Studies in Educational Evaluation 76 (2023) 101233

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Why increasing the number of raters only helps sometimes: Reliability and validity of peer assessment across tasks of different complexity

Yimin Tong [a,*], Christian D. Schunn [b], Hong Wang [a]

[a] School of Foreign Languages, Dalian University of Technology, No.2 Linggong Road, Ganjingzi District, Dalian 116024, China
[b] Learning Research and Development Center, University of Pittsburgh, 3420 Forbes Ave., Pittsburgh, PA 15260, USA

ARTICLE INFO

Keywords:
Validity
Reliability
Number of raters
Task complexity
Peer assessment

ABSTRACT

Number of raters is theoretically central to peer assessment reliability and validity, yet rarely studied. Further, requiring each student to assess more peers' documents both increases the number of evaluations per document but also assessor workload, which can decline performance. Moreover, task complexity is likely a moderating factor, influencing both workload and validity. This study examined whether changing the number of required peer assessments per student / number of raters per document affected peer assessment reliability and validity for tasks at different levels of task complexity. 181 students completed and provided peer assessments for tasks at three levels of task complexity: low complexity (dictation), medium complexity (oral imitation), and high complexity (writing). Adequate validity of peer assessments was observed for all three task complexities at low reviewing loads. However, the impacts of increasing reviewing load varied by reliability vs. validity outcomes and by task complexity.

1. Introduction

Peer assessment is increasingly adopted for a number of reasons (Cho & MacArthur, 2011; Sadler & Good, 2006). As a pedagogical technique, students can learn from providing constructive feedback to their peers (Lundstrom & Baker, 2009; Huisman et al., 2019). On a pragmatic level, when facing large numbers of students, peer assessment can address teachers' limited capacity to assess students' assignments and provide elaborated comments in a timely fashion (Comer et al., 2014). Peer assessment is sometimes distinguished from peer review or peer feedback in that it provides scoring-based evaluation for assessing students' learning achievements (Zhang et al., 2020), and, as a collaborative learning activity, it can be effectively used in wide variety of contexts, including both English as a foreign language (EFL) and non-EFL classes (Cheng et al., 2014; Cho et al., 2006; Li et al., 2012).

Student and instructor concerns about the validity and reliability of peer assessments are regularly highlighted as hindering the broader adoption of peer assessment (Chang et al., 2011; Kaufman & Schunn, 2011; Luo et al., 2017; Topping, 1998). A number of factors have been shown to influence the validity and reliability of peer assessment, such as sufficient training (Ozogul & Sullivan, 2009; Saito, 2008), having clear rating criteria (Ashton & Davies, 2015; Falchikov & Goldfinch, 2000), anonymity (Vanderhoven et al., 2015), and class grade level (Zhang et al., 2020). However, the number of reviews completed per reviewer is often viewed as having a strong influence on the accuracy of peer review: when reliability is low, simply increase the number of raters to produce a more reliable average score (Jeffery et al., 2016). But what exactly is the relationship between the number of raters and the validity and reliability of multi-peer assessment? Increasing the number of raters does not always significantly increase the correlation of average peer ratings with expert ratings (Cho & Schunn, 2018), perhaps because students will not continue to put significant effort into each assessment when asked to assess many peers' contributions.

In addition, task complexity may be a critical factor having an overall influence and moderating the relationship between number of raters and validity/reliability of peer assessment. For example, if a task of assessing speaking is less complex than a task of assessing writing, will fewer peers be needed to obtain reliable and valid peer assessments? However, as the previous research focus in peer assessment has been mainly on writing (Cho et al., 2006; Saito & Fujita, 2004; Yu & Lee, 2014), the variation in the validity and reliability of peer assessment across assessment tasks of different complexity remains to be addressed. Therefore, the current study examined the effect of number of peer raters on validity and reliability for peer assessment task of varying

* Corresponding author.
E-mail addresses: tym999@dlut.edu.cn (Y. Tong), schunn@pitt.edu (C.D. Schunn), 17854173761@163.com (H. Wang).

https://doi.org/10.1016/j.stueduc.2022.101233
Received 20 June 2022; Received in revised form 30 October 2022; Accepted 15 December 2022
Available online 27 December 2022
0191-491X/

complexity.

2. Literature review

An efficient assessment method should meet the two conditions of high validity and reliability (Gay & Airasian, 2003). Validity in research on peer assessment refers to agreement of peer ratings with either teacher ratings or expert ratings (Falchikov & Goldfinch, 2000). This validity can sometimes be measured in terms of the degree of agreement (e.g., percent of scores with exact agreement, or mean distance) between peer ratings and teacher or expert ratings (Falchikov, 1986; Rushton et al., 1993; Li et al., 2016). Others have examined validity in terms of the consistency across rated objects (e.g., correlations) between peer and teacher or expert ratings (e.g., Cho et al., 2006). By contrast, reliability is the extent to which assessments will produce the same values if re-assessed; it is typically measured in terms of inter-rater reliability (Shrout & Fleiss, 1979; Li et al., 2016). Conceptually, it is possible to have assessment methods that are low on reliability but high on validity or low on validity but high on reliability. However, in practice, low reliability of a measure can be thought of as a kind of measurement noise which lowers typical measures of validity. Indeed, one analysis of over a thousand peer assessment rubrics found a moderate relationship between reliability of the rubric and validity of the rubric after controlling for other differences in the assignment and the course context (Xiong & Schunn, 2021).

2.1. Number of raters

In theory, both the reliability of an average of multiple peer ratings and the validity of the average of multiple peer ratings should be affected by the number of peer raters assessing each document. Simply from the law of large numbers, the greater the number of contributing data points, the more stable the average (Cho et al., 2006). In particular, the reliability of a mean of multiple assessments should follow a simple, monotonically-increasing curve based upon the reliability of a single assessor (Cho et al., 2006). However, few studies have empirically examined the effect of the number of raters on validity and reliability of peer assessment, and they have had mixed conclusions.

