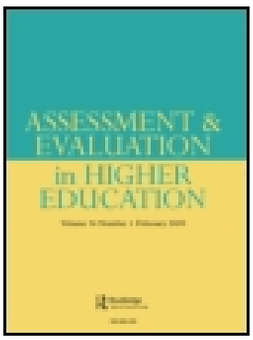


Routledge
Taylor & Francis Group

Assessment & Evaluation in Higher Education

ISSN: 0260-2938 (Print) 1469-297X (Online) Journal homepage: https://www.tandfonline.com/loi/caeh20

Changes in the reliability and validity of peer assessment across the college years

Fuhui Zhang, Christian Schunn, Wentao Li & Miyin Long

To cite this article: Fuhui Zhang, Christian Schunn, Wentao Li & Miyin Long (2020): Changes in the reliability and validity of peer assessment across the college years, Assessment & Evaluation in Higher Education, DOI: 10.1080/02602938.2020.1724260

To link to this article: https://doi.org/10.1080/02602938.2020.1724260

Published online: 07 Feb 2020.

ASSESSMENT & EVALUATION IN HIGHER EDUCATION
https://doi.org/10.1080/02602938.2020.1724260

Changes in the reliability and validity of peer assessment across the college years

Fuhui Zhang[a], Christian Schunn[b], Wentao Li[a] and Miyin Long[a]

[a]School of Foreign Languages, Northeast Normal University, Changchun, China; [b]Learning Research and Development Center, University of Pittsburgh, Pittsburgh, USA

ABSTRACT

Although a variety of learning benefits of peer assessment have been documented, many concerns remain regarding its reliability and validity, especially for peers at early learning stages and for English as a foreign language (EFL) learners in general. Such concerns may prevent adoption of peer assessment for instruction. To help localize the appropriate contexts for peer assessment, this study examined whether reliability and validity of peer assessment changes over years in a program either for overall scores or specifically for high-level dimensions or language conventions. Participants were 118 English major undergraduates in a comprehensive university in Northeast China. We found that the peer assessments for both 1st year and 4th year students had high reliability, similar to reliability levels shown by teachers. However, the validity of peer assessment showed developmental growth: the peer assessments of 1st year EFL students had low validity, especially for assessments of language conventions, but relatively strong validity of assessments of higher-level dimensions in both 1st and 4th year students. This study suggests peer assessment may be valid for assessments of higher-level concerns across broad developmental levels, but may require stronger supports when used for assessing lower levels of language, particularly for students early on in their language development.

KEYWORDS

Peer assessment; reliability; validity; EFL writing

Introduction

Peer assessment, sometimes also called peer review, peer evaluation or peer feedback, is a collaborative learning activity that can be effectively used in wide variety of contexts, including both English as a foreign language (EFL) and non-EFL classes (Cho, Schunn, and Wilson 2006; Li, Liu, and Zhou 2012; Cheng, Hou, and Wu 2014). Peer assessment is sometimes distinguished from peer review or peer feedback in that it provides scoring-based evaluation for assessing students' learning achievements. Peer assessment as a strategy of "assessment as learning" is increasingly widely adopted, with benefits for cognitive, social and language development (e.g. Liu and Hansen 2002; Lu and Law 2012).

One factor that frequently prevents broader adoption of peer assessment is concern by students and instructors about whether students have sufficient expertise to assess their peers (Boud 1989; Lynch and Golen 1992; Mowl and Pain 1995; Cheng and Warren 1997; Saito and Fujita 2004). Given the strong possible pedagogical benefits as well as pragmatic benefits of reducing grading burden (Sadler and Good 2006; Cho and MacArthur 2011), understanding

CONTACT Wentao Li liwt283@nenu.edu.cn

when peer assessment can be used is of great interest to teachers. A number of investigations have demonstrated good reliability and validity of peer assessments on average, especially for web-based peer assessment with clear rubrics and a structured assessment process (Cho, Schunn, and Wilson 2006; Paré and Joordens 2008; Li et al. 2016; Schunn, Godley, and DeMartino 2016). However, given that peer assessment can be used in a wide variety of contexts with learners varying greatly in ability, variation in reliability and validity by context remains an open question (Li et al. 2016).

What kinds of contexts might be particularly problematic for peer assessment? Since expertise of peers is the primary source of concern, contexts with especially low levels of student expertise need close scrutiny. The focus of the current investigation is on peer assessment of writing, which is a challenging assessment task even for experts (Weigle 1994; Cho, Schunn, and Wilson 2006), and more specifically writing in EFL context, where suspicion regarding validity and reliability remains high (Marcoulides and Simkin 1995; Mowl and Pain 1995; Saito and Fujita 2004; Bai 2013; Liu and Ji 2018). These learners are at low overall levels of expertise in evaluating the lower-level aspects of writing. However, even EFL is a varied context, with some students only beginning their studies with English and other students having already experienced multiple years of instruction in the foreign language. The current study examines how reliability and validity of peer assessment changes across years of instruction.

Research on validity and reliability of peer assessment

Validity refers to agreement between peer ratings and teacher or expert ratings (Falchikov and Goldfinch 2000). Research on peer assessment often measures validity in terms of the degree of agreement (e.g. percent agreement or correlation) between peer ratings and a single expert's rating (Falchikov 1986; Rushton, Ramsey, and Rada 1993; Stefani 1994). However, using one expert's rating as the criterion to evaluate peer assessment artificially lowers validity estimates since expert ratings sometimes show evidence of inconsistency (Ward, Gruppen, and Regehr 2002). Instead, it is better to examine consistency with the mean score generated by multiple expert scorers. In addition, the common metric of percentage agreement, although easy to understand, is not ideal because it is artificially influenced by distributional patterns (e.g. if one rating is especially common). Instead, the Pearson correlation coefficient presents a better measure of validity that is much less influenced by distributional patterns (Haaga 1993; Cho, Schunn, and Wilson 2006; Han 2018).

