Received: 9 December 2022 | Revised: 12 July 2023 | Accepted: 24 July 2023

DOI: 10.1111/jcal.12861

ORIGINAL ARTICLE

Journal of Computer Assisted Learning WILEY

What predicts variation in reliability and validity of online peer assessment? A large-scale cross-context study

Yao Xiong[1] | Christian D. Schunn[2] | Yong Wu[3]

[1]Roblox Corporation, San Mateo, California, USA

[2]Learning Research and Development Center, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

[3]School of Humanities, Beijing University of Posts and Telecommunications, Beijing, China

Correspondence
Yong Wu, School of Humanities, Beijing University of Posts and Telecommunications, Beijing 100876, China.
Email: wuyong@bupt.edu.cn

Abstract

Background: For peer assessment, reliability (i.e., consistency in ratings across peers) and validity (i.e., consistency of peer ratings with instructors or experts) are frequently examined in the research literature to address a central concern of instructors and students. Although the average levels are generally promising, both reliability and validity can vary substantially from context to context. Meta-analyses have identified a few moderators that are related to peer assessment reliability/validity, but they have lacked statistical power to systematically investigate many moderators or disentangle correlated moderators.

Objectives: The current study fills this gap by addressing what variables influence peer assessment reliability/validity using a large-scale, cross-context dataset from a shared online peer assessment platform.

Methods: Using multi-level structural equation models, we examined three categories of variables: (1) variables related to the context of peer assessment; (2) variables related to the peer assessment task itself; and (3) variables related to rating rubrics of peer assessment.

Results and Conclusions: We found that the extent to which assessment documents varied in quality on the given rubric played a central role in mediating the effect from different predictors to peer assessment reliability/validity. Other variables that are significantly associated with reliability and validity included: Education Level, Language, Discipline, Average Ability of Peer Raters, Draft Number, Assignment Number, Class Size, Average Number of Raters, and Length of Rubric Description. The results provide information to guide practitioners on how to improve reliability and validity of peer assessments.

KEYWORDS

peer assessment, reliability, validity

1 | INTRODUCTION

Peer assessment is an educational activity where students assess the quality of work by other students of approximately equal status. It has become increasingly common in performance assessments (i.e., opposed to traditional assessments such as assessments with mostly multiple-choice questions) where students are required to complete more open-ended tasks (e.g., essays, papers, or projects) and human judgement is essential for providing feedback and evaluation. Earlier implementation of peer assessment was mainly paper-and-pencil-based in small classrooms (Luxton-Reilly, 2009). The development of online peer assessment systems made it easier to implement peer assessment in large classes with many desirable functionalities embedded, which would not be possible or easy with only a paper-and-pencil mode. In addition, the functions (e.g., anonymity, random assignment, and judgements of comment helpfulness) embedded in online peer assessment systems have changed peer assessment processes and thereby influenced peer assessment results. For example, online peer assessment facilitates random assignment of raters to ratees, anonymity and accountability in assessment, and implementation of different levels of structure in supporting peer assessment (Cho & Schunn, 2007; Gielen & De Wever, 2015; Luxton-Reilly, 2009; Zong et al., 2021). Participation in back-evaluation (i.e., ratees evaluated the quality of feedback they received from raters by rating the helpfulness of the feedback) also improved peer assessment quality (Patchan et al., 2017; Wu & Schunn, 2021). In addition, back-evaluation is included in peer assessment to give raters accountability pressures to provide quality feedback.

Research attention on peer assessment has often focused on peer rating reliability and validity (e.g., Chang et al., 2011; Cho et al., 2006; Falchikov & Goldfinch, 2000; Li et al., 2016). Reliability (e.g., the consistency of evaluations across raters or time) and validity (e.g., the consistency of evaluations with external criteria) are two related but different concepts, although the peer assessment literature sometimes failed to differentiate the two (Topping, 1998). When the quality of peer assessment results is studied as compared to expert assessment (e.g., correlation between peer and expert ratings; e.g., Falchikov & Goldfinch, 2000; Li et al., 2016), it is predominantly a validity issue in which the expert assessments (e.g., instructor assessments) are the external criteria, which is one type of validity evidence as defined in current psychological and educational measurement standards (AERA et al., 2014). By contrast, it is a reliability issue when the quality of peer assessment results is evaluated in terms of inter-rater consistency among peer raters (e.g., Cho et al., 2006). While reliability and validity issues in peer assessment can be much broader than what has been documented in literature, we focus here on these prevailing conceptual and operational definitions of reliability (i.e., the inter-rater consistency among peer raters) and validity (i.e., the consistency between peers and experts).

Online peer assessment reliability and validity have been investigated by calculating the consistency among peer ratings and comparing peer ratings and instructor ratings (e.g., Schunn et al., 2016; Tseng & Tsai, 2007; Zhang et al., 2020). However, little empirical research has been conducted to analyse what may influence online peer assessment reliability and validity except several meta-analyses that investigated variables that may impact peer rating reliability and validity through moderator analyses (e.g., Double et al., 2019; Li et al., 2016; Zheng et al., 2019). A moderator analysis takes a step towards explaining the observed variation in effect sizes of different studies. However, most moderator analyses in meta-analysis studies had power limitations, which resulted in finding relatively few significant moderators (e.g., Li et al., 2016, 2020) and not being able to address potential statistical confounds among different moderators (e.g., Double et al., 2019; Li, Bialo et al., 2021, Li, Zhao et al., 2021; Zheng et al., 2019). To sum up, meta-analyses are excellent from an external validity perspective—distilling common patterns across a wide range of study contexts. However, the known drawbacks of meta-analysis are also related to this context variation: (1) threats to internal validity when synthesising research results from very different contexts, and (2) the challenge of explaining the different results in different studies (Cortina, 2016; Glass, 2000).

In the current study, we address the research question of what factors influence peer assessment reliability and validity using a large-scale, cross-context dataset (i.e., 374 courses from 88 different institutions, representing 19,722 total students) collected from an online peer assessment system. This specific research design has important complementary methodological advantages in comparison with general meta-analysis for investigating variation in peer assessment reliability and validity. Although validity seems to be more important in a general sense due to its connection to assessment result interpretation, reliability is important in a practical sense because the disagreement in peer ratings can confuse assessees and harm their motivation and trust in this pedagogical effective approach (Cho et al., 2006). By setting a common ground for comparison, such as using the same online peer assessment system and using the exact same measures of reliability/validity across different courses, it addresses the problem of incomparability of effect sizes in meta-analyses as well as reducing extraneous context factors that threaten internal validity and reduce analysis power.

Consistent with the typologies used in the literature to characterise the diversity of approaches, applications, and contexts of peer assessment (Gielen et al., 2011; Topping, 1998), we examined different categories of variables that may influence peer assessment reliability and validity (i.e., a form of moderator analysis), including (1) variables related to the context; (2) variables related to the task itself; and (3) variables related to rating rubrics. From (1) to (3), variables ranged from measures of a broader context to those related to more granular procedures. While (1) and (2) have been studied more in literature (e.g., Li et al., 2016), (3) was rarely studied. We used this framework to both align with the existing typologies in peer assessment literature as well as to emphasise the unique contribution of the current study. The variables collected in the current study were extracted directly from the online system rather than being human-coded, which further reduces a source of noise that is common in meta-analyses (i.e., errors in moderator coding). In addition, thanks to the wide use of the system, we collected thousands of instances of reliability/validity measures, which produced much greater power for the moderator analysis, as compared to only dozens to a few hundreds of effect sizes that are typically included in meta-analyses. With those advantages of the current study, we believe the results will shed light on important factors driving peer assessment reliability and validity, which can guide practitioners on when peer assessment is likely to be reliable and how to effectively conduct peer assessment activities, as well as inform theories of performance and learning from peer assessment.

2 | LITERATURE REVIEW

2.1 | Online peer assessment reliability and validity

To be consistent with current educational and psychological measurement standards (AERA et al., 2014), peer assessment reliability is operationally defined in this study as the inter-rater consistency among peer raters while validity is operationally defined as the consistency between peers and experts. Defined this way, the reliability and validity of peer assessment have been found to be, on average, at or above acceptable levels (e.g., Schunn et al., 2016; Tseng & Tsai, 2007; Xiao & Lucking, 2008; Zhang et al., 2020).

Inter-rater reliabilities, in particular, were reported to be medium to high in different online peer assessment studies in both K-12 and higher education settings. For example, Tseng and Tsai (2007) carried out a study of online peer assessment in a high school computer science course. They found that the inter-rater reliability (IRR) of peer ratings (among 10 peers) was between 0.70 and 0.81 (using Cronbach's alpha), and the alpha coefficients were the highest for the second round of peer assessment and the lowest for the third round of peer assessment. In high school AP writing classes, Schunn et al. (2016) examined a large number of secondary school students (1215 students) enrolled in AP writing courses, and reported intraclass correlations (ICCs) above 0.4 among peer raters. Similar patterns were found in higher education settings as well. For example, an online peer assessment study of 16 different science courses in higher education reported an IRR among peer raters that was generally medium to high, ranging from 0.45 to 0.88 (Cho et al., 2006).