A potentially important factor underlying these mixed effects involves the two consequences of increasing the number of raters assigned to each document: 1) each document receives more assessments; and 2) each assessor has more work. Note that this implication does not always apply. Within an assignment, because some students fail to complete peer assessments or some students do additional assessments beyond what was assigned, within one class the workload of each student can vary and the number of assessments each document receives can vary. Across assignments or classes, instructors can create situations in which the number of assessments completed and the number of assessments received are not equal, such as group assignments where groups submit but individuals review. However, most commonly across assignments and classes, there is a close relationship between the two elements number of assessments per document and assessor workload, and an instructor needs to make pragmatic decisions about how many documents to assign to each peer based on the impact on both elements.

The inherently correlated aspects of assessor workload and number of assessments per document could provide an explanation for the mixed research findings on this topic. On the one hand, in support of the simple theoretical prediction for the reliability of the average rating increasing with number of peer raters, a number of researchers found that increasing the number of raters can greatly improve the reliability (Cho et al., 2006; Luo et al., 2017), or that at least three raters could make online peer assessment reliable (Jeffery et al., 2016; Luo et al., 2017; Cho & Schunn, 2018). On the other hand, other scholars have focused on the student workload perspective and have proposed that the number of peer raters should not be too many, suggesting that substantially increasing the number of raters might change the assessor's effort and therefore quality of peer ratings, thus leading to the opposite effect (Suen, 2014; Cho & Schunn, 2018). Looking at the reliability of a single peer's ratings should be especially sensitive to measuring workload effects. However, even the reliability of the average rating across multiple raters could show declines when the negative workload effects are large enough.

Suen (2014), without supporting empirical evidence, proposed that each student rater should rate no more than a handful of other students' assignments. Li et al.'s (2016) meta-analysis on the validity of peer assessment found that having more than 6 peer raters did not improve the validity of the assessments. However, moderator analyses in meta-analyses depend upon correlations between studies, and there could have been confounds between, for example, the number of raters and the complexity of the tasks being assessed. Indeed, one would expect instructors to naturally avoid requiring students to assess many peer documents when they are being asked to assess especially complex documents. Examining natural variation in number of raters within a classroom, Cho and Schunn (2018) suggested that reliability and validity declined when there were more than 5 raters because student effort declines when students are given too many documents to assess. However, again natural variation could also be subject to confounds (e.g., additional reviews were produced near the reviewing task deadline and were rushed; Xiong & Schunn, 2021). In addition, as we argue below, there are reasons to suspect that such simple conclusions may not generalize across tasks of varying complexity.

2.2. Task complexity

Task complexity has not been directly examined in relation to peer assessment reliability and validity. More generally, task complexity has been found to be an important factor that influences and predicts human performance and behaviour. In the literature related to task complexity, there are various definitions and models of task complexity (Liu & Li, 2012). According to Wood's (1986) influential paper, task complexity is a function of the number of task elements of which the task is composed and the interdependence between those elements; a task with high complexity has more elements and increased interconnections between the elements. Robinson's (2001) more specific analysis of task complexity in linguistics focused on resource requirements. In particular, this framework specifies that task complexity is "the result of the attentional, memory, reasoning, and other information demands imposed by the structure of the task on the language learner". This framework also notes that task complexity can be heavily affected by the structure and design of the task, which influences the cognitive processing requirements of the task. In the context of peer assessment, the details of not only the performance task but also the breadth and depth of the assessment rubric will determine the complexity of the assessment task. For example, if the assessment rubric is narrowly focused upon only one aspect of the performance task, the complexity of the assessment task is lowered.

Since peer assessment tasks of differing complexity may require different amounts of cognitive resources, they should lead to different qualities of performance of peer raters. van Zundert et al. (2010) experimentally manipulated task complexity and found a strong negative effect on the accuracy of peer assessment of comments (i.e., ability to accurately detect problems in fictitious peers' solutions), but they examined neither reliability nor validity of ratings. Further, the interaction with review load / number of peer raters per document is unknown. Increased task complexity may lead to each peers noticing a smaller subset of existing problems within a document and therefore specifically require a greater number of peer raters to accurately evaluate the document.

It is also important to note that there can be differential effects of reviewing load / number of raters on reliability vs. validity across tasks of differing complexity. Reliability can be high either when students consistently find the same problems or when students consistently miss

the same problems (Gao et al., 2018). If students choose to put less effort into the assessment task as number of documents to rate becomes high, this shift in effort would hurt the validity of the assessments, but it would not necessarily hurt the reliability of the assessments. Such a motivational shift might be especially pronounced in more complex tasks. Therefore, our research explores the effect of task complexity and its interaction with number of raters for both validity and reliability of peer assessment.

2.3. Peer assessment in EFL

There has been a surge of interest in peer assessment in EFL contexts (Liu & Ji, 2018; Yu & Lee, 2014; Zhang et al., 2020), and this is the context of our study. Studies of EFL peer assessment have included a range of performance tasks, such as writing (Wang, 2014; Yu & Lee, 2016), translation (Korol, 2019; Tseng et al., 2018), interpretation (Han, 2018; Su, 2018), and speaking (Li et al., 2022; Chien et al., 2020). Generally, peer assessment in EFL has been found to be well correlated with instructor assessments (Jafarpur, 1991; Patri, 2002; Saito & Fujita, 2004) depending on many factors such as class grade level (Zhang et al., 2020) and training (Ozogul & Sullivan, 2009; Cui et al., 2022). However, limited attention has been given to two critical factors: number of raters and task complexity. Given the established research gap and the increasing popularity of online multi-peer assessment in EFL, it is useful to explore the effect of number of raters and task complexity in an EFL context.

In general, this study aims to examine whether and how the number of peer raters (number of raters per document / number of assessments per peer) and complexity of assessment tasks affect reliability and validity of multi-peer assessment in EFL courses. Three different typical performance tasks from an EFL course are used: dictation, oral imitation, and writing. The specific assessment tasks applied to these performance tasks were analysed using the Robinson framework and found to fall into three different complexity levels: the dictation assessment tasks had low complexity, the oral imitation assessment tasks had medium complexity, and the writing assessment tasks had high complexity. In this EFL context, the study addresses the following specific research questions:

1. What is the effect of reviewing load/number of raters per document and assessment task complexity on assessment *reliability* at the single-rater and multi-rater levels of analysis?
2. What is the effect of reviewing load/number of raters per document and task complexity on assessment *validity* in terms of consistency with instructor ratings and exact agreement with instructor ratings?
3. To what extent are the changes in peer assessment validity attributable to changes in reliability?