An older meta-analysis of 48 studies published between 1959 and 1999 found the mean correlation between peer and teacher ratings to be .69 (Falchikov and Goldfinch 2000). A more recent meta-analysis conducted on the results of 69 studies since 1999 found a similar mean correlation of .63 (Li et al. 2016), which was considered to be strong. However, the observed correlations varied between .33 and close to .86 across different scenarios that were investigated, with a number of contextual factors significantly moderating the observed correlations. For example, graduate courses showed higher validity correlations than did undergraduate courses, suggesting assessing skills may improve with additional instruction or expertise. However, the students that are enrolled in graduate courses are systematically different from the students enrolled in undergraduate courses, so developmental change is only one possible explanation.

Only a few studies of peer assessment validity have specifically taken place in the EFL context, a context with high concern about this issue. In studies of more advanced students, validity was found to be good. Azarnoosh (2013) found there was no significant difference between teacher and peer scoring with 3rd year Iranian students. Shao (2009) found high validity correlations with Chinese 2nd year students for analytic scoring, and Liu and Ji (2018) found that holistic scoring had adequate validity with 2nd year students. But Bai (2013) studied 1st year Chinese students and found mean teacher-peer score differences in ratings of language use and genre-specific

elements, but no teacher-peer score differences at the content level. These few studies suggest the earlier levels of language development may be of special concern, but there are many possible differences across the populations in these few prior studies. In general, more studies of analytic scoring agreement, especially looking across college levels, need to be conducted.

One common cause of low validity is low reliability (Ward, Gruppen, and Regehr 2002): if students cannot agree with each other, they will be less likely to agree with teacher or expert ratings. However, in the case of multi-peer review, it is possible that the mean correlation across peers is relatively well correlated with expert ratings, even when peers are disagreeing with each other. That is, following the Law of Large Numbers in statistics, the mean becomes increasingly more stable as more ratings are obtained. And yet, a mean across three to five peers might not be enough to overcome a very noisy rating process.

Sometimes reliability is assessed based on the correlation between pairs of raters. A more elegant solution involves an intraclass correlation coefficient (ICC), which can evaluate the inter-observer reliability across an arbitrary number of raters (Stemler 2004). ICC, (with values ranging from 0 to 1), is computed by a variance calculation, and .4 is considered an acceptable value (Shrout and Fleiss 1979). Cho, Schunn, and Wilson (2006) observed peer assessment ICC values ranging from .3 to .6, and Paré and Joordens (2008) reported a mean overall ICC value at .6. However, reliability of peer ratings has not been examined in the EFL context. Thus, based on the prior studies, it is unclear whether EFL learners (especially at beginning levels) are simply noisier raters or whether they systematically cannot produce valid ratings; if the issue is only noise, then additional raters can address the problem. If the issues are more systematic validity challenges, additional training or more appropriately-selected dimensions of evaluation may be necessary.

Overall, the study of validity and reliability of peer assessments in EFL writing remains largely unexplored, though peer assessment has been taken as a frequent strategy of formative assessment in writing or other contexts (Min 2006). From a writing instructor's perspective, peer assessment might be considered to be worthwhile as long as it benefits students in at least some ways (e.g. produces learning via providing feedback); however, students might be resistant to participating in peer assessment differently if validity or reliability is low.

What results might be expected of EFL learners in terms of validity and reliability? If provided good rubrics, they might be able to produce reliable and valid ratings since rubrics are generally helpful in that way (Berg 1999; Min 2006; Topping 2010; Panadero, Romero, and Strijbos 2013; Greenberg 2015; Russell et al. 2017). However, even with rubrics, learner ability can influence reviewing ability. Patchan and Schunn (2015) found that high ability reviewers pointed out more problems in low-quality papers than in high-quality papers. Huisman et al. (2018) found that reviewer ability was positively related to their own essay writing performance. But within a given course, reviewer ability and essay performance could be influenced by learner motivation for the overall course. Thus, the performance of EFL learners (or other learners low on the expertise continuum) cannot be predicted just on the basis of such patterns. Furthermore, EFL learners may be particularly weak in lower-level writing issues such as grammar and word choice, but be at similar levels in writing performance on higher level writing aspects such as coherence and supporting evidence. However, writing in a second language may introduce additional cognitive efforts and complexities that impact reviewing of high-level as well as low-level writing issues. Finally, EFL learners also exist along a continuum. A fourth-year student may be at a much more advanced level than a first-year student. In general, no studies have examined peer assessment validity in learners at different points along a developmental pathway.

Specifically, this study will address the following research questions:

1. Does the validity and reliability of peer assessment differ across years of university studies?
2. Do these patterns differ by high and lower-level aspects of writing?

These general research questions are examined in an EFL context where there may be especially weak initial performance and especially large changes across the years of university studies.

Methods

Participants

Data were collected from 118 English major students (91% female) in two undergraduate courses in a research-intensive university in Northeast China: a first-year course called English grammar and writing ($n = 48$, 88% females), and a fourth-year course called English academic writing ($n = 70$, 93% females). These two courses are required for English majors in this university and were given by the same writing instructor, who had five years of teaching writing experience and an educational background in EFL writing.

Materials

Assignments

In both courses, participants had to complete writing and online peer review assignments. These tasks accounted for 40% of course grades, and the writing grades were largely based upon peer assessments. Reflecting differences in proficiency levels that naturally come from three intervening years of studying English, the course contents naturally vary in genres, rubrics, and the required length of the writing assignment. For example, 1st years were only required to write a 300-word essay on an open topic ("Should I_____?"), while 4th years wrote a 1,000-word essay on an open topic ("A literature review on _____"; see Appendix A for details).

Peerceptiv

Distinguished from face-to-face peer discussions of drafts, Peerceptiv is a web-based peer assessment application (https://go.peerceptiv.com/). Similar to a number of online peer assessments systems (Paré and Joordens 2008), students are given opportunities to interact with other peers in anonymous fashion as both authors and reviewers. The three core features of the Peerceptiv system are: (a) multiple peer review, (b) anonymous peer review, and (c) accountability mechanisms for accurate ratings and constructive comments. Papers are dynamically assigned to students. Each student is asked to complete a required number of reviews (in this case, four). Peerceptiv assigns one randomly-selected document to a student for review at a time, giving another new document to review as each review is completed, until the required number of reviews are completed. This dynamic assignment process allows students to submit late documents and still receive reviews. In addition, students are held accountable for: 1) accurate ratings through grades given for ratings that are generally consistent with what other peers gave to the same papers; and 2) detailed, constructive comments through ratings given by authors regarding whether they found the comments to be helpful (Gao, Schunn, and Yu 2019).