Similarly, when examining validity, correlations between peer ratings and expert ratings were reported to be medium to high. In a high school computer course, Tseng and Tsai (2007) reported that the correlations between peer ratings and expert ratings (i.e., validity of online peer grading) were medium to high, ranging from 0.49 to 0.79. In the high school AP writing class, the correlations between instructor ratings and mean student ratings were between 0.4 and 0.6 on different assessment dimensions (i.e., the different aspects that student work was rated on) (Schunn et al., 2016). Strong correlations have also been found between anonymous peer ratings and expert ratings in undergraduate writing classes of English native speakers (e.g., Xiao & Lucking, 2008) and English as a Foreign Language (EFL) learners (e.g., Zhang et al., 2020). By performing an experiment of online peer assessment in higher education, Bouzidi and Jaillet (2009) compared peer ratings to instructor ratings, and found the correlations were high, ranging from 0.88 to 0.91. They claimed that online peer assessment could be trusted when at least four peers rated the same exam paper. A more recent meta-analysis found a high average correlation between peer and instructor ratings of 0.63 (Li et al., 2016). However, not all studies in the meta-analysis reported high levels of validity. The varied levels may be related to how the peer assessment was carried out (Patchan et al., 2017; Schunn et al., 2016).

2.2 | Variables influencing peer assessment reliability and validity

Few studies have formally examined what variables might influence online peer assessment reliability and validity. This section will consider both online peer assessment and traditional peer assessment studies given the scarcity of moderator analyses in peer assessment in general. In this broader literature, a wide variety of variables has been found to be related to peer assessment reliability or validity. The variables are organised in three broad categories: (1) variables related to the context of peer assessment; (2) variables related to the peer assessment task itself; and (3) variables related to rating rubrics. Variables in category 1 are at a larger grainsize and are usually a constant within a course, while variables in category 3 are at more fine grainsize and may vary even within one peer assessment task. Within these categories, we focus on the specific variables that can be studied within the peer assessment system we examined. Some variables could be conceptually meaningful at multiple grainsizes but could not be simultaneously included at each grainsize because of collinearity issues. These variables were included at the finest grainsize because it would capture more information.

2.2.1 | Variables related to the context

There has been growing emphasis on the role of context in influencing peer assessment results (Topping, 1998). Contextual variables are about the broader context under which the peer assessment activity is conducted. This category includes variables related to school/institution and those related to course characteristics. Context-level variables examined in the present study included Education Level, Language, and Discipline.

Education level

Peer assessment has been used in elementary, secondary, undergraduate, and graduate education, but it may be especially common in tertiary education because college students may be more likely to have the skills and expertise needed for effective peer assessment. Research results on peer assessment in high school suggests that students are able to provide reliable and valid assessment results to their peers (Bouzidi & Jaillet, 2009; Sanchez et al., 2017; Schunn et al., 2016; Tseng & Tsai, 2007), particularly when guided by a carefully designed rubric and supportive peer assessment system. Sadler and Good (2006) even found a very high correlation between middle school students' peer evaluations and teachers' evaluations.

Two meta-analyses have found that advanced-level or graduate-level courses had higher correlations with instructor ratings than did introductory or lower-level courses (Falchikov & Goldfinch, 2000; Li et al., 2016). However, with increasing level, there are also likely to be increasingly difficult problems to detect in student work; therefore, from a theoretical perspective, it may be that reliability and validity hold relatively constant across levels. The findings in the meta-analyses may have involved confounds between education level and other characteristics. For example, tertiary education might have been more likely to use online systems that include useful scaffolds for raising reliability. Overall, there is no data directly comparing K-12 students to tertiary students in terms of peer assessment reliability or validity. There are theoretical arguments for and against an effect of education level coupled with limited amount of evidence across many levels.

Language

Another contextual level variable is the language students use for peer assessment. In addition to being a common instructional technique for students in their native language, peer assessment is also widely used in EFL and ESL classrooms (e.g., Birjandi & Hadidi Tamjid, 2012; Gao et al., 2019; Matsuno, 2009; Saito, 2008). However, if peer assessment is conducted in a language that is different from students' native language, it may increase the difficulty of providing informative feedback in this process.

The literature on peer assessment in EFL or ESL classrooms does report some concerns about validity and reliability. For example, Matsuno (2009) reported peer raters in an EFL writing class of a Japanese university were lenient (i.e., tended to rate higher than instructors) in evaluating their peers in general, but also tended to grade high-achieving writers lower and low-achieving writers higher—a central tendency effect. By contrast, Rasouli and Esfandiari (2022) found Iranian EFL students were too severe (i.e., tended to rate lower than instructors) in rating English essays, and advanced peer raters showed more variability in their severity compared with their intermediate counterparts. In EFL classrooms, different patterns have also been observed when reliability and validity were considered on separate dimensions of rubrics, with some dimensions targeting higher-level aspects (i.e., content and structure) while other dimensions targeting lower-level aspects (i.e., grammar and vocabulary). For example, Zhang et al. (2020) found that peer assessment reliability among EFL students was generally high for all assessment dimensions, but validity was much higher for the higher-level aspects (e.g., unity, support, and coherence) than that for the lower-level aspects (e.g., wording and sentence skills).

The existing literature does not suggest a consistent pattern for the relationship between language level and assessment accuracy. Generally, students in EFL or ESL classrooms are often given simpler English writing assignments, so their reliability and validity of peer assessments are not necessarily lower in general, although EFL or ESL students on average have less fluency in English as English native speakers. Indeed, some studies have found good reliability and validity levels, especially with some trainings provided to peer raters (Saito, 2008; Zhang et al., 2020). However, it has not been systematically investigated whether the fluency of the language used for assessment is associated with the reliability or validity of assessment results. We addressed this question in the current study by comparing reliability and validity in EFL/ESL classrooms versus native English student classrooms.

Discipline

Peer assessment has been conducted in many different subject areas, including social science, science, engineering, business, and medical courses. Two meta-analyses showed similar effects of Discipline: peer assessments conducted in medical areas had lower validity in terms of correlations with expert assessment than did peer assessment in all other areas (Falchikov & Goldfinch, 2000; Li et al., 2016). The complexity of the medical subject discipline requires expertise for high-quality evaluation, which made using peer assessment challenging in medical subjects, especially for summative purpose. Leaving out the medical subjects, no significant differences were reported between social science/arts and science/engineering. In the Jeffery et al. (2016) study, subject matter (i.e., hospitality and tourism management, human health and nutritional sciences, molecular and cellular biology) was not a significant factor influencing peer assessment reliability and validity.

However, the results reported in those meta-analyses are syntheses of studies conducted in different contexts and implemented with various peer assessment processes. The non-significant differences might have been influenced by confounds not controlled across the studies. For example, supports offered by the peer assessment system or the tasks (e.g., simple questions vs. essays) in different disciplines to be reviewed might influence peer assessment reliability and validity. Therefore, larger datasets using more closely equivalent metrics and with some commonalities in peer assessment process across disciplines would be better source to examine the discipline effect. Theoretically, a greater emphasis on objective rather than subjective dimensions within assignments in science/engineering could produce greater reliability.

2.2.2 | Variables related to the peer assessment task itself

While contextual variables surround peer assessments, task variables involve the peer assessment task itself: the goals and characteristics of the objects that are being evaluated, as well as the setup of the peer assessment process. Each peer assessment activity is focused on a certain task or assignment that the instructor directs students to complete. This task can take many different formats, such as essay writing, computer programming assignments, projects, and oral presentations. The instructor can also make many choices regarding how the peer assessment process unfolds, such as peer rater training, involving peer raters in the rubric development, written versus verbal feedback, in-text comments versus summary comments, the number of peer raters per document, and author and reviewer anonymity.

Class Size

Class Size refers to the number of students who participate in peer assessment. For example, a large class could include hundreds of students who submit their work to an online peer assessment system and review peers' work in the context of large lecture courses or even massive open online courses (MOOCs, e.g., Piech et al., 2013). A small class might involve only dozens of students engage in peer assessment in a writing class (e.g., Gao et al., 2019). Few studies have been done to investigate the effects of Class Size on peer assessment. Jeffery et al. (2016) did not find correlations between peer assessment reliability/validity and Class Size. In their study, class sizes of three undergraduate courses were divided into small (<30 students), medium (40–60 students), and large (>100 students) class sizes. The results might be different if peer assessment were conducted in more courses at each Class Size. Broader research on Class Size effects found that small classes can have fewer disengaged students because teachers are better able to incentivise or support low-effort students effectively and students are more likely to connect to the class setting (Babcock & Betts, 2009; Dynarski et al., 2013). Similarly, Class Size may negatively influence peer assessment reliability and validity because it would be more difficult to build a shared understanding of the rubrics in a large class in online peer assessment contexts. In addition, students in small class may be more likely to engage with the peer assessment process because of closer connection to other students and the instructors.

Rater training/experience

Raters' familiarity and understanding of the rating rubrics are considered to be important in peer assessment (Saito, 2008). In a recent meta-analysis (Li et al., 2016), rater involvement in the rubric development had a large effect on peer assessment validity. However, rater training in general did not show any effect in the same study. It may be the case that Rater Training needs to be conducted in certain ways to be effective; well-structured rater training with exemplars and practices was shown to be helpful for improving peer assessment results (Jonsson & Svingby, 2007; van Zundert et al., 2010). It is also important to note that training may have other benefits than improving reliability and validity of ratings. For example, Rater Training was also reported to be the single most important factor associated with learning outcomes in peer assessment (Li et al., 2020).