3. Method

3.1. Participants and course context

Data were collected from 181 masters students who were non-English majors (44% female and 56% male) in a course called English Reading and Writing (32 h across 16 weeks) in a public research university in Northeast China. All six sections (with approximately 30 students each) of the course were taught by the same instructor, who had approximately 30 years of EFL teaching experience. The primary goals of the course were to teach short essay writing skills, reading comprehension skills, and pronunciation skills for English as a Foreign Language. The students were placed into this course based upon having lower performance on an English placement test, thereby producing a set of students that were roughly at similar levels of English competence across the six sections.

3.2. Study design

In this course, all students had to complete online assignments involving dictation (two assignments), oral imitation (two assignments), and writing (three assignments). Students also had to complete peer assessments for each of these assignments, but the number of assessments each student had to complete for a given assignment was varied across assignments in a counter-balanced fashion. The main research design involves examining the reliability and validity of peer assessments using different measures of reliability and validity and then analysing the patterns according to assessment task complexity and number of raters per document.

For the dictation assignments, students were required to dictate an audio recording on the topics of love and family, involving about 200–300 English words. For the oral imitation assignments, students were required to imitate a 1-minute recording in English on the topics of 1) travel and food, and 2) popular science. Finally, for the writing assignments, students were asked to write a five-paragraph essay in English, gradually expanding in length across writing assignments, on one topic of their choosing from the possible topics of science fiction, adolescent rebellion, human beings and viruses, and happiness. For the first assignment, they wrote the introduction. In the second assignment, they revised the introduction and added the body parts of the essay. In the final assignment, they revised the prior essay while adding a conclusion. The final essay was expected to be approximately 300 words in length.

For each of the peer assessment tasks, students were asked to provide a single holistic grade for each assessed object. The specific rubrics for each peer assessment task varied in ways that reflected the different emphasis and intended outcome of each assignment (see Appendix A for rubrics given to students and their English translations). The rubric for dictation addressed spelling and completeness, asking reviewers to estimate the percentage of the given audio file that the students had written down correctly; assessors were provided with a transcript of the audio file. The rubric for oral imitation focused on whether students had correct pronunciation and intonation, and whether they were fluent, which was defined as speaking at a similar speed as in the audio. The rubric for writing had separate dimensions for ideas (whether there were interesting ideas and supporting evidence), language competence (wording and sentence skills), and organization (coherence and cohesion of sentences, structure within a paragraph and between paragraphs).

Applying Robinson's Triadic Framework for task complexity, the specific peer assessment tasks in this course (i.e., assessment rubrics applied to the specific performance tasks) were categorized into three different levels of relative task complexity (see Table 1). Note that this analysis depends on the details of the assigned performance tasks and rubric; it is not the case that all possible assessment-of-writing tasks are necessarily of greater complexity than all assessment-of-imitation tasks. Not shown in the table are the ways in which the three peer assessments are similar in terms of task complexity: students respond to information presented in the moment, there is no required perspective taking,

Table 1
Dimensions of task complexity that varied across the specific peer assessment tasks in the study (+ = higher complexity, – = lower complexity).

Dimensions of Task Complexity	Task Complexity		
	Low Peer Assessment of Dictation	Medium Peer Assessment of Oral Imitation	High Peer Assessment of Writing
Multiple task	—	—	+
Reasoning demands	—	+	+
Multiple elements	—	+	+

students can plan their response, and no outside prior knowledge is required. The specific assigned tasks did differ along three task complexity dimensions shown in the table: single vs. multiple tasks, presence of reasoning demands, and number of elements. Peer assessment of dictation was low on all three elements, peer assessment of oral imitation was low on one of the three elements, and peer assessment of writing was high on all three elements.

The primary reason why the peer assessment of dictation was considered to be a low complexity task (LCT) was that raters were expected to do grading only on the accuracy of the submitted dictations, with little subjective reasoning since a transcript answer was provided, and relatively few elements (of spelling and completeness) were involved. By contrast, the assessment of oral imitation task was regarded as a medium complexity task (MCT) because students were expected to grade the pronunciation and intonation as a whole (i.e., as a single task), but there was subjectivity of evaluating pronunciation and intonation that added reasoning demands, and there were more elements involved such as the pronunciation of vowels, word stress, rhythm, tone, and linking. The assessment of writing task was regarded as a high complexity task (HCT) because it included grading language competence, organization, and ideas (which are each relatively distinct tasks), each of those component tasks had multiple elements (e.g., just language competence involved spelling, grammar, and word choice), and the evaluation of ideas involved substantial reasoning.

3.3. Measures

3.3.1. Rating reliability

The peer ratings data were analyzed for inter-rater reliability on each rubric in terms of the Intraclass Correlation Coefficient (ICC; Zhang et al., 2020). There are many different forms of ICC (Koo & Li, 2016; McGraw & Wong, 1996). We apply the ICC naming convention proposed by McGraw and Wong. When data is analyzed as a one-way random analysis, which is necessarily a kind of absolute agreement, there is one parameter to consider, which is whether the focus is on the reliability of a single rating, ICC(1), or the reliability of the average of *k* ratings, ICC (k). When data is analyzed as a two-way random or mixed analysis, there are more variations possible, and they involve two parameters. The first parameter indicates whether the reliability is based upon absolute agreement, A, or consistency in relative scores, C. The second parameter is the same as the one for one-way random analysis: single rating vs. average of *k* ratings. For example, ICC(C,1) is the consistency reliability of a single rating. See McGraw and Wong for specific formulas associated with each ICC type.