Writing rubrics

Papers were evaluated by peers (during the semester) and by instructors (afterwards for the research study) on multiple shared rubrics using 7-point scales. The rubrics involved different dimensions at a detailed level depending on the course level, as would normally be the case for courses at such different years of study. However, both sets of rubrics contained dimensions that focused on having a main claim, justifying the claim, and attending to low-level issues.

The 1st year course had three dimensions focused on high-level aspects of writing (unity, support, coherence) and one focused on low level aspects of writing (wording and sentence skills) (Langan 2013). The *unity* dimension was concerned with whether students advanced a single point and stuck to that point in the paper. The *support* dimension was concerned with whether the paper supported the central point with the specific evidence, and whether the specific evidence was relevant, typical and adequate. The *coherence* dimension addressed whether the paper organized and connected the specific evidence smoothly. The *wording and sentence skill* dimension examined the effective use of specific words, active verbs, and avoiding slang, cliché, pretentious language and wordiness.

The 4th year course had also three dimensions focused on high-level aspects of writing (introduction, analyzing and synthesizing the literature domestic and abroad, concluding literature review), and one focused on the low level aspects of writing (APA style in-text citation and reference conventions). The *introduction* dimension was concerned with how well the literature review paper presented the central topic and established its academic importance. The *analyzing and synthesizing the literature domestic and abroad* dimension evaluated whether the paper addresses all the key words and issues in the central topic, whether the paper reviewed references from research domestic and abroad, whether the cited literature was relevant to the writer's research focus, whether the paper clarified the evolution in the literature of the central issues, and, finally, whether the paper critiqued the argument, research design, methodology or conclusions of prior work. The *concluding literature review* dimension was concerned with whether the paper addresses a need for the current study and clarifies a point of departure from previous studies. The last dimension was *APA style*, which focused on low levels of writing like in-text citation convention and reference convention. The full rubrics are presented in Appendix B.

Measures

Rating reliability

The ratings across peers were analyzed for inter-rater reliability on each dimension using an ICC. The aggregate, consistency ICC was used, which is similar to a Cronbach alpha. It measures the reliability of the mean rating produced across student raters, focused on the consistency in ordering of ratings (i.e. was their consistency in the relative strengths and weaknesses of the set of documents) rather than the exact agreement. Reliabilities above .7 are considered good (similar to trained expert reliability for evaluation of writing), and reliabilities below .4 are considered poor (high likelihood of substantially different scores if evaluated by a different set of peers). Rating reliability for expert ratings was also calculated using this ICC measure.

Rating validity

To test the validity of the student ratings, four instructor raters rated two batches of papers, with two rating the 1st year papers and another two rating the 4th year papers. The validity of the peer rating means on each dimension was measured as the Pearson correlation between the mean peer score and the mean instructor score. In writing, instructor ratings often correlate with one another at only around r=.3 (Cho, Schunn, and Wilson 2006). Thus, correlations of mean ratings across student raters with instructor ratings of .3 are acceptable and ones above .5 are excellent.

Perceived ability

To learn about student's changing self-perceptions across 1st and 4th year students, students were given a survey involving 5-point Likert ratings. The survey included one question about their perceived English writing ability (As a writer, please review your writing ability), and one question about their perceived reviewing ability (As a reviewer, please rate your reviewing ability). A mean value for each scale was computed.

Procedure

Students submitted a first draft to Peerceptiv, then provided ratings and comments for four peers' documents via Peerceptiv, and finally submitted the second draft. Each phase (submission, reviewing, revision) involved approximately one week.

After the second draft was submitted, students were invited to complete the survey anonymously, but in class to increase the response rates. Time to complete the survey was typically 5 min. Excluding the incomplete responses, there were 108 valid surveys producing a high return rate of 92%.

Analysis

Statistical differences in peer versus instructor correlations and student inter-rater ICC values (e.g. between 1st and 4th year data, high-level versus low-level writing dimensions) were assessed using a Fischer r-to-z transformation. T-tests were used to compare 1st and 4th year students' perceived writing and reviewing abilities.

Results

The reliability of the teachers' ratings (leftmost group in Figure 1) was very similar and at acceptable levels across 1st and 4th year student data, overall and by sub-dimension. Although the research questions were not focused on teacher reliability, this data is a critical assumption test regarding the key contrasts across dimensions and courses: the quality of 1st and 4th year students' documents were equally easy to discriminate, at least in terms of the rubric dimensions applied to each set of documents. Similarly, the high and low-level dimensions were equally easy to discriminate. Thus, contrasting validity and reliability in students' ratings across the courses is a fair assessment of their relative abilities, rather than inherently confounded by rubrics that were much harder or easier to apply to the given documents.

Just as importantly, the moderately high teacher reliabilities establish that this assessment task was far from trivial. However, as shown in the middle group of Figure 1, the reliabilities of the mean student judgments were about the same or only slightly lower than those of the teachers. Thus, while the reliability of the student ratings are not perfect, they are no less reliable than teacher ratings, an important pragmatic point for instruction.

Although there was a trend for the 4th year ratings to have lower reliability than 1st year ratings for overall and high-level writing dimensions (middle group in Figure 1), none of these observed differences were statistically significant ($ps > .2$). The lower reliability for high-level writing ratings than for the low-level writing rating was statistically significant in the 4th year data ($z = 2.04$, $p < .05$).

Interestingly, a different pattern was observed in the validity of the student ratings (right group in Figure 1). The 4th year ratings had much higher validity overall ($z = 3.29$, $p < .01$) and for the low-level writing dimensions ($z = 3.52$, $p < .01$). The trend was for greater validity in higher level writing dimensions as well, but this difference was not statistically significant ($z = 1.11$, $p > .25$). This pattern was primarily driven by variation by dimension within the 1st years: having lower validity for the low-level dimension than for higher-level writing dimensions. Most saliently, the variation in validity was not driven by variation in reliability. That is, the near zero validity of the low-level ratings in the 1st years' assessments was not because of low reliability of these ratings. To put it simply, for assessments of low-level writing, 1st year raters align well with each other but not with their teachers' assessments: the 1st year students appear to be systematically (rather than randomly) missing certain kinds of low-level writing competence.