Within one course, a student might experience just one or two peer assessment tasks by attending one or more assignments or completing one or more drafts. Having students provide peer assessments on more occasions promoted consensus among peer raters and produced greater reliability in peer assessments of athletic training (Marty et al., 2010), of computer projects (Tseng & Tsai, 2007), and of essay writing (Xiao & Lucking, 2008). Extensive practice with peer assessment within a course may be particularly useful for improving reliability and validity as a kind of deliberate practice (Ericsson, 2008).

While prior peer assessment experience could be a positive factor in promoting reliability/validity of assessment results, the opposite effect could happen. For example, student performance may tend to have less variability with more practice (e.g., lower-performing students close gaps with higher-performing students with practice), which would reduce the observed reliability and validity when assessed via correlations due to restricted range problems (i.e., attenuated correlations due to a narrower range of observed performances). Consistent with that prediction, negative relationship of raters' prior experience and peer assessment reliability have been observed (e.g., Li et al., 2020; Tseng & Tsai, 2007). Different findings alongside potential confounds of rater experience with restricted range concerns call for more research on this factor. Assignment Number and Draft Number were used to measure Rater Experience in the present study.

Number of raters per document

Mathematically, more accurate inferences come from a larger number of independent observations (i.e., the reliability of the mean across N ratings increases with N). Correspondingly, several studies have found that reliability or validity coefficients increased from 1 to 4 raters (Bouzidi & Jaillet, 2009; Cho & Schunn, 2018; Sung et al., 2010), with equivalent or close to instructor levels somewhere between 2 and 4 (Magin & Helmore, 2001; Schunn et al., 2016; Xiao & Lucking, 2008). However, Cho and Schunn (2018) noted that there were diminishing returns as the number of reviewers increase beyond that point; they speculated that high reviewer workload might result in lower reliability and validity as students rush to complete all the reviewing tasks. A recent meta-analysis (Li et al., 2016) showed there was no difference in peer assessment validity across different numbers of peer raters per document. In this meta-analysis, the number of peer raters was bucketed into three categories: below 6 raters, 6–10 raters, and above 10 raters. To sum up, the Number of Raters seems to impact the peer assessment results positively when the number is relatively small (e.g., <6). However, the effect is likely to disappear above 4 or 5.

2.2.3 | Variables related to rating rubrics

The third set of variables is about rubric-related variables. Rubrics are essential in peer assessment, describing the expectations of each competence level. Rubrics can increase rating consistency (Jonsson & Svingby, 2007). In the literature, rubric-related variables are usually measured at the assignment level since each assignment can have its own rating rubrics. The variables mainly include indicators of rubric appropriateness or quality, such as whether the rating rubric is analytic or holistic, whether the rating rubric fits the targeted student population, and whether the rating rubric is specific and sufficiently comprehensive.

Total number of rubric dimensions

Little research has been done to directly investigate the effect of Total Number of Rubric Dimensions on peer assessment results. However, a more extreme case has been studied in literature—analytic versus holistic rating rubrics. Analytic rating rubrics involved rating each document on multiple dimensions, while holistic rating rubrics involve only one overall evaluation criterion. In the assessment literature, analytic versus holistic rating rubrics are comparable in terms of rating results and reliability (Klein et al., 1998). Given the longer time it requires to develop analytic rubrics and to complete each rating, a single holistic rating rubric may be preferred given time and resource constrains. Interestingly, Falchikov and Goldfinch (2000) reported that peer assessment produced more accurate results when holistic rating (i.e., global judgement) was used than when analytic rating involving multiple individual dimensions was used. This finding is supported by the study by Schunn et al. (2016) who found the correlations between instructor and student ratings were higher for the overall essay rating than for the rubric dimensions. However, a more general review on rating rubrics showed that analytic rubrics led to more reliable ratings (Jonsson & Svingby, 2007). As analytic rating rubrics tend to provide more specific information for assessment, they may be more beneficial to students if implemented properly, especially when the assessment is for formative purpose (i.e., providing feedback to improve learning).

Specificity of rating rubrics

A rubric needs to be explicit and specific on what and how peer raters are expected to assess, which can be done through giving a detailed overview of the rubric or by given substantive anchor statements for levels in the rating rubric (Jonsson & Svingby, 2007; Panadero et al., 2013). Specificity of Rating Rubrics can prevent misunderstanding or different understandings among peer raters. Prior research has found that rubrics with a sufficient level of specificity could thereby increase peer scoring reliability and construct validity (Jonsson & Svingby, 2007; Miller, 2003; Panadero et al., 2013; Russell et al., 2017).

Sensitivity of rating rubrics

Different from specificity, sensitivity involves the level of discriminating power of the rating rubrics for the set of documents being evaluated. In measurement theory, having enough variability of assessed documents on a measure is a desired quality of any measure, which represents the purpose of measurement (Brennan, 2005)—to differentiate the different qualities/abilities/skills. More specifically for peer assessment, Miller (2003) raised a concern regarding an observed restricted range in peer ratings, which was argued to be largely due to the poor fit of the rating rubric for the assessed population. More generally, if most students in a class are given similar scores on a given rubric (i.e., there is low document variability on the rubric), it does not necessarily mean they have the same performance, but more likely that the rating rubric was not sensitive enough to determine appropriate, more subtle performance differences among the student population. A high-quality rubric should ensure both specificity and sensitivity.

Average ability of peer raters

The Ability of Peer Raters is not directly a characteristic of a rating rubric. But it could be measured at the rubric level if analytic peer rating (i.e., multiple dimensions to be assessed on) was utilised. That is, even within one assignment, mean student performance could be high on one rubric and low on another rubric. This mean level of student performance could shape competence of assessors for the assessment task or trigger different kinds of biases in ratings. It was reported that the bottom 20% of peer reviewers had weaker review accuracy in the context of MOOCs (de Alfaro & Shavlovsky, 2016). In addition, Matsuno (2009) reported that there appeared to be a central tendency effect in peer rating wherein peer raters tended to rate high-quality essays lower and low-quality essays higher. However, the opposite effect was observed in the context of MOOCs wherein high-quality work rated even higher and low-quality work rated even lower (Piech et al., 2013). Although results are not consistent, the ability of peer raters appears to be related to rating accuracy. However, it has not yet been studied how the average ability level of a class on a rubric is associated with the overall reliability or validity of peer assessment results with that rubric.

2.3 | The present study

Based on the gaps in the literature on peer assessment, the present study explores a wide range of variables, across the three different levels that might influence peer assessment reliability and validity. It leverages a large dataset including a variety of schools, disciplines, and students using the same online peer assessment system. Each data point is a rubric in an assignment within a course. One assignment could then contribute multiple rubrics, and one course could contribute multiple assignments. Reliability is assessed for every data point. Validity is also assessed for every data point when instructors also provided assessments. Multi-level (rubrics nested within courses) structural equation models were used to estimate the relationships of the range of variables at different levels with peer assessment reliability and validity. Relationships between the variables with each other, along with the relationship between reliability and validity are explored as well.

Correspondingly, three research questions were tested: (1) what predicts peer assessment reliability?; (2) what predicts peer assessment validity?; and (3) does variation in peer assessment reliability explain validity? For research questions 1 and 2, from the literature review as summarised in the prior section, we hypothesised significant effects for some predictors (Discipline, Class Size, Average Number of Raters per Document, Length of Rubric Description, and Number of Substantive Anchor Statements in the Rubrics) and treated the others as exploratory because of mixed prior findings or theoretical predictions in both directions (see Table 1). Related to research question 3, we also predicted a strong relationship between reliability and validity, but the extent to which each predictor is specifically related to validity or mediated through reliability was treated as an open question. However, through exploratory model building, Document Quality Variability (DQV) was positioned as a critical mediator in the analytical models.

3 | METHODS

3.1 | Dataset

The data were collected from an online peer assessment system, *Peerceptiv* (Cho & Schunn, 2007; Schunn, 2016; Schunn et al., 2016). The peer assessment process in the system includes several steps as shown in Figure 1. First, instructors design rating rubrics for a given assignment, and set up the peer assessment task including the rubrics in the system. At least one rubric was required by the system, but one or all of the seven levels of the rating rubrics could be left blank, thereby including cases that had no meaningful rubric. Second, students use the system to submit their documents before a specified deadline. The assessed documents were most typically written documents, such as essays or science reports, but could also be posters or projects with extensive visual components. Third, students are asked to assess a handful of their peers' documents randomly assigned to them during a reviewing period, providing both ratings and comments according to the rubrics. Fourth, authors are sometimes asked to revise their

TABLE 1 Summary of measures.