This study measured reliability in terms of agreement of a single rater, ICC(1), and the average rating across multiple raters, ICC(k). Reliability was measured in terms of agreement because different reviewers were assigned different documents, so the formula for reliability in terms of consistency of ratings across a shared set of documents could not be calculated. Also from a practice perspective, the variation in standards that each peer uses adds nuisance variance that necessarily degrades measure reliability, and it is thus important to not ignore such variation in standards by using consistency measures. The single-rater reliability measure shows the impact of task and reviewing load on an individual reviewer. The multiple-rater reliability measure shows the net impact of task and reviewing load/number of raters per document on the resulting mean rating across raters. The former speaks to psychological processes in the reviewer. The latter speaks to pragmatic impacts on assessment reliability.

Landis and Koch (1977) suggested that the following standards for ICC: 0.80 - 1.00 is considered to be almost perfect agreement; 0.61 - 0.80 is substantial agreement; 0.41 - 0.60 is moderate agreement; 0.21 - 0.40 is fair agreement; 0.01 - 0.20 is slight agreement, and < 0 is poor agreement.

3.3.2. Rating validity

Validity of the mean rating for a document across student raters was calculated using two different methods. First, the consistency of the mean student rating with instructor scores was calculated using the Pearson correlation coefficient with instructor scores, equivalent to a consistency-type ICC(C,1). Second, the agreement of the mean student rating with instructor scores was calculated using ICC(1), the same inter-rater (one-way) agreement reliability formula, now applied to the mean student rating vs. instructor ratings rather than student vs. student ratings. The same standards for validity ICCs were applied as with student inter-rater reliability ICCs.

3.4. Procedure

Students were allocated a randomly-selected and anonymized set of peer documents for grading and comments according to holistic scoring rubrics in the week after the documents were submitted. The instructor gave a lecture on each new rubric as each new type of peer assessment task was assigned; however, given the wide range of assessment tasks implemented in the course, it was not possible to also include time for guided practice on each assessment task. Both document submission and peer assessment were implemented on the Chaoxing Platform, a popular online teaching and learning platform in Chinese universities. The mean score across raters for each document was generated by the platform, while the instructor also graded all the documents.

While each student was required to submit a document for all seven assignments (two dictation, two oral imitation, and three writing), some students failed to submit a document for some of the assignments. As a result, the 181 students only produced 1048 documents in total (rather than the expected 1267 possible documents). Submitted documents were then randomly assigned to peers according to reviewing load condition, varying from 3 to 9 documents. For analysis, we treat each student as being randomly assigned to complete either a low (3−4), medium (5−7), and high (8−9) number of peer assessments within a given task. Creating the three categories generated sufficient data within each category for sufficient statistical power, while also reducing the number of contrasts that would be made in analysis. However, having different exact numbers within each category allows for better generalization to the full range within each category.

Taking into account both the varying number of submitted documents for each assignment and the varying number of assignments per task type (i.e., two dictation, two oral imitation, and three writing), the mean number of documents for analysis in each exact number of raters (3−9) was approximately 40 for dictation (LCT), 50 for oral imitation (MCT), and 55 for writing (HCT), respectively. Note then that the grouping of exact number of raters into ranges (i.e., 3–4, 5–7, or 8–9) doubles or triples the number of documents for the analysis categories, which then vary from $N = 79$ to $N = 173$ (see Appendix B for table of exact N for each case).

3.5. Analysis

To visually describe the effects of Reviewing Load/Number of Raters per Document and Task Complexity on peer assessment reliability and validity, mean ICC values for each condition were calculated using the different types of ICCs. To evaluate the statistical significance of effects of Reviewing Load/Number of Raters per Document and Task Complexity on reliability and validity ICC values, we applied a Fischer r-to-z transformation to each ICC value because ICCs are not normally distributed. Pairwise contrasts of mean ICC values then involve simple difference scores in pairs of transformed z-values. The resulting different scores are themselves z-scores that can then compared to z-score cut-off values at different p-values ($z = 2.58$, 3.29, and 3.82 respectively for p-values =0.01,0.001, and.0001). Because of the number of comparisons being made, we set $p < .01$ as the alpha level.

4. Results

4.1. Single and multi-rater reliability of tasks varying complexity

Fig. 1 presents the mean single-rater and multi-rater reliabilities as a function of number of documents assigned to each student and task complexity. For single-rater reliability, there is a clear and strong interaction between task complexity and reviewing load. For the Low Complexity Task, single-rater reliability declined as reviewing load increased (3–4 vs. 8–9 $z = 2.44$, $p < .05$). By contrast, for the High Complexity Task, single-rater reliability actually increased as reviewing load increased (3–4 vs. 8–9 $z = 2.67$, $p < .01$). The Medium Complexity Task fell in the middle, with relatively little impact of reviewing load on reliability ($zs < 1.1$, $ps > 0.3$). Examined from the perspective of effects of Task Complexity, only in the case of low reviewing load were the effects on single-rater reliability statistically significant, with the Low Complexity task having higher reliabilities than the Medium Complexity task ($z = 2.85$, $p < .01$) and the High Complexity Task ($z = 3.56$, $p < .001$).

Relative to the single-rater reliabilities, multi-rater reliabilities of course show a net benefit of having more raters per document, a simple consequence of the law of large numbers. But this pattern played out in a differential way across Task Complexity. For the Low Complexity Task, there was almost perfect agreement across all reviewing loads ($zs < 1.2$, $ps > 0.3$). For the High Complexity Task, reliability varied greatly across reviewing load (3–4 vs. 8–9 $z = 5.71$, $p < .0001$), ranging from fair agreement with few documents per reviewer to considerable agreement with a medium number of documents per review to almost perfect agreement with a large number of documents per reviewer. This time, the pattern for the Medium Complexity task was relatively similar to the High Complexity Case, with a substantial positive effect of reviewing load on reliability (3–4 vs. 8–9 $z = 3.01$, $p < .01$). Again, from the perspective of effects of Task Complexity on multi-rater reliability, there was only a significant effect in the low reviewer load case, logically mirroring the single-rater reliability pattern: the Low Complexity task had significantly higher reliabilities than did the Medium Complexity task ($z = 4.04$, $p < .0001$) and the High Complexity Task ($z = 5.66$, $p < .0001$).