These measures of reviewing ability can be contrasted with students' perception of their writing and reviewing skills as measured by the survey. Analyses of the survey results showed that both 1st years and 4th years had similar views of their own reviewing abilities (means of 3.4 and 3.5

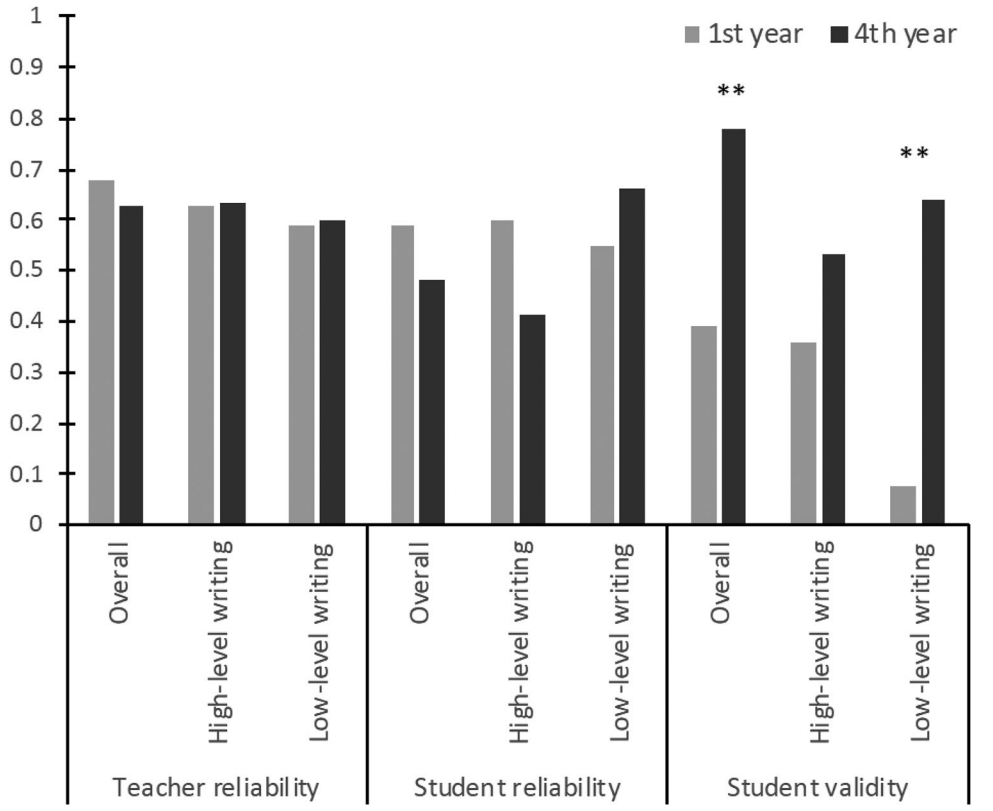


Figure 1. Reliability (teacher and student ratings) and validity (mean teacher versus mean student) for 1st year and 4th year students, for overall ratings, high-level writing ratings, and low-level writing ratings. ** $p<.01$ for 1st versus 4th year contrasts.

respectively on the five-point Likert scale from "absolutely unable to review" to "highly able to review"). That is to say, they generally rated themselves as having moderate levels of reviewing skills. A very similar pattern was found in the ratings of writing ability (i.e. no significant change, with means of 3.2 and 3.4). Self-efficacy ratings often showed little growth over experience because the experiences not only lead to improvements in skill but also change students' benchmarks for what being strong entails (Pratt, McGuigan, and Katzev 2000; Drennan and Hyde 2008).

General discussion

Previous research has systematically documented the benefits of peer assessment for both first and second language (L1 and L2) student writers (Liu and Hansen 2002, Zhao 2010; Gao, Schunn, and Yu 2019). Further, validity and reliability of peer assessments has generally been found to be high. However, they have not been found to be universally high (Li et al. 2016), and it is therefore important for instructors to know when peer assessment validity is likely to be acceptably high and when students will need extra supports in order to produce valid and reliable assessments. The current study examined the effects of growth in expertise, contrasting 1st and 4th year EFL students. The small amount of prior work with EFL populations suggested that 1st year students might have less reliable and valid ratings, but no direct contrast within one context had been previously made. The current study did reveal a developmental difference but in a very specific way: within the validity of student assessments and only for lower-level dimensions of writing. Otherwise, reliability across groups was generally high and validity was generally high for higher-level writing dimensions.

Importantly, EFL learners appear to be similar to L1 student writers in terms of peer assessment validity and reliability (Cho, Schunn, and Wilson 2006, Paré and Joordens 2008). That EFL peer assessment validity correlations were generally similarly high, at least for higher level aspects of writing, in comparison with L1 groups, illustrates that peer assessment effectiveness is probably not centrally dependent upon students' L2 English skills (Haaga 1993; Paré and Joordens 2008; Schunn, Godley, and DeMartino 2016). At the same time, the study illustrates that the low-level issue, which is most directly connected to L2 skill, can cause validity problems for peer assessment.

The research finding like the current ones may lessen the worries writing instructors might have about EFL peer reviewing (e.g. afraid to integrate peer assessment into final grading). It is interesting to note that in the current study students were generally willing to take active roles as assessor: there was a 100% submission rate of reviews. This finding is contrasted with Zou et al. (2018) who used the same system for peer assessment with Chinese EFL learners, but found that students were often unwilling to participate in peer assessment. A possible explanation is that this study focuses on predominantly-female English major student population, unlike the more male-dominated engineering population examined in Zou et al. (2018)'s research. In any case, this variation suggests some replication of finding is likely required, particularly including populations more reticent to engage in peer assessment.

What might explain the patterns of developmental differences and non-differences found in this study? The differences might be attributed to task differences: the 4th year students' literature review was more challenging compared with the 1st year student's short essay. However, such task complexity variation is inherently part of an educational pathway; to examine the validity of 4th year's assessments of the simple documents typically assigned to 1st years would have little instructional applicability. In coursework, tasks given to more advanced students are inherently more complex than tasks given to less advanced students. Thus, the instructionally-relevant question is whether students at different points along a pathway are able to meaningfully participate in peer assessments of typically assigned writing tasks. The current study found that both 1st and 4th year students were able to reliably and validly assess the kinds of documents that are typically assigned, at least in terms of high-level writing dimensions.

