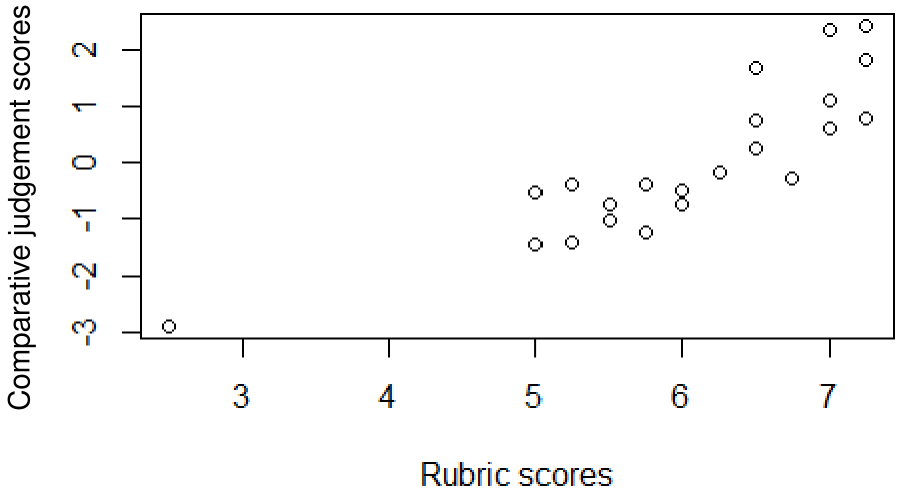

FIG. 3. Relation between the rank orders stemming from rubrics and from comparative judgment

**Appendix 1. Rubrics Self-reflection**

| criteria | insufficient (1 to 3/8) | sufficient (4/8 or 5/8) | good (6/8) | very good (7/8 or 8/8) |
|---|---|---|---|---|
| **1. Relevancy of the choice of topic** | The topic and experience are general and do not add something to the student's learning process. | The topic and experience are general but add small insights concerning the student's own functioning and learning process. | The topic and experience are personal and trigger the student to investigate his/her own functioning and learning process to initiate change. | The topic and experience are personal and original; in this way they provide very valuable insights concerning the student's own functioning and learning process and really initiate change. |
| **2. Use of all phases of Korthagen (ALACT) – *content is conform to the phases*** | Multiple phases are not present or insufficiently described | Maximum 1 phase is missing (excl. phase 5) or multiple phases are only moderately described | Phases 1 until 4 are described with a clear focus and fluently follow one after the other (phase 5 can be lacking) | Phases 1 until 4 or 5 are explicitly and concretely described with a clear focus and fluently follow one after the other. The attention to all phases is balanced. |
| **3. Personal point of view** | The student describes a situation and gives his/her own opinion on this situation. The situation is not used to investigate his/her own functioning. | Personal point of view varies across the different phases. Moderate exploration of the student's own thinking, feeling, willing and acting. | Personal point of view is clearly present in the different phases. The student has a clear view on his/her own functioning. | *'Authentic'*<br>In all phases the student focusses on his/her own process. Exploration of own experiences on the dimensions of thinking, feeling, willing and acting. Critical for him/herself, in a healthy way. |