Category	Variable (hypothesised Effect)[a]	Measurement level	Measure description
Variables related to the context	Education Level (o)	Institution	University vs. High School
	Language (o)	Institution	English (including Foreign Easy) and Foreign Difficult
	Discipline (+ for Science and Professions)	Course	Humanities (including Business), Science, and Professions
Variables related to the task itself	Assignment # (o)	Assignment/Draft	First Assignment vs. Later Assignments (indicator of rater experience)
	Draft # (o)	Assignment/Draft	First Draft vs. Later Drafts (indicator of rater experience)
	Class Size (−)	Assignment/Draft	Total Number of Documents Submitted
	Average # of Raters per Document (+)	Assignment/Draft	Mean Number of Completed Peer Reviews per Document, ranging from 3 to 6
Variables related to rating rubrics	Total # of Dimensions (o)	Assignment/Draft	Total number of dimensions in the rubrics
	Length of Rubric Description (+)	Rubric Dimension	Total number of words in the rubric description after pre-processing (i.e., removing punctuation and stop words) (indicator of rubric specificity)
	Number of Substantive Anchor Statements in the Rubrics (+)	Rubric Dimension	Number of Substantive Anchor Statements on the 7-point Rating Scale, ranging from 0 to 7 (indicator of rubric specificity)
	Average Ability of Peer Raters (o)	Rubric Dimension	Mean of weighted average peer ratings
	Document Quality Variability (+)	Rubric Dimension	Standard deviation across documents of the weighted average peer ratings for each document (indirect results of rubric sensitivity)

[a]Hypothesised effect of the variables on Reliability/Validity: o, open; +, positive effect; −, negative effect.

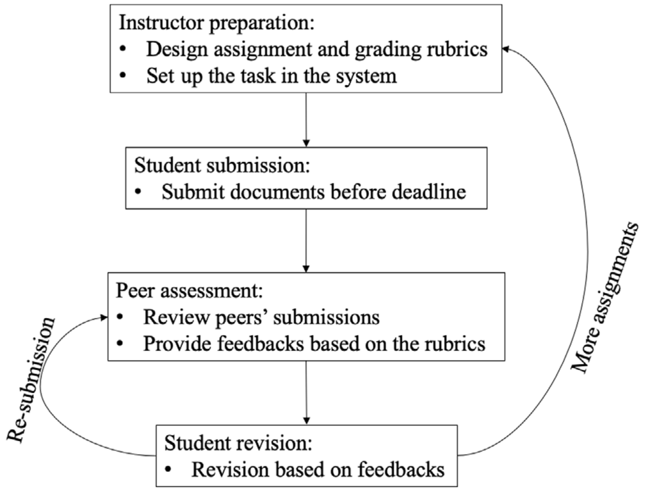
Instructor preparation:
• Design assignment and grading rubrics
• Set up the task in the system
Student submission:
• Submit documents before deadline
Peer assessment:
• Review peers' submissions
• Provide feedbacks based on the rubrics
Student revision:
• Revision based on feedbacks
Re-submission
More assignments

FIGURE 1 A peer assessment process in the online peer assessment system.

documents based on the feedback and resubmit a second (or third) draft for peer assessment.

In the current study, 374 courses from 88 different institutions were used for analyses, representing a total of 19,722 students. The courses that met the following three criteria were selected: (1) at least 10 students submitted documents because reliability and validity are difficult to estimate in extremely small classes; (2) (at the assignment level) fewer than six raters per document so as not to confounding rater workload with group assignments in which multiple students submit a shared document but do reviewing work individually, which results in a large number of raters per document; and (3) (at the rubric level) the rubric used a 7-point rating scale because this was the most common case in the peer assessment system and it is difficult to normalise predictors across scales with varying levels. Although a student could contribute data to multiple courses, in practice this was quite rare (<1%), and thus student-level data across courses was treated as independent, and analyses were conducted instead at the assignment/rubric level.

Those 88 institutions include 33 high schools and 55 universities, with 73 institutions from the United States and 15 institutions from other countries (e.g., Australia, Cambodia, Canada, China, England, Estonia, Singapore, and South Korea). Those 374 courses, given from 2009 to 2014, included 194 humanities courses, 5 business courses, 135 science courses, and 40 professions-related courses (e.g., law, medical, communication disorders, pre-service teacher education, and instructional design). The Human Subjects Review Board at the University of Pittsburgh reviewed the larger project and declared the analysis of anonymised peer assessment data as non-human subject research.

3.2 | Measures

3.2.1 | Inter-rater reliability

IRR was measured using an aggregate consistency type of ICC (Cho et al., 2006) adapted for sparse reviewer × document matrices (i.e., only a few reviewers among many reviewers rate each of the many submitted documents):

$$\mathrm{IRR} = \frac{\mathrm{MS}_{\mathrm{document}} - \mathrm{MS}_{\mathrm{error}}}{\mathrm{MS}_{\mathrm{document}}} \quad (1)$$

The two mean square terms were calculated as below:

$$\mathrm{MS}_{\mathrm{document}} = \frac{\sum_i (x_i - x_{\mathrm{mean}})^2}{N_{\mathrm{document}} - 1} \quad (2)$$

$$\mathrm{MS}_{\mathrm{error}} = \frac{\sum_i \sum_j (x_{ij} - x_i)^2}{N_{\mathrm{rating}} - 1} \quad (3)$$

where *i* and *j* refer to document and rater respectively; x_{mean} is the grand mean for all documents; x_i is the mean for document *i*; x_{ij} is the rating for document *i* by rater *j*; N_{document} is the total number of documents; and N_{rating} is the total number of ratings. Therefore, $\mathrm{MS}_{\mathrm{document}}$ is a measure of variation across documents while $\mathrm{MS}_{\mathrm{error}}$ is a measure of disagreement among raters on the same document. IRR can be negative when the $\mathrm{MS}_{\mathrm{document}}$ is smaller than $\mathrm{MS}_{\mathrm{error}}$ (but the negative cases should be rare), and it has an upper bound of 1, but no lower bound.

IRRs were calculated separately for each rubric within an assignment to investigate the rubric-level IRRs and to examine the effects of the rubric-level predictors on the IRRs. The vast majority of the peer assessment activities used analytic scoring with multiple grading rubrics rather than just one overall holistic rating. Analytic peer rating was utilised in 370 courses out of the 374 courses studied. The 374 courses yielded 3907 IRRs (i.e., approximately 10 IRRs per course). Figure 2 presents a frequency histogram of the IRR values. A small number (1.6%) of IRRs smaller than −1 were winsorised to −1 to adjust for extreme cases. The mean IRR was 0.36. However, the distribution was skewed, with a median of 0.48.

3.2.2 | Validity

In a subset of assignments, instructors rated at least some of the same documents using the system (i.e., using the same interface and the same rubrics). Instructors could choose to rate all the documents, or they could spot check a subset of documents. Validity was measured using the Pearson correlation between averaged peer ratings and instructor ratings (Cho et al., 2006; Li et al., 2016) for a given rubric dimension, treating these instructor ratings as 'ground truth'. Such correlations are bounded by −1 and +1. We included only correlations for the cases in which at least 10 documents were rated by the instructor. There were 1116 such correlations, coming from 97 courses at 38 different schools. Figure 3 presents the frequency histogram of the correlations. The mean of these correlations was 0.45; there was a small skew in the distribution, with a median of 0.51.

The Reliability and Validity measures were at the rubric level within an assignment within a course. They were aggregated measures from a class of students' peer rating data.

3.2.3 | Predictors related to the context

To predict variations in Reliability and Validity, we included a range of measures based on our theoretical framework (see Table 1). Those measures include three contextual variables: Education Level

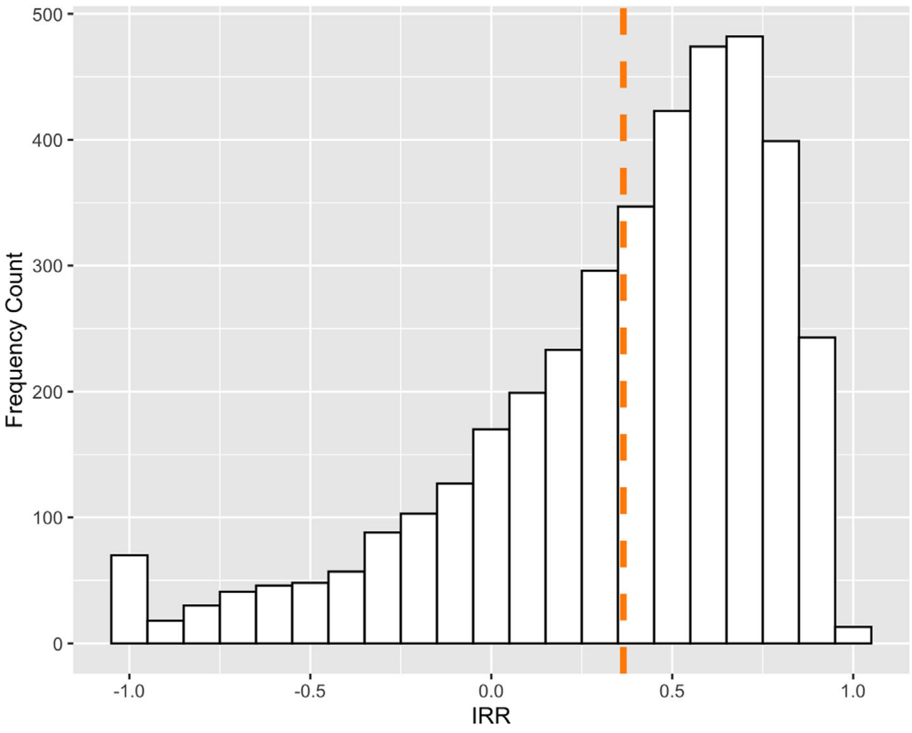

FIGURE 2 Frequency histogram of inter-rater reliability (IRR) values, with the mean value indicated with the dotted line.