4.2. Agreement and consistency validity of mean student ratings

Fig. 2 presents the mean validity values in terms of consistency and agreement. In general, consistency values are higher than agreement values (Cohen's d=0.97; $p < .001$): students more consistently replicate instructors' relative ratings of document quality than they do the exact ratings.

Consistency validity is near perfect in the Low Complexity Task regardless of number of reviewing load ($zs < 1.6$, $ps > 0.2$). For the Medium Complexity Task, consistency validity was gradually increasing from substantial with low reviewing load to near perfect for high reviewing load (3–4 vs. 8–9 $z = 2.85$, $p < .01$). By contrast, for the High Complexity Task, consistency validity trended towards declining from substantial to moderate with increasing reviewing load (3–4 vs. 5–7 $z = 2.37$, $p < .05$). From the perspective of effects of Task Complexity on consistency validity, the effects were marginal within low reviewer load (Low vs. High Complexity Tasks $z = 6.91$, $p < .0001$) and very large with medium reviewer load (Low vs. High Complexity Tasks $z = 5.44$, $p < .00001$) and high reviewer load (Low vs. High Complexity Tasks $z = 2.01$, $p < .0001$).

Validity agreement scores showed similar overall effects of both conditions, likely reflecting a close correspondence between the two kinds of validity measures; the mean agreement scores and mean validity scores within each condition were correlated at $r = 0.84$. The main difference in results, beyond agreement scores simply being lower, was

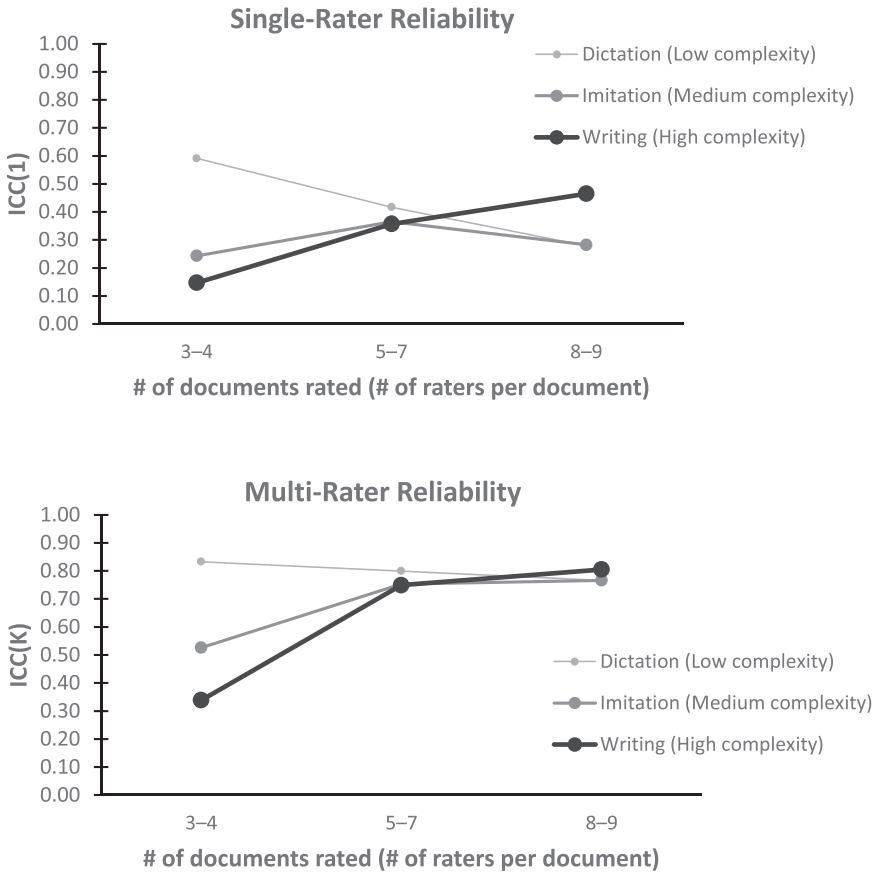

Fig. 1. Mean single-rater reliability (top) and multi-rater reliability (bottom) as a function of task complexity and number of documents rated by each student.

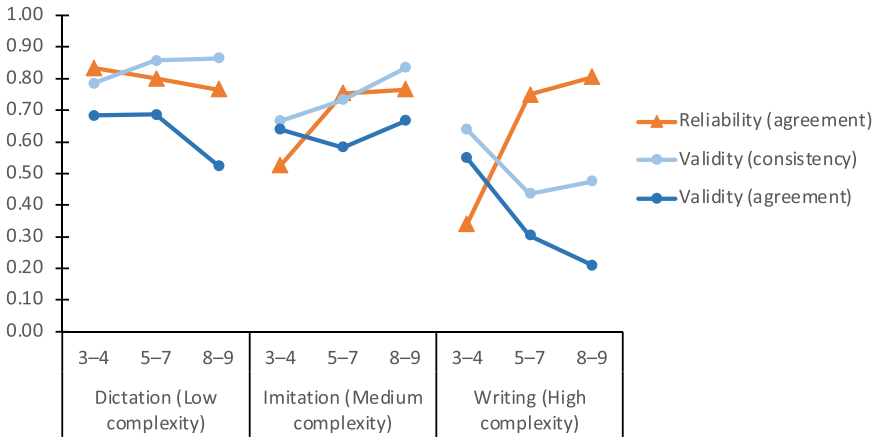

Fig. 2. Consistency and agreement validity (blue), along with multi-rater reliability (orange), as a function of task complexity and number of documents rated by each student.

that reviewing load more often had a negative effect on agreement across tasks. This negative effect of reviewing load was only statistically significant for the High Complexity Task (3–4 vs. 8–9 $z = 3.02, p < .01$). Conversely, the effects of Task Complexity continued to only be statistically significant within medium review loads (Low vs. High Complexity Tasks $z = 4.48$, $p < .00001$) and high reviewing loads (Medium vs. High Complexity Tasks $z = 4.32, p < .00001$).