Another possibility might involve variability with each population: 1st year students may have more uniformity immediately after college entrance examinations given the relatively selective university, whereas 4th years have more diversity after three consecutive years of self-sustained learning. Reliability and validity of judgments are usually higher when the variability in quality is larger; it is easier to make consistent judgements about large quality differences than small quality differences. However, the instructor reliabilities were similarly high across dimensions and years, arguing against such an explanation.

The final possibility involves the situation of learners at the early stage of learning a second language. Human languages involve a very large collection of relatively arbitrary and complex rules and exceptions to be mastered, particularly when the languages have little in common, as in the case of Chinese and English. In this specific case, early on in the language learning process, peer assessments of language conventions might have systematically lower validity. By contrast simpler sets of conventions, like APA conventions for writing, can be more validly judged by peers. Future research is needed to further explore the underlying causes of the patterns revealed in the current study.

Conclusion

Using the case of EFL students in China, the current study examined developmental changes in peer assessment reliability and validity. When using a carefully structured peer assessment process, reliability was generally high across years and writing dimensions, even in this challenging

EFL context. Thus, it is likely that peer assessment in non-EFL contexts will also show high reliability across levels, at least when the peer assessment process is also well structured.

At the same time, the study also uncovered some important variations by writing dimension. 1st year students showed low validity in their assessments of writing conventions like wording and sentence skills. These findings point out specific areas which need more instructional scaffolding and support.

Implications for instructors using peer assessment

Overall the findings illustrate the general asymmetry between reliability and validity. When validity is high, reliability is necessarily high, but when reliability is high, validity is not necessarily high. Therefore, when instructors see low reliability scores, they should be concerned. But when they see high reliability scores, some spot checking is likely necessary to make sure the validity is also high.

Establishing the "for when" and "for whom" of peer assessment validity is important for understanding how and when to integrate peer assessment into formative and summative assessments. Combined with the obvious benefits of receiving immediate, frequent and large amounts of peer feedback, demonstrating its reliability and validity further strengthens the value of peer assessment as an important teaching and learning strategy, whether as a process of Assessment for Learning or Assessment as Learning. Having acceptable reliability and validity levels is likely an important goal in the context of teaching and learning academic writing. Further, various authors have spoken to the importance of empowering students by giving them more opportunities to assess as a peer and participate in the assessment process (Gunersel et al. 2008; Li, Liu, and Zhou 2012; Liu and Lee 2013).

The low-level convention dimension fell well under .4 for validity correlations for 1st year EFL students. This result is likely concerning for instructors, since this is pedagogically important for writing instructors, especially in relation to the goals for students at this stage. It may be that instructors need to illustrate those review dimensions by: 1) giving more model writing samples; 2) using separate and more-detailed reviewing prompts for specific aspects of low-level conventions; or 3) focusing on only a subset of low-level writing conventions in any given writing assignment. Since the reliability was relatively high in this case, it is likely the case that students are systematically missing particular kinds of errors rather than not understanding how the dimension is being defined. An instructor could systematically investigate mismatch cases (e.g. students rated the document with a high score, but they rated it with a low score) to determine which more specific kinds of problems were not noticed.

Limitations and future work

Several limitations need to be addressed in future work. In the current study, peer assessment validity and reliability are explored in number of specific ways within the much larger context in which peer assessment occurs. First, it was examined only in first drafts. In later drafts, student expertise could develop and familiarity with peer assessment could increase. And yet, later drafts might have reduced variability across documents and thus reliability could decrease.

Second, these students were all English majors, and the findings may not extend to non-English majors. For example, such students are likely stronger in English, in addition to having higher interest levels regarding learning English, which may increase the effort levels and hence performance levels obtained during peer assessment. Similarly, these students were from a relatively selective university. Students at less selective universities are likely to have lower English abilities and perhaps other kinds of academic skill and motivation differences which could impact peer assessment performance.

Third, these students came from the Chinese context. Students from other (non-English speaking) countries might vary in terms of typical levels of familiarity with English, linguistic distance between their native language and English, and level of familiarity with peer assessment (and other student-centered teaching methods) during high school instruction.

Finally, future work should also examine the potentially moderating role of rating rubrics; it maybe be that 4th year EFL students have low validity with poorly defined evaluation rubrics and 1st year EFL students have consistently high validity with systematically developed rubrics. Other moderating factors might also include amount of prior training on the skills of reviewing or on the peer assessment process itself.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Work on this project was funded by grants No. JJKH20180054SK, JJKH20190295SK, and JJKH20190306SK from the Education Department of Jilin Province.

Notes on contributors

Fuhui Zhang, Ph.D., is a professor working with School of Foreign Languages, Northeast Normal University, China. She has published on writing and peer assessment in SSCI journal Instructional Science and CSSCI journals like Foreign Language Education. Her enduring research interest lies in how writing and peer assessment help improve students' writing and thinking skills. She has visited Learning Research & Development Center (LRDC) at the University of Pittsburgh for a year.

Christian Dieter Schunn, Ph.D., is a professor and senior cognitive scientist working with Learning Research & Development Center (LRDC) at the University of Pittsburgh. His research interest extends to a wide range of cognitive studies involving STEM reasoning and learning, web-based peer interaction and instruction, neuroscience of complex learning and engagement and learning. He is the founder of an online peer assessment system (Peerceptiv), which is widely used in the US, Canada, China, and some European countries.

Wentao Li is a student in School of Foreign Languages, Northeast Normal University, China. His research interests include second language writing, peer assessment, and self-regulated learning. He has been doing research on second language writing and peer assessment for three years under the supervision of Professor Fuhui Zhang. He has visited Department of Linguistics at Southern Illinois University (SIU) for a year.

Miyin Long is a student in School of Foreign Languages, Northeast Normal University, China. Her research interests include English teaching and learning. She has been doing research on second language writing and peer assessment for three years under the supervision of Professor Fuhui Zhang.