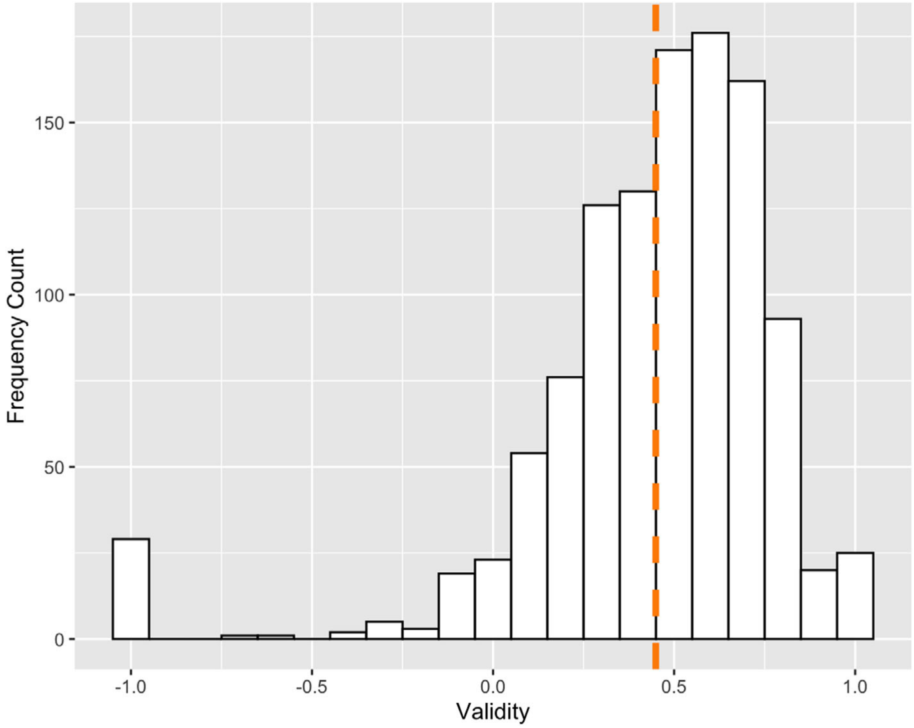

FIGURE 3 Frequency histogram of correlations between peer and instructor ratings, with the mean value indicated with the dotted line.

(University vs. High School), Language (of native country), and Course Discipline.

Language was initially divided into three categories based on its distance to English since the web platform interface was English (English, Foreign Easy, and Foreign Difficult). The categorisation of the language variable was based on the language difficulty ranking (Foreign Service Institute, 2017). Foreign Difficult corresponds to their Category V, which requires 88 weeks (or approximately 2200 h) of studying. Courses conducted in China and Korea were in this category. For analysis, English and Foreign Easy categories were collapsed to be one category due to the small number of cases for easy foreign languages and also due to the high average proficiency with English in the countries represented in the easy foreign languages category.

Course Discipline was measured in four categories (Humanities, Business, Science, and Professions-related courses). Humanities and Business courses were collapsed due to the small number of Business courses and the frequent content overlap of Business courses with humanities topics.

3.2.4 | Predictors related to the peer assessment task itself

Four task-level measures were included. To test effects of Rater Training or Experience, Assignment Number (First Assignment vs. Later Assignment) and Draft Number (First Draft vs. Later Draft) was included with later assignments or drafts indicating that raters had more prior experience and better understanding about the rating rubrics and process: 'first' means the peer assessment was on the first draft submitted by students while 'later' means it was on the revised and resubmitted drafts.

In the system, instructors could specify how many reviews each student was required to do on an assignment. They could also allow students to do additional reviews beyond that number. Sometimes students did not complete the review task. These features led to variation in how many reviews each document received. We simply calculated a mean to produce the Average Number of Raters per Document variable. Note that the small proportions of incomplete responses from individual students did not impact the data analyses because our measures were mainly at the assignment or rubric levels and there were no missing values with the aggregate measures.

We did not have direct access to Class Size in the system. Instead, Class Size was approximated by the Total Number of Documents Submitted.

3.2.5 | Predictors related to rating rubrics

Three characteristics of grading rubrics were also included: Total Number of Dimensions in the assignment rubric; Length of Rubric Descriptions measured by the number of meaningful words (described below); and Number of Substantive Anchor Statements (describe below) for each of the points on the 7-point rating scale. Most of the assignment rubrics involved analytic rather than holistic ratings. Instead of distinguishing rating rubrics in terms of holistic versus analytic as is commonly done in the literature, we used a more granular measure of Total Number of Dimensions. Length of Rubric Descriptions and Number of Substantive Anchor Statements are indicators of rating rubric specificity. The Total Number of Dimensions in the rubrics was measured at the draft level while the other two variables were measured at the rubric dimension level.

Meaningful words were defined as words that are not stop words or punctuation. There were a few (3%) rubric descriptions that had more than 40 meaningful words, which were winsorised to 40 to minimise the statistical biasing effect of extreme cases.

Substantive anchor statements were defined as statements with more than five words. For example, this would exclude statements like '7—Excellent' or '1—Poor' which provided no information about what characteristics a document would need to have in order to receive a given rating. By contrast, anchor statements like '5—Most arguments are well developed and logic and may be supported with independent evidence (references and facts)' would be counted as substantive. Since every included rubric had exactly seven levels and the opportunity to include seven corresponding anchors, the number of substantive anchor statements for a rubric ranged from 0 to 7.

Average Ability of Peer Raters was measured as a weighted mean for each rubric dimension. Weights were based upon each peer reviewer's overall reliability on the assignment, defined by the correlation between that peer reviewer's ratings and mean ratings of the other peer reviewers' rating for each paper in the pool of papers that peer reviewed. This weighting function is used by the online peer assessment system to remove the effects of students who rate randomly or give high ratings to every paper. However, weighted and unweighted means are typically quite similar.

DQV (measured by standard deviation of ratings on documents) in a specific rubric dimension was included in the analysis as a resulting effect of rubric sensitivity. However, it was conceptualised a mediator (as shown in Figures 4 and 5) because (1) it is not directly controlled by the instruction; (2) it is likely influenced by a number of other features of a rubric, like Average Ability of Peer Raters and Draft Number; and (3) early model explorations revealed that it played an important mediating role.

Tables 2 and 3 present the descriptive statistics for all the measures for the full dataset ($N = 3907$) and subset with validity values ($N = 1116$). The Validity and Total Number of Documents measures are leptokurtic, and several variables are skewed; these non-normality issues were addressed in the analytic methods. Table 4 presents the correlations among the continuous variables (full dataset in the lower left; subset with validity values in the upper right). Generally speaking, the two datasets showed very similar descriptive values and variable correlations, suggesting that the validity subset was representative of the larger dataset. Consistent with conceptualisation of DQV as a mediator, it was strongly correlated with reliability and average peer ability.

3.3 | Data analysis procedures

Two path models were constructed: a reliability model and a validity model. The validity model is an extension of the first model but with a smaller sample size because validity data was available in only a subset of the cases. The reliability model is shown in Figure 4 with DQV as a mediator and IRR as the outcome. The validity model is shown in Figure 5, with both DQV and IRR as mediators and Validity as the outcome.

The path models were fitted in Mplus 8 (Muthén & Muthén, 2017). Robust maximum likelihood estimation was used to address the non-normality of the data. In addition, the data had a multilevel structure: rubric dimensions were nested inside drafts, which were nested inside assignments, which were nested inside courses. However, only the nesting of rubrics inside courses was formally modelled because the majority of the courses only had one assignment per course (66%) and one draft per assignment (84%). Therefore, the assignment-/draft-level was not modelled in the analysis. We used

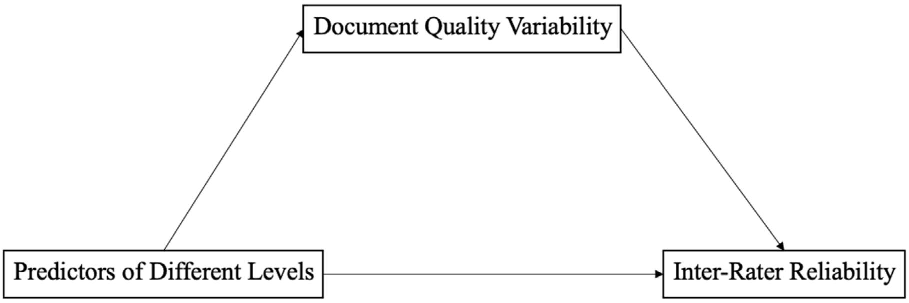

FIGURE 4 Hypothetical model for inter-rater reliability.

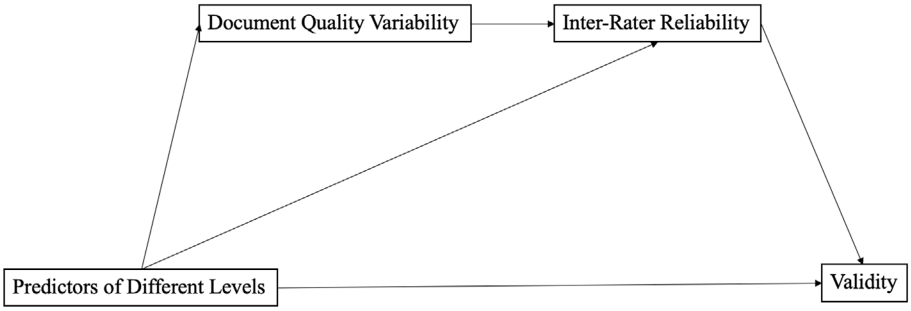

FIGURE 5 Hypothetical model for Validity.