4.3. Relationship of validity to reliability

Fig. 2 also showed the mean multi-rater reliabilities alongside the validity values. In general, at the level of condition means, there was very little relationship between the reliability of the mean student ratings and their validities in terms of consistency ($r = 0.19$) and agreement ($r = -0.10$). In other words, the variation in validity by condition was not due to variation in reliability. This divergence was particularly salient in the High Complexity Task, which had very high reliability and very low validity with high reviewing loads.

5. Discussion

As expected, assessment task complexity substantially influenced many aspects of reliability and validity of peer assessment. First, tasks of higher complexity tended to have lower validity of peer assessments. Second, task complexity moderated the relationship of number of raters to the reliability of peer assessment: number of raters had little effect on reliability at low complexity (all cases were high reliability), but it had a large effect on reliability at high complexity. Third, task complexity moderated the relationship between changes in reliability and validity across number of raters: at medium task complexity, both reliability and validity increased with increasing reviewing load, but at high complexity, reliability increased and validity decreased with increasing reviewing load.

From a pragmatic perspective, if validity of.6 in terms of agreement or.7 in terms of consistency is considered sufficient, having a reviewing load of 3–4 was sufficient for adequate validity across all task complexities. Those levels of validity are similar to the reliability of single instructor scores (Cho et al., 2006). Increasing the reviewing load only improved validity in the case of the medium complex task, and the gain was relatively modest. Increasing the reviewing load had large negative effects on validity for the high complexity task. Understanding exactly which tasks will be of moderate or high complexity may be challenging for instructors given the many factors that enter into determining task complexity (Robinson, 2001). Therefore, the safest general recommendation for instructors could be to only assign 3–4 reviews per student.

From a theoretical perspective, the current study has several implications. First, the current study highlights the importance of considering the assessment task being given to peers. Particularly for examinations of assessment reliability and validity, generalizable claims in future research are unlikely to be possible from studies that use just one assessment task. Task complexity not only overall increased or decreased reliability and validity of peer assessments, it also moderated the effects of another variable (i.e., reviewing load).

Second, the current study drew attention to the two different impacts of requiring more students to review more documents: a growing load for the assessor and an increasing number of assessments per document. Prior research has tended to emphasis one or the other. For expert scoring, new experts can be added such that the number of ratings per document can be increased without added to each expert's load, and thus the number of ratings per document can be the focus of decision making: how many expert ratings are needed to produce acceptable validity? By contrast in peer assessment, there is typically a natural connection between number of raters per document and reviewing load, as in our study, where students were required to do both individual assignment and individual assessment (not do the performance task as a group assignment and then peer assessments as individuals). In such situations, every additional rating for a document means an additional review for a peer assessor.

Third, by examining validity from both consistency and agreement perspectives, it was possible to observe some important differences in effects. Prior research in peer assessment has tended to focus on one or the other. From an agreement perspective, a number of researchers have discussed overall biases in peer assessments. For example, students are sometimes overall too positive in their assessments (van Hattum-Janssen et al., 2004; Hafner & Hafner, 2003; Xiong & Schunn, 2021) and sometimes too negative (English et al., 2006). We observed that agreement validity was generally lower than consistency validity, likely reflecting overall biases in ratings. Since such biases can be addressed through grading a few benchmark documents or curving of results, other researchers have focused on validity in terms of consistency. Beyond the overall bias issue, we found some similarities in patterns of findings across consistency and agreement (e.g., lower validity with high complexity assessment tasks), but also important differences (e.g., more sensitivity to reviewer load in agreement scores than in consistency scores). It is therefore recommended that future research continue to measure validity from both perspectives.

Fourth, the current study highlights how divergent peer assessment reliability and validity can be. In particular, increasing reliability was sometimes associated with large decreases in validity, as might be predicted from a cognitive load perspective (van Zundert et al., 2010). We suggest that separating superficial agreement from deep agreement is likely to be critical to understand this separation. When students are

given a very complex task and conditions that lead them to assess peers' contributions relatively quickly (Usher & Barak, 2018), they may be in superficial agreement with one another based upon detecting only relatively obvious problems in the documents. By contrast, when students are given less complex tasks and have sufficient time to deeply process each document, then they may more consistently identify more subtle issues in the documents. Future research will be needed to test this interpretation. For example, one approach could involve examining the types of problems being detected through looking at comments rather than just ratings. Another approach could involve a lab study in which students are given a fixed set of essays to review that have a mixture of obvious and subtle problems.

Fifth, peer assessment validity was sometimes actually higher than its reliability (i.e., reliability was not necessarily an upper bound on validity). How can this be? Conceptually, reliability here is operationalized in terms of reliability across assessors (as opposed to across measures or time). Past research has found that disagreements among peer comments often take the form of complementary observations (Patchan et al., 2013): each student finds different errors in the document. If such a pattern also occurred in this dataset, then an aggregate evaluation across peers could produce a good holistic assessment even when individual ratings disagree (Cheng & Warren, 2000; Suen, 2014).

5.1. Caveats and future directions

Several caveats regarding the current study must be acknowledged. First, three different tasks were used to represent differing complexity levels, but complexity differences were not the only differences among these tasks. Task complexity is driven by many aspects of a task. Future research will need to examine a broader range of tasks to test whether task complexity per se is critical to understand reliability and validity of peer assessment at differing reviewing loads. Further, future research should also examine whether specific dimensions of task complexity are particularly important (e.g., inclusion of multiple tasks vs. reasoning demands).

The current study examined ranges of reviewing load, but not with large enough samples to precisely examine reliability and validity at specific reviewing loads. From a pragmatic perspective, testing reviewing loads of 3, 4, and 5 specifically could be important for giving useful guidance to instructors. These are the more common reviewing loads assigned by instructors (Cho et al., 2006; Li et al., 2016).