ORCID

Fuhui Zhang http://orcid.org/0000-0003-3494-2162
Christian Schunn http://orcid.org/0000-0003-3589-297X
Wentao Li http://orcid.org/0000-0001-7472-3517
Miyin Long http://orcid.org/0000-0002-9183-5809

References

Azarnoosh, M. 2013. "Peer Assessment in an EFL Context: Attitudes and Friendship Bias." *Language Testing in Asia* 3 (1): 1–10. doi:10.1186/2229-0443-3-11.

Bai, L. 2013. "The Feasibility and Validity of Adopting Peer Revision in English Writing Process." *Journal of PLA University of Foreign Languages* 1: 51–56.

Berg, B. C. 1999. "The Effects of Trained Peer Response on ESL Students' Revision Types and Writing Quality." *Journal of Second Language Writing* 8 (3): 215–241. doi:10.1016/S1060-3743(99)80115-5.

Boud, D. 1989. "The Role of Self-Assessment in Student Grading, Assessment and Evaluation." *Assessment & Evaluation in Higher Education* 14: 20–30. doi:10.1080/0260293890140103.

Cheng, K.-H., H. T. Hou, and S. Y. Wu. 2014. "Exploring Students' Emotional Responses and Participation in an Online Peer Assessment Activity: A Case Study." *Interactive Learning Environments* 22 (3): 271–287. doi:10.1080/10494820.2011.649766.

Cheng, W., and M. Warren. 1997. "Having Second Thoughts: Students Perceptions before and after a Peer Assessment Exercise." *Studies in Higher Education* 22 (2): 233–239. doi:10.1080/03075079712331381064.

Cho, K., and C. MacArthur. 2011. "Learning by Reviewing." *Journal of Educational Psychology* 103 (1): 73–84.

Cho, K., D. C. Schunn, and R. Wilson. 2006. "Validity and Reliability of Scaffolded Peer Assessment of Writing from Instructor and Student Perspectives." *Journal of Educational Psychology* 98 (4): 891–901. doi:10.1037/0022-0663.98.4.891.

Drennan, J., and A. Hyde. 2008. "Controlling Response Shift Bias: The Use of the Retrospective Pre-Test Design in the Evaluation of a Master's Programme." *Assessment & Evaluation in Higher Education* 33 (6): 699–709. doi:10.1080/02602930701773026.

Falchikov, N. 1986. "Product Comparisons and Process Benefits of Collaborative Self and Peer Group Assessments." *Assessment & Evaluation in Higher Education* 11 (2): 146–166. doi:10.1080/0260293860110206.

Falchikov, N., and J. Goldfinch. 2000. "Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks." *Review of Educational Research* 70 (3): 287–322. doi:10.3102/00346543070003287.

Gao, Y., D. C. Schunn, and Q. Yu. 2019. "The Alignment of Written Peer Feedback with Draft Problems and Its Impact on Revision in Peer Assessment." *Assessment & Evaluation in Higher Education* 44 (2): 294–308. doi:10.1080/02602938.2018.1499075.

Greenberg, K. P. 2015. "Rubric Use in Formative Assessment: A Detailed Behavioral Rubric Helps Students Improve Their Scientific Writing Skills." *Teaching of Psychology* 42 (3): 211–217. doi:10.1177/0098628315587618.

Gunersel, A. B., N. J. Simpson, K. J. Aufderheide, and L. Wang. 2008. "Effectiveness of Calibrated Peer Review for Improving Writing and Critical Thinking Skills in Biology Undergraduate Students." *Journal of the Scholarship of Teaching and Learning* 8 (2): 25–37.

Haaga, D. A. F. 1993. "Peer Review of Term Papers in Graduate Psychology Courses." *Teaching of Psychology* 20 (1): 28–32. doi:10.1207/s15328023top2001_5.

Han, C. 2018. "A Longitudinal Quantitative Investigation into the Concurrent Validity of Self and Peer Assessment Applied to English-Chinese Bi-Directional Interpretation in an Undergraduate Interpreting Course." *Studies in Educational Evaluation* 58: 187–196. doi:10.1016/j.stueduc.2018.01.001.

Huisman, B., W. Admiraal, O. Pilli, M. van de Ven, and N. Saab. 2018. "Peer Assessment in MOOCs: The Relationship between Peer Reviewers' Ability and Authors' Essay Performance." *British Journal of Educational Technology* 49 (1): 101–110. doi:10.1111/bjet.12520.

Langan, J. 2013. *College Writing Skills with Readings.* New York, NY: Tata McGraw-Hill Education.

Li, H., Y. Xiong, X. L. Zang, M. Kornhaber, Y. Lyu, K. S. Chung, and H. K. Suen. 2016. "Peer Assessment in the Digital Age: A Meta-Analysis Comparing Peer and Teacher Ratings." *Assessment & Evaluation in Higher Education* 41 (2): 245–264.

Li, L., X. Y. Liu, and Y. C. Zhou. 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. doi:10.1111/j.1467-8535.2011.01180.x.

Liu, E. Z. F., and C. Y. Lee. 2013. "Using Peer Feedback to Improve Learning via Online Peer Assessment." *The Turkish Online Journal of Educational Technology* 12 (1): 187–199.

Liu, J., and J. G. Hansen. 2002. *Peer Response in Second Language Writing Classrooms*. Ann Arbor: The University of Michigan Press.

Liu, L., and X. Ji. 2018. "A Study on the Acceptability and Validity of Peer Scoring in Chinese University EFL Writing Classrooms." *Foreign Language World* 5: 63–70.

Lu, J. Y., and N. Law. 2012. "Online Peer Assessment: Effects of Cognitive and Affective Feedback." *Instructional Science* 40 (2): 257–275. doi:10.1007/s11251-011-9177-2.

Lynch, D. H., and S. Golen. 1992. "Peer Evaluation of Writing in Business Communication Classes." *Journal of Education for Business* 68 (1): 44–48. doi:10.1080/08832323.1992.10117585.

Marcoulides, G., and M. Simkin. 1995. "The Consistency of Peer Review in Student Writing Projects." *Journal of Education for Business* 70 (4): 220–223. doi:10.1080/08832323.1995.10117753.