TABLE 2 Descriptive statistics (Mean, *SD*, Skewness, and Kurtosis) for each continuous measure for the full dataset (upper row) and the validity dataset (lower row).

Measures	Mean	*SD*	Skewness	Kurtosis
IRR	0.36	0.44	−1.15	3.99
	0.40	0.43	−1.38	4.52
Validity	0.45	0.34	−2.01	9.27
Total # of dimensions	7.48	3.24	0.74	3.50
	7.99	2.54	−0.08	3.48
Total # of documents submitted	48.68	59.83	3.06	13.86
	49.53	60.78	2.73	10.59
Average # of raters per document	3.93	0.67	0.52	2.62
	3.92	0.63	0.56	2.87
Average Ability of Peer Raters	5.35	0.56	−0.41	4.28
	5.17	0.58	−0.07	3.55
Document Quality Variability	1.27	0.31	0.98	4.99
	1.28	0.29	0.98	5.43
Length of Rubric Description	12.54	9.76	1.40	4.00
	12.49	10.12	1.55	4.06

TABLE 3 Relative frequencies for each level of the categorical measures for the full dataset (top row) and the validity dataset (bottom row).

Measures	*N*	Frequency
Education level	3907	High School: 17%; University: 83%
	1116	High school: 19%; University: 81%
Discipline	3907	Humanities: 53%; Science: 38%; Professions: 9%
	1116	Humanities: 71%; Science: 25%; Professions: 3%
Language	3907	English: 81%; Foreign Difficult: 19%
	1116	English: 58%; Foreign Difficult: 42%
Assignment #	3907	First: 62%; Later: 38%
	1116	First: 59%; Later: 41%
Draft #	3907	First: 85%; Later: 15%
	1116	First: 79%; Later: 21%
Number of substantive Anchor statements	3907	0: 14%; 1: 1%; 2: 5%; 3: 13%; 4: 33%; 5: 2%; 6: 2%; 7: 31%
	1116	0: 2%; 1: 0.4%; 2: 2%; 3: 7%; 4: 34%; 5: 1%; 6: 2%; 7: 51%

'TYPE = COMPLEX' function in Mplus to adjust standard errors due to the nesting structure of the data (Muthén & Muthén, 2017).

Models were built in an iterative fashion to find best fitting parsimonious models, using the fit indices/criteria recommended by Hu and Bentler (1999): a comparative fit index (CFI) value ≥0.90; a root-mean-square-error of the approximation (RMSEA) ≤ 0.06; and the standardised root mean square residual (SRMR) ≤ 0.08. We examined all the three indices and compared them to the recommended cutoffs in order to control both Type I and Type II errors (Hu & Bentler, 1999). We did not use the χ^2 test for model evaluation purpose because even strong models will produce a significant deviation in fit with large enough sample sizes (Markland, 2007). Models that included all the predictors generally had poor model fits. By dropping weak predictors, acceptably fitting models were obtained. However, it is important to note that the core predictors included in the final models were robust predictors across all model variations.

4 | RESULTS

4.1 | Inter-rater reliability model

4.1.1 | Overall model quality

The original model with all variables (see Appendix A) had non-satisfactory fit (CFI = 0.88; RMSEA = 0.03; and SRMR = 0.05), suggesting the need for a better model. After deleting two non-significant variables from the original model (Total # of Dimensions in the Rubrics and Number of Substantive Anchor Statements), the model fit became acceptable (CFI = 0.90; RMSEA = 0.03; and SRMR = 0.05), and this model is presented in Figure 6 and discussed in detail below (see Appendix B for model details). Both the intermediate outcome, DQV, and the final outcome, IRR, were largely explained by the included predictors ($R^2 = 0.37$ and 0.45 respectively; see Figure 6).

4.1.2 | The central role of Document Quality Variability

The existence of a positive path from DQV to IRR is to be expected. However, that DQV is such a strong and primary predictor of IRR ($\gamma = 0.63$, $p < 0.01$) is the important empirical finding of this study. Further, the strength of this one variable allows for other variables to be grouped by those that also have a direct effect on IRR and those that have indirect effects via DQV (i.e., influence IRR levels because they influence DQV levels).

4.1.3 | Other direct predictors of IRR

Three variables were only directly related to IRR. The largest predictor other than DQV was Average # of Raters per Assignment. Interestingly, it negatively predicted IRR ($\gamma = -0.14$, $p < 0.01$). As the number of raters per assignment increases, reviewer work load necessarily increases (Cho & Schunn, 2018), which may reduce the effort students are willing to put into each review. As shown in Figure 7a, the best performance levels were found with approximately three reviewers per document.

Class Size as measured by Total # of Documents submitted for a certain assignment/draft had a small positive effect on IRR ($\gamma = 0.12$, $p < 0.01$), meaning that larger classes had higher IRRs, which is unexpected. Figure 8 shows the bivariate relation between Class Size and IRR (Figure 8a) or DQV (Figure 8b), to explore whether larger classes

TABLE 4 Correlation matrices for the full dataset (lower left) and the validity dataset (upper right).

	2	3	4	5	6	7	8	9
1. Val	0.28***	0.06*	−0.05	0.08**	−0.10**	0.31***	0.01	0.02
2. IRR	–	0.03	0.25***	−0.04	−0.05	0.67***	0.25***	−0.22***
3. #Dim	0.02	–	0.26***	0.08*	0.12***	0.06	0.14***	0.06
4. #Doc	0.13***	0.08***	–	0.22***	0.29***	0.15***	0.47***	−0.22***
5. #Rater	−0.10***	−0.08***	0.29***	–	0.08**	0.12***	−0.08**	−0.24***
6. Mean	−0.23***	0.00	0.13***	0.02	–	−0.26***	0.21***	−0.41***
7. DQV	0.64***	0.04*	0.08***	0.03	−0.43***	–	0.11***	−0.21***
8. Length	0.12***	0.01	0.12***	−0.04*	0.09***	0.05**	–	−0.13***
9. #Anch	−0.04*	0.27***	0.03*	0.08***	−0.11***	−0.06***	−0.05**	–

Note: 1. Validity; 2. Inter-rater reliability (IRR); 3. Total # of dimensions; 4. Total # of documents submitted; 5. Average # of raters per document; 6. Average ability of peer raters; 7. DQV of document quality; 8. Length of rubric description; and 9. Number of substantive anchor statements.
*p < 0.05; **p < 0.01; ***p < 0.001.

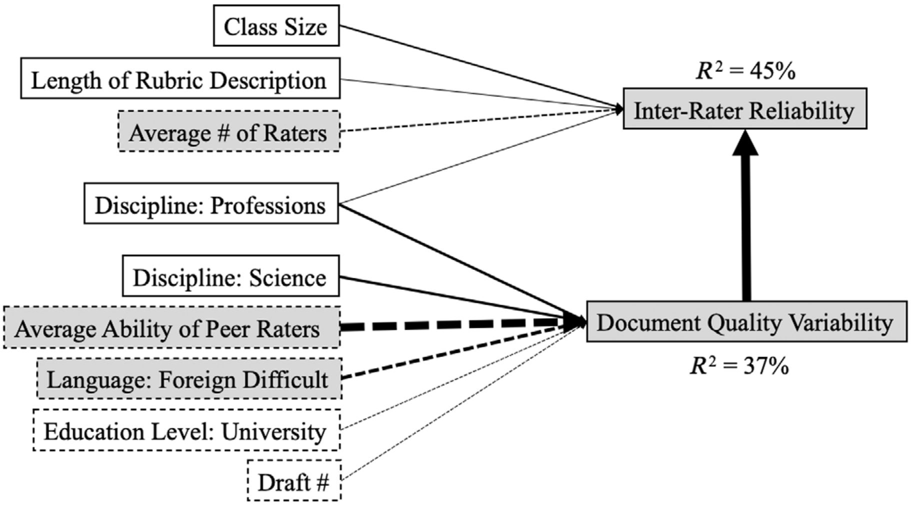
Class Size
Length of Rubric Description
Average # of Raters
Discipline: Professions
Discipline: Science
Average Ability of Peer Raters
Language: Foreign Difficult
Education Level: University
Draft #
$R^2 = 45\%$
Inter-Rater Reliability
Document Quality Variability
$R^2 = 37\%$

FIGURE 6 Inter-rater reliability model results with the dotted line meaning negative coefficient and the width of the line meaning the magnitude of the coefficient. Dotted lines indicate negative effects while solid lines indicate positive effects. Dotted box outline indicates a total negative effect while solid box outline indicates a total positive effect. Shadowed variables have the largest effects overall on inter-rater reliability.

also had more variability. However, the figures show a clear pattern that Class Size was more impactful on IRR than on DQV. Further, the smaller classes with less than 30 submitted papers appeared to have much lower IRRs while only having slightly lower DQV.

As expected, Length of Rubric Description also positively predicted IRR ($\gamma = 0.07$, $p < 0.01$), but with a small magnitude; it is plausible that more qualitative features of rubric descriptions (e.g., the presence of student friendly language; whether the teacher explained the rubrics in class) could be stronger predictors.