The current study only provided a moderate amount of training on peer assessment: one lecture on each rubric. Some researchers have tested more detailed forms of peer assessment training (Cui et al., 2021; Min, 2006), which sometimes involves guided practice and whole class discussion. This approach also can take considerable class time overall in the case of a class using multiple rubrics across many assignments, like in the current setting. However, other courses can focus on a particular kind of assessment task, reusing a rubric. Higher levels of reliability and validity may result when more time is spent on training.

Finally, the current study involved holistic rubrics (i.e., a single overall rating for a peer's contribution) and did not hold peers accountable for the accuracy of their ratings. Providing more scaffolding for the assessment process, either in the form of more detailed, analytic assessment rubrics (Chi, 2001) or accountability to produce accurate ratings (Patchan et al., 2018) may have reduced the negative impact of reviewing load on assessment validity or generally increased assessment validity overall.

Funding

The research presented in this paper was supported by grants from Education Plan Bureau of Liaoning Province (Grant No. JG20DB096), and China National MTI Education Committee (Grant No. MTIJZW202151).

Declaration of Competing Interest

The authors declare that there is no conflict of interest.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.stueduc.2022.101233.

References

Ashton, S., & Davies, R. S. (2015). Using scaffolded rubrics to improve peer assessment in a MOOC writing course. *Distance Education, 36*(3), 312–334. https://doi.org/10.1080/01587919.2015.1081733

Chang, C., Tseng, K., Chou, P., & Chen, Y. (2011). Reliability and validity of web-based portfolio peer assessment: a case study for a senior high school's students taking computer course. *Computers & Education, 57*(1), 1306–1316. https://doi.org/10.1016/j.compedu.2011.01.014

Cheng, K., Hou, H., & Wu, S. (2014). Exploring students' emotional responses and participation in an online peer assessment activity: A case study. *Interactive Learning Environments, 22*(3), 271–287. https://doi.org/10.1080/10494820.2011.649766

Cheng, W., & Warren, M. (2000). Making a difference: Using peers to assess individual students' contributions to a group project. *Teaching in Higher Education, 5*(2), 243–255. https://doi.org/10.1080/135625100114885

Chi, E. (2001). Comparing holistic and analytic scoring for performance assessment with many-facet Rasch model. *Journal of Applied Measurement, 2*(4), 379–388.

Chien, S., Hwang, G., & Jong, M. (2020). Effects of peer assessment within the context of spherical video-based virtual reality on EFL students' English-Speaking performance and learning perceptions. *Computers & Education, 146*, 1–20. https://doi.org/10.1016/j.compedu.2019.103751

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of Educational Psychology, 103*(1), 73–84. https://doi.org/10.1037/a0021950

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. https://doi.org/10.1016/j.stueduc.2017.12.001

Cho, K., Schunn, C. D., & Wilson, R. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98*(4), 891–901. https://doi.org/10.1037/0022-0663.98.4.891

Comer, D. K., Clark, C. R., & Canelas, D. A. (2014). Writing to learn and learning to write across the disciplines: Peer-to-peer writing in introductory-level MOOCs. *The International Review of Research in Open and Distributed Learning, 15*(5). https://doi.org/10.19173/irrodl.v15i5.1850

Cui, Y., Schunn, C. D., & Gai, X. (2022). Peer feedback and teacher feedback: A comparative study of revision effectiveness in writing instruction for EFL learners. *Higher Education Research & Development, 41*(6), 1838–1854. https://doi.org/10.1080/07294360.2021.1969541

Cui, Y., Schunn, C. D., Gai, X., Jiang, Y., & Wang, Z. (2021). Effects of trained peer vs. teacher feedback on EFL students' writing performance, self-efficacy, and internalization of motivation. *Frontiers in Psychology, 12*, 6659. https://doi.org/10.3389/fpsyg.2021.788474

English, R., Brookes, S. T., Avery, K., Blazeby, J. M., & BenShlomo, Y. (2006). The effectiveness and reliability of peer-marking in first-year medical students. *Medical Education, 40*(10), 965–972. https://doi.org/10.1111/j.1365-2929.2006.02565.x

Falchikov, N. (1986). Product comparisons and process benefits of collaborative peer group and self-assessment. *Assessment & Evaluation in Higher Education, 11*(2), 146–166. https://doi.org/10.1080/0260293860110206

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. https://doi.org/10.3102/00346543070003287

Gao, Y., Schunn, C. D., & Yu, Q. (2018). The alignment of written peer feedback with draft problems and its impact on revision in peer assessment. *Assessment and Evaluation in Higher Education, 44*(2), 294–308. https://doi.org/10.1080/02602938.2018.1499075

Gay, L. R., & Airasian, P. (2003). *Educational research: Competencies for Analysis and Applications* (7th ed.,). Upper Saddle River, NJ: Merrill Prentice Hall,.

Hafner, J. C., & Hafner, P. M. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education, 25*(12), 1509–1528. https://doi.org/10.1080/0950069022000038268

Huisman, B., Saab, N., van den Broek, P., & van Driel, J. (2019). The impact of formative peer feedback on higher education students' academic writing: A Meta-Analysis. *Assessment & Evaluation in Higher Education, 44*(6), 863–880. https://doi.org/10.1080/02602938.2018.1545896

Jafarpur, A. (1991). Cohesiveness as a basis for evaluating compositions. *System, 19*(4), 459–465. https://doi.org/10.1016/0346-251X(91)90026-L

Jeffery, D., Yankulovb, K., Crerarc, A., & Ritchie, K. (2016). How to achieve accurate peer assessment for high value written assignments in a senior undergraduate course. *Assessment & Evaluation in Higher Education. 41*(1), 127–140. https://doi.org/10.1080/02602938.2014.987721

Kaufman, J. H., & Schunn, C. D. (2011). Students' perceptions about peer assessment for writing: Their origin and impact on revision work. *Instructional Science, 39*(3), 387–406. https://doi.org/10.1007/s11251-010-9133-6

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Korol, T. (2019). Translation project as an assessment tool: Ukrainian context. *The Journal of Teaching English for Specific and Academic Purposes, 7*(1), 115–123. https://doi.org/10.22190/JTESAP1901115K