Min, H. T. 2006. "The Effects of Trained Peer Review on EFL Students' Revision Types and Writing Quality." *Journal of Second Language Writing* 15 (2): 118–141. doi:10.1016/j.jslw.2006.01.003.

Mowl, G., and R. Pain. 1995. "Using Self and Peer Assessment to Improve Students' Essay Writing: A Case Study from Geography." *Innovations in Education and Training International* 32 (4): 324–335. doi:10.1080/1355800950320404.

Panadero, E., M. Romero, and J. Strijbos. 2013. "The Impact of a Rubric and Friendship on Peer Assessment: Effects on Construct Validity, Performance, and Perceptions of Fairness and Comfort." *Studies in Educational Evaluation* 39 (4): 195–203. doi:10.1016/j.stueduc.2013.10.005.

Paré, D. E., and S. Joordens. 2008. "Peering into Large Lectures: Examining Peer and Expert Mark Agreement Using PeerScholar, an Online Peer Assessment Tool." *Journal of Computer Assisted Learning* 24 (6): 526–540. doi:10.1111/j.1365-2729.2008.00290.x.

Patchan, M. M., and D. C. Schunn. 2015. "Understanding the Benefits of Providing Peer Feedback: How Students Respond to Peers' Texts of Varying Quality." *Instructional Science* 43 (5): 591–614. doi:10.1007/s11251-015-9353-x.

Pratt, C. C., W. M. McGuigan, and A. R. Katzev. 2000. "Measuring Program Outcomes: Using Retrospective Pretest Methodology." *American Journal of Evaluation* 21 (3): 341–349. doi:10.1177/109821400002100305.

Rushton, C., P. Ramsey, and R. Rada. 1993. "Peer Assessment in a Collaborative Hypermedia Environment: A Case Study." *Journal of Computer-Based Instruction* 20 (3): 73–80.

Russell, J., S. Van Horne, A. S. Ward, E. A. Bettis, and J. Gikonyo. 2017. "Variability in Students' Evaluating Processes in Peer Assessment with Calibrated Peer Review." *Journal of Computer Assisted Learning* 33 (2): 178–190. doi:10.1111/jcal.12176.

Sadler, P. M., and E. Good. 2006. "The Impact of Self- and Peer-Grading on Student Learning." *Educational Assessment* 11 (1): 1–31.

Saito, H., and T. Fujita. 2004. "Characteristics and User Acceptance of Peer Rating in EFL Writing Classrooms." *Language Teaching Research* 8 (1): 31–54. doi:10.1191/1362168804lr133oa.

Schunn, D. C., A. Godley, and S. DeMartino. 2016. "The Reliability and Validity of Peer Review of Writing in High School AP English Classes." *Journal of Adolescent & Adult Literacy* 60 (1): 13–23. doi:10.1002/jaal.525.

Shao, M. 2009. "A Study of Peer Assessment in EFL Instruction." *Foreign Language Learning and Practice* 2: 47–53.

Shrout, P. E., and J. L. Fleiss. 1979. "Intraclass Correlations: Uses in Assessing Rater Reliability." *Psychological Bulletin* 86 (2): 420–428. doi:10.1037/0033-2909.86.2.420.

Stefani, L. A. J. 1994. "Peer, Self and Tutor Assessment: Relative Reliabilities." *Studies in Higher Education* 19 (1): 69–75. doi:10.1080/03075079412331382153.

Stemler, S. E. 2004. "A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability." *Practical Assessment, Research & Evaluation* 9 (4): 1–11. http://PAREonline.net/getvn.asp?v=9&n=4.

Topping, K. 2010. "Methodological Quandaries in Studying Process and Outcomes in Peer Assessment." *Learning and Instruction* 20 (4): 339–343. doi:10.1016/j.learninstruc.2009.08.003.

Ward, M., L. Gruppen, and G. Regehr. 2002. "Measuring Self-Assessment: Current State of the Art." *Advances in Health Sciences Education* 7 (1): 63–80. doi:10.1023/A:1014585522084.

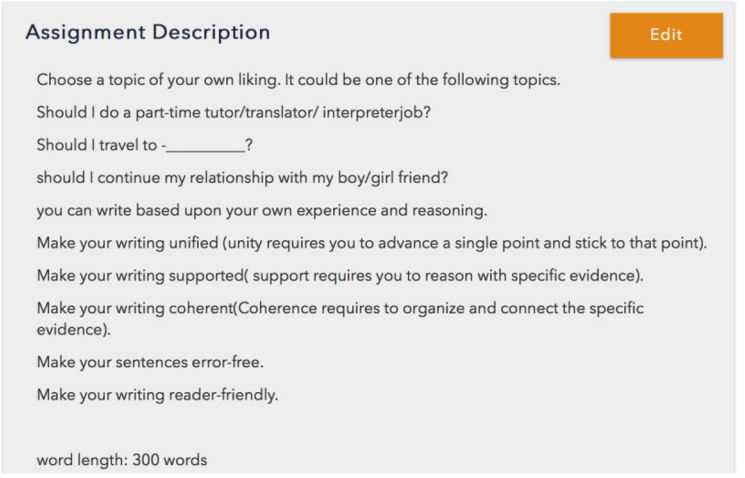
Weigle, S. C. 1994. "Using Facets to Model Rater Training Effects." *Language Testing* 15 (2): 40. doi:10.1177/026553229801500205.

Zhao, H. 2010. "Investigating Learners' Use and Understanding of Peer and Teacher Feedback on Writing: A Comparative Study in a Chinese English Writing Classroom." *Assessing Writing* 15 (1): 3–17. doi:10.1016/j.asw.2010.01.002.

Zou, Y., D. C. Schunn, Y. Wang, and F. Zhang. 2018. "Student Attitudes That Predict Participation in Peer Assessment." *Assessment & Evaluation in Higher Education* 42: 1–12. doi:10.1080/02602938.2017.1409872.

Appendix A

Writing tasks for the 1st year (top) and 4th year (bottom)



Assignment Description

Edit

Choose a topic of your own liking. It could be one of the following topics.

Should I do a part-time tutor/translator/ interpreterjob?