4.1.4 | Indirect predictors of IRR

Average Ability of Peer Raters had the strongest indirect effects on IRR through DQV ($\gamma = -0.36$, $p < 0.01$), and it had no direct effects on IRR. The effect of Average Ability of Peer Raters on DQV was generally a ceiling effect. As shown in Table 2, the Average Ability of Peer Raters across all courses was 5.4 out of 7, which was approaching the high end of the scales. Further, Figure 7b shows a severe drop in IRR when the mean was larger than 6 out of 7.

In addition, courses offered in countries where English is a difficult foreign language had a substantially lower IRR levels (see Figure 7c) because they had smaller variations in document quality ($\gamma = -0.16$, $p < 0.01$). This specific group (foreign difficult) was using peer assessment in English language courses, which may have often involved short writing assignments and students with relatively similar basic English writing skills. Note that there was no direct effect on IRR; rather, the problem was one of restricted variation in document quality.

The Discipline variable (Professions-related and Science courses) tended to have a positive effect on IRR that was at least partially indirect via DQV, which is consistent with our hypotheses. Professions-related courses led to higher IRR both through a direct path ($\gamma = 0.09$, $p < 0.01$) and through an indirect path via DQV ($\gamma = 0.11$, $p < 0.01$), while Science had only indirect effect via DQV ($\gamma = 0.11$, $p < 0.01$). That is, the increased level of IRR found in courses from the professions and from the sciences occurred primarily because those courses tended to have higher variability in document quality.

Finally, Education Level (University) and Draft # (Later Draft) had small negative indirect effects on IRR via DQV ($\gamma = -0.06$, $p < 0.05$; and $\gamma = -0.06$, $p < 0.01$). That is, IRR was lower in university courses

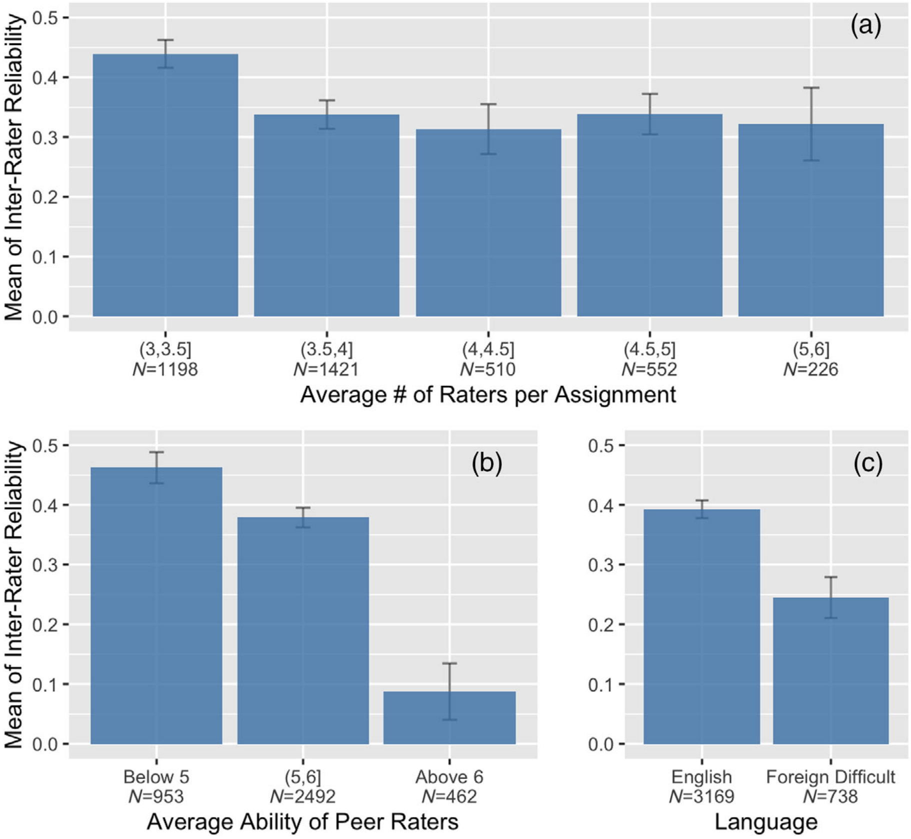

FIGURE 7 Mean of inter-rater reliability as a function of Average # of Raters per Assignment, Average Ability of Peer Raters, and Language. The error bars are shown at the top.

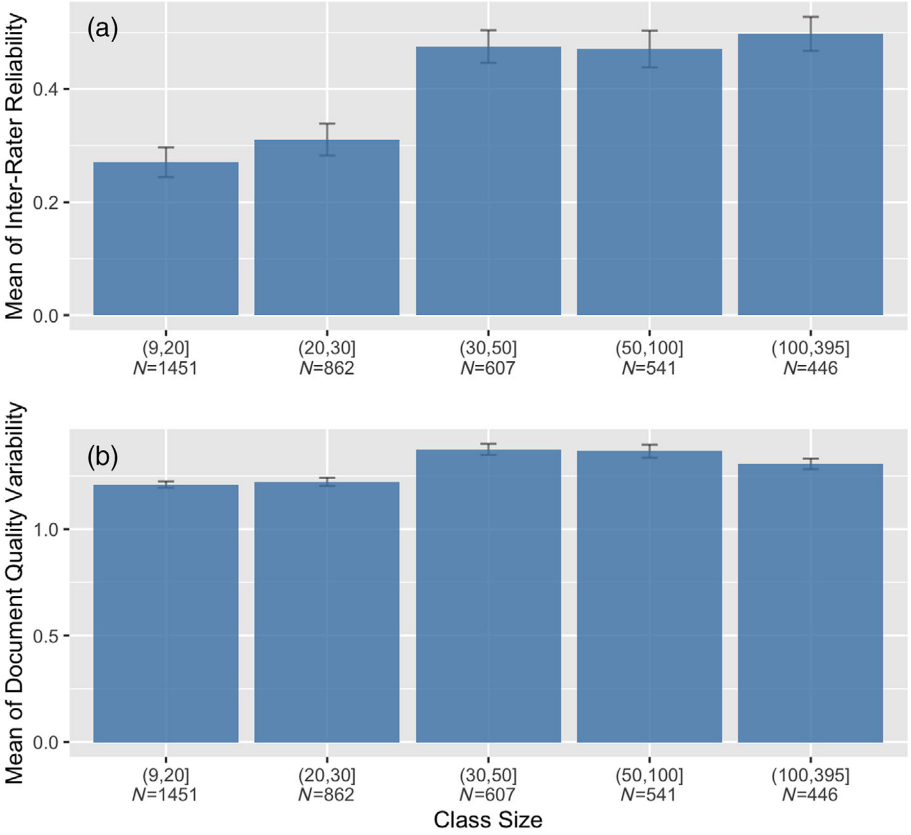

FIGURE 8 Mean of inter-rater reliability and Document Quality Variability as a function of Class Size. The error bars are shown at the top.

through the less document variability in those courses relative to K-12 courses, and later drafts of a given assignment had lower IRR through the decrease of document variability. However, we also note that the magnitude of these variables on IRR was relatively small and perhaps not of practical importance.

4.2 | Validity model

4.2.1 | Overall model quality

The results of the validity model with all variables (see Appendix C) had non-satisfactory model fit ($p < 0.01$; CFI = 0.73; RMSEA = 0.07; and SRMR = 0.11), indicating a need for modification. Five variables from the full validity model were deleted sequentially either due to collinearity or non-significance: Language (Foreign Difficult), Discipline, Education Level (University), Total # of Dimensions in the Rubrics, Length of Rubric Description, and Number of Substantive Anchor Statements. After that, the model fit became good (CFI = 0.93; RMSEA = 0.05; and SRMR = 0.07), and this model is presented in Figure 9 and discussed in detail below (see Appendix D for model details). DQV, IRR, and Validity were explained by the included predictors to some extent ($R^2 = 0.25$, 0.58, and 0.16, respectively) as shown in Figure 9.

4.2.2 | The role of the intermediate variables

Consistent with the IRR model, the strong path from DQV to IRR was found in the validity model as well ($\gamma = 0.56$, $p < 0.01$). In addition, the path from IRR to Validity was positive ($\gamma = 0.24$, $p < 0.01$), which is consistent with the theory that reliability is related to validity in a positive way in that it can set an upper bound for measures of validity (Weiner et al., 2003). Figure 10a shows the positive bivariate relation between IRR and Validity. The mean validity appeared to be highest for the cases with IRR larger than 0.5. Interestingly, the direct effect of DQV on Validity was also significant ($\gamma = 0.21$, $p < 0.01$), even after controlling for the indirect effect from DQV to Validity via IRR. The central role of DQV in mediating the effects on both IRR and Validity is an important finding. In sum, the effect of different predictors on Validity could be either direct or indirect via DQV or IRR or both.

4.2.3 | Direct predictors of validity

Two direct effects on Validity were found. Assignment # and Average # of Raters both have positive direct effect on Validity ($\gamma = 0.16$, $p < 0.01$; and $\gamma = 0.12$, $p < 0.05$). Later assignments tend to have a higher validity as compared with that of the first assignment. Average # of Raters was shown to have negative effect on IRR in the previous section, but it showed a direct positive effect on Validity.