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Li, H., Xiong, Y., Zang, X., Kornhaber, L., Lyu, M., Chung, Y., K, S., & Suen, H. K. (2016). Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education, 41*(2), 245–264. https://doi.org/10.1080/02602938.2014.999746

Li, J., Huang, J., & Cheng, S. (2022). The reliability, effectiveness, and benefits of peer assessment in college EFL speaking classrooms: Student and teacher perspectives. *Studies in Educational Evaluation, 72*, 1–11. https://doi.org/10.1016/j.stueduc.2021.101120

Li, L., Liu, X., & Zhou, Y. (2012). Give and take: A re-analysis of assessor and assessee's roles in technology facilitated peer assessment. *British Journal of Educational Technology, 43*(3), 376–384. https://doi.org/10.1111/j.1467-8535.2011.01180.x

Liu, L., & Ji, X. (2018). A study on the acceptability and validity of peer scoring in Chinese university EFL writing classrooms. *Foreign Language World, 5*, 63–70.

Liu, P., & Li, Z. (2012). Task complexity: A review and conceptualization framework. *International Journal of Industrial Ergonomics, 42*(6), 553–568. https://doi.org/10.1016/j.ergon.2012.09.001

Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing, 18*(1), 30–43. https://doi.org/10.1016/j.jslw.2008.06.002

Luo, H., Zuo, M., & Robinson, A. (2017). An empirical study on the effect of peer assessment in massive open online learning. *Open Education Research, 23*(1), 73–82. https://doi.org/10.13966/j.cnki.kfjyyj.2017.01.009

Min, H. (2006). The effects of trained peer review on EFL students' revision types and writing quality. *Journal of Second Language Writing, 15*, 118–141. https://doi.org/10.1016/j.jslw.2006.01.003

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Ozogul, G., & Sullivan, H. (2009). Student performance and attitudes under formative evaluation by teacher, self and peer evaluators. *Educational Technology Research and Development, 57*(3), 393–410. https://doi.org/10.1007/s11423-007-9052-7

Patchan, M. M., Hawk, B. H., Stevens, C. A., & Schunn, C. D. (2013). The effects of skill diversity on commenting and revisions. *Instructional Science, 41*(2), 381–405. https://doi.org/10.1007/s11251-012-9236-3

Patchan, M. M., Schunn, C. D., & Clark, R. J. (2018). Accountability in peer assessment: Examining the effects of reviewing grades on peer ratings and peer feedback. *Studies in Higher Education, 43*(12), 2263–2278. https://doi.org/10.1080/03075079.2017.1320374

Patri, M. (2002). The influence of peer feedback on self- and peer assessment of oral skills. *Language Testing, 19*(2), 109–131. https://doi.org/10.1191/0265532202lt224oa

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*(1), 27–57. https://doi.org/10.1093/applin/22.1.27

Rushton, C., Ramsey, P., & Rada, R. (1993). Peer assessment in a collaborative hypermedia environment: A case study. *Journal of Computer-Based Instruction, 20*(3), 73–80.

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

Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing, 25*(4), 553–581. https://doi.org/10.1177/0265532208094276

Saito, H., & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research, 8*(1), 31–54. https://doi.org/10.1191/1362168804lr133oa

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Su, W. (2018). Interpreting quality as evaluated by peer students. *The Interpreter and Translator Trainer. 13*(2), 1–13. https://doi.org/10.1080/1750399X.2018.1564192

Suen, H. K. (2014). Peer assessment for massive open online courses (MOOCs). *The International Review of Research in Open and Distributed Learning, 15*(3), 312–327. https://doi.org/10.19173/irrodl.v15i3.1680

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276. https://doi.org/10.3102/00346543068003249

Tseng, W., Su, T., & Nix, J. L. (2018). Validating translation test items via the many-facet Rasch model. *Psychological Reports, 122*(2), 748–772. https://doi.org/10.1177/0033294118768664

Usher, M., & Barak, M. (2018). Peer assessment in a project-based engineering course: Comparing between on-campus and online learning environments. *Assessment & Evaluation in Higher Education, 43*(5), 745–759. https://doi.org/10.1080/02602938.2017.1405238

van Hattum-Janssen, N., Pacheco, J. A., & Vasconcelos, R. M. (2004). The accuracy of student grading in first-year engineering courses. *European Journal of Engineering Education, 29*(2), 291–298. https://doi.org/10.1080/0304379032000157259

van Zundert, M., Sluijsmans, D., & van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction, 20*(4), 270–279. https://doi.org/10.1016/j.learninstruc.2009.08.004

Vanderhoven, E., Raes, A., Montrieux, H., Rotsaert, T., & Schellens, T. (2015). What if pupils can assess their peers anonymously? A quasi-experimental study. *Computers & Education, 81*, 123–132. https://doi.org/10.1016/j.compedu.2014.10.001

Wang, W. (2014). Students' perceptions of rubric-referenced peer feedback on EFL writing: A longitudinal inquiry. *Assessing Writing, 19*, 80–96. https://doi.org/10.1016/j.asw.2013.11.008

Wood, R. E. (1986). Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Process, 37*(1), 60–82. https://doi.org/10.1016/0749-5978(86)90044-0

Xiong, Y., & Schunn, C. D. (2021). Reviewer, essay, and reviewing-process characteristics that predict errors in web-based peer review. *Computers & Education, 166*. https://doi.org/10.1016/j.compedu.2021.104146

Yu, S., & Lee, I. (2014). An analysis of EFL students' use of first language in peer feedback of L2 writing. *System, 47*, 28–38. https://doi.org/10.1016/j.system.2014.08.007

Yu, S., & Lee, I. (2016). Peer feedback in second language writing (2005–2014). *Language Teaching, 49*(4), 461–493. https://doi.org/10.1017/S0261444816000016

Zhang, F., Schunn, C. D., Li, T., & Long, M. (2020). Changes in the reliability and validity of peer assessment across the college years. *Assessment & Evaluation in Higher Education, 45*(2), 1–15. https://doi.org/10.1080/02602938.2020.1724260