Should I travel to -_____?

should I continue my relationship with my boy/girl friend?

you can write based upon your own experience and reasoning.

Make your writing unified (unity requires you to advance a single point and stick to that point).

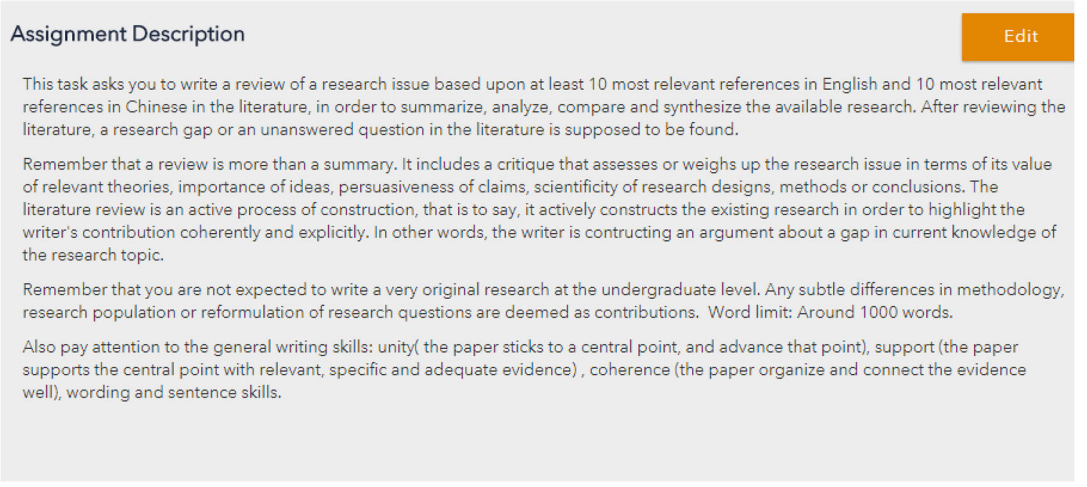
Make your writing supported(support requires you to reason with specific evidence).

Make your writing coherent(Coherence requires to organize and connect the specific evidence).

Make your sentences error-free.

Make your writing reader-friendly.

word length: 300 words



Assignment Description

Edit

This task asks you to write a review of a research issue based upon at least 10 most relevant references in English and 10 most relevant references in Chinese in the literature, in order to summarize, analyze, compare and synthesize the available research. After reviewing the literature, a research gap or an unanswered question in the literature is supposed to be found.

Remember that a review is more than a summary. It includes a critique that assesses or weighs up the research issue in terms of its value of relevant theories, importance of ideas, persuasiveness of claims, scientificity of research designs, methods or conclusions. The literature review is an active process of construction, that is to say, it actively constructs the existing research in order to highlight the writer's contribution coherently and explicitly. In other words, the writer is contructing an argument about a gap in current knowledge of the research topic.

Remember that you are not expected to write a very original research at the undergraduate level. Any subtle differences in methodology, research population or reformulation of research questions are deemed as contributions. Word limit: Around 1000 words.

Also pay attention to the general writing skills: unity(the paper sticks to a central point, and advance that point), support (the paper supports the central point with relevant, specific and adequate evidence) , coherence (the paper organize and connect the evidence well), wording and sentence skills.

Appendix B. Peer assessment rubrics

1st years' anchor points for each of the four evaluation dimensions

Dimension	Rating	Rubrics
Unity	7	Very clear central idea, with clear sub-ideas, and very convincing.
	6	between 7 and 5
	5	Very clear central idea, with rather clear supporting ideas, but not convincing enough.
	4	between 5 and 3
	3	Very clear central idea, not with clear supporting ideas.
	1	No central idea at all.
Support	7	The supportive evidence is very specific, typical and adequate.
	6	between 7 and 5
	5	The supportive evidence is specific, typical and adequate.
	4	between 5 and 3
	3	The supportive evidence is not very specific, typical or adequate.
	1	Little supportive evidence.
Coherence		
	7	The supporting ideas and sentences connect very well, with good transitions and a clear order.
	6	between 7 and 5
	5	The supporting ideas and sentences connect with each other, though the sentence order is not very clear.
	4	between 5 and 3
	3	The supporting ideas and sentences connect loosely, with a few disordered sentences and ideas.
	2	The supporting ideas and sentences do not connect.
	1	no flow
Wording and sentence skills		
	7	The wording and sentences are used very effectively.
	6	between 7 and 5
	5	The wording and sentences are used effectively.
	4	between 5 and 3
	3	Most of the wording and sentences are used correctly.
	1	The wording and sentences are used incorrectly.

4th years' anchor points for each of the four evaluation dimensions

Dimension	Rating	Rubrics
Introduction	7	Introducing the central topic, importance of central topic, necessity of further investigation very clearly and precisely, in a highly well-understood manner.
	6	Introducing the central topic, importance of central topic, necessity of further investigation clearly and precisely, in a well- understood manner.
	5	Introducing the central topic, importance of central topic, necessity of further investigation clearly in an easy-to-follow manner.
	4	Introducing the central topic, importance of central topic, necessity of further investigation clearly.
	3	Introducing the central topic, importance of central topic, necessity of further investigation.
	2	Introducing the central topic, importance of central topic, necessity of further investigation poorly.
	1	Difficult to read at all.
Analyzing and synthesizing the literature domestic and abroad	7	Present the literature in a highly critical and objective manner, with clear focus and problem evolvement.
	6	Present the literature in a critical and objective manner, with clear focus and problem evolvement.
	5	Present the literature with clear focus and problem evolvement.
	4	Present the literature with clear focus.
	3	Present the literature by comparison and contrast.
	2	List the literature with no clear focus.
	1	Difficult to read at all.
Concluding literature review	7	End the literature review with an identified gap, very operational and manageable for further research.
	6	End the literature review with an identified gap, operational and manageable for further research.
	5	End the literature with an identified gap or point of departure in a logical way.
	4	End the literature with an identified gap or point of departure.
	3	End the literature review with an identified gap.
	2	End the literature review.
	1	Difficult to read at all.
APA style	7	Excellent
	6	Very Good
	5	Good
	4	Ok
	3	Almost Done
	2	Half-done
	1	Failure to do at all