4.2.4 | Indirect predictors of Validity

The two intermediate variables, DQV and IRR, created three possible paths from the predictors to Validity in the model: (a) predictors → DQV → IRR → Validity; (b) predictors → DQV → Validity; and (c) predictors → IRR → Validity. Path (a) was the more indirect path going through both intermediate variables, which often results in a small indirect effect. Path (b) and (c) only had one mediator.

The two variables that had direct effects on Validity also had indirect effects. The indirect effect of Assignment # on Validity was significant only through path (a) ($\gamma = -0.02$, $p < 0.05$) with a trivial magnitude, which means that the total effect of Assignment # on Validity was mainly the direct positive effect. The indirect effect of Average # of Raters on Validity was only through path (c) ($\gamma = -0.05$, $p < 0.01$) with a small magnitude. Therefore, the main total effect of Average # of Raters on Validity was also the direct positive effect. The bivariate relations between the two predictors and Validity are shown in Figure 10b,c, both showing a small positive relation.

Draft #, Average Ability of Peer Raters, and Class Size only had indirect effects on Validity. The indirect effects of Draft # went through all three paths with a total indirect effect of −0.19 ($p < 0.01$). Later drafts had lower Validity, and the effect was through both IRR and DQV. The revised drafts (later drafts) tended to have less variation from one author to another because they had received feedback and made revisions. At the same time, the revised drafts also directly led to lower IRRs.

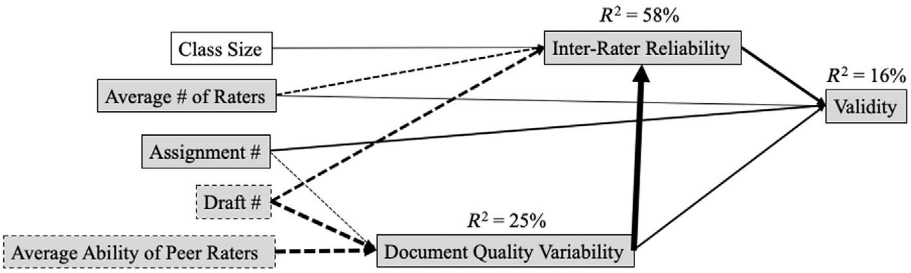

FIGURE 9 Validity model results with the dotted line meaning negative coefficient and the width of the line meaning the magnitude of the coefficient. Dotted lines indicate negative effects while solid lines indicate positive effects. Dotted box outline indicates a total negative effect on Validity while solid box outline indicates a total positive effect on Validity. Shadowed variables have the largest overall effects on Validity.

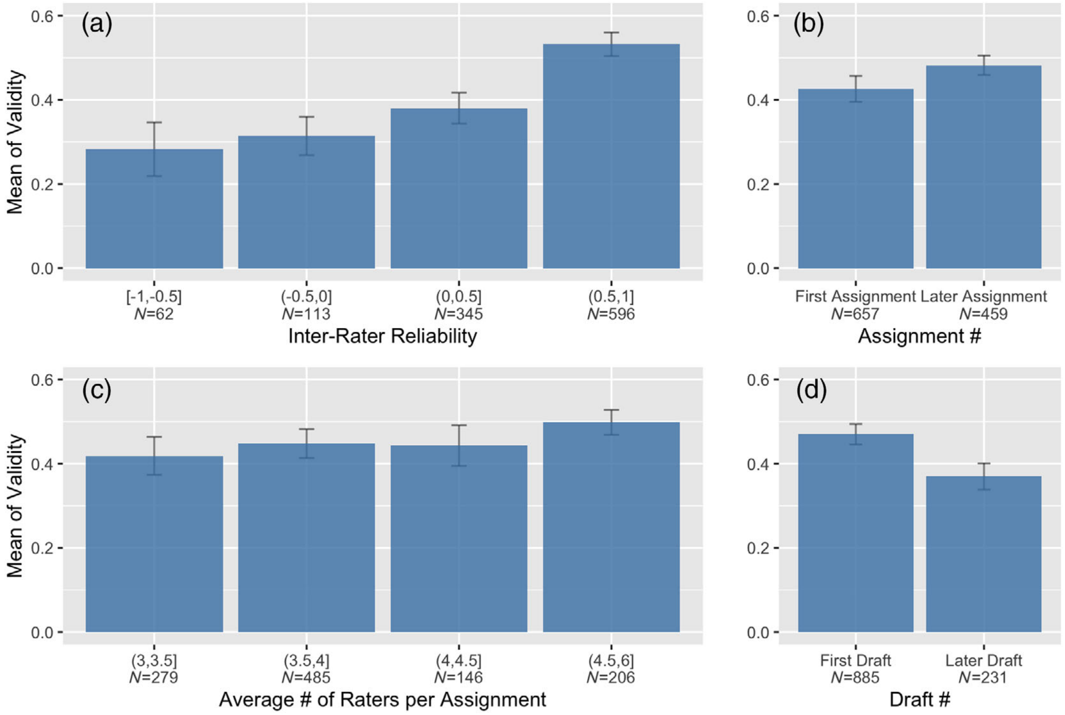

FIGURE 10 Mean of Validity as a function of inter-rater reliability, Assignment #, Average # of Raters per Assignment and Draft #. The error bars are shown at the top.

Average Ability of Peer Raters had a total indirect effect of −0.10 ($p < 0.01$) through path (a) and (b). Again, Average Ability of Peer Raters had a ceiling effect on DQV, which was the main reason for the negative effect on Validity as well. The indirect effect of Class Size was only through path (c) ($\gamma = 0.03$, $p < 0.05$), which was a rather small effect.

5 | DISCUSSION

5.1 | The central pathway: Document Quality Variability

Consistent with the hypothesised IRR model, DQV was found to be a strong direct predictor of IRR and a particularly important mediator of the effects of many other predictors. The strong direct relationship of DQV to IRR was consistently shown in both the Reliability and Validity SEM models. Mathematically, it is consistent with the way IRR is calculated: comparing peer rater disagreement to the variability of the measured documents. As the variability of the measured documents increases, the resulting proportion for a given level of disagreement is therefore less, which leads to a larger IRR. This is also consistent with Generalizability Theory (Brennan, 2005), where reliability is boosted if the purpose of measurement has a larger share of the overall variation. The purpose of measurement is the 'good' variation that we desire, for example, student performance variation; while the other facets are the 'bad' variation that we want to minimise, for example, disagreement among raters. Conceptually, it is also consistent with cognitive processes in human assessment: given larger variability of peer performance, it is usually easier for peer raters to consistently see such large distinctions. By contrast, subtle differences would be a source of disagreement.

In addition to the strong direct effect of DQV on IRR, a direct positive effect was also detected from DQV to Validity. Conceptually, larger distinctions in quality are likely easier to involve agreement, either for the agreement among peer reviewers or the agreement between peer reviewers and instructors.

5.2 | Consistency of the two models: Predictors of both Reliability and Validity

In addition to the important finding on the central role of DQV, the Reliability model and the Validity model also agreed on several other findings, indicating the robustness of the models for the strongest relationships, despite involving substantially different amounts of data. For example, most of the effects on IRR across context, task, and rubric levels were consistently found in both the IRR and Validity models. Here, we discuss each of the consistent relationships.

First, Average # of Raters per Document consistently showed direct negative effects on IRR, which contradicted with our hypothesised positive effect. This negative relationship might be related to workload effects: when each student is asked to rate more documents, they may put less work into each review (Cho & Schunn, 2018). When dealing with concerns about IRR, it is typical to respond by increasing the number of raters to produce a higher effective reliability of the average rating across raters, with the assumption that reliability of each rater is held constant as one increases the number of raters. The findings of our model show that this assumption does not hold. Instead, effective reliability of a set of raters depends upon a tradeoff between the traditional benefits of more raters with negative workload effects. From this large dataset, having three reviewers seemed to be likely to offer a good balance as shown in Figure 7a. However, other contexts not involving the same number of scaffolds in the currently studied peer review system might produce a different ideal number of reviewers.

Second, Average Ability of Peer Raters # had a large negative relationship to both IRR and Validity, directly or through mediation via DQV. This observed large negative effect of Average Ability of Peer Raters was essentially a ceiling effect (i.e., no further variation is possible when the maximum of the scale is reached). Prior research also found that students tended to give high scores in online learning contexts (e.g., Garcia-loro et al., 2020; Li et al., 2020). Pragmatically speaking, it is not important that every dimension avoid this problem. When an assignment has a mixture of rubrics varying widely in average ratings, the rubrics that have ratings clustered at the top will contribute little to variation in overall document ratings, and the rubrics that have more variation will contribute more to overall document ratings. The current analyses show that rubrics with more pragmatic influence on overall scores will also tend to have greater IRR.

Although it is useful for student learning to have students revise and improve their work by implementing peer feedback (Wu & Schunn, 2020a, 2020b), the negative effect of Draft # on IRR and Validity tells a different story from the perspective of rating accuracy. Li et al. (2020) also observed an adverse correlation between the number of reviews completed by each rater and peer assessment reliability in the context of MOOCs. The effect of Draft # on Validity was substantial in size and involved indirect pathways through both DQV and IRR: later drafts were associated with both lower DQV and lower IRR, both of which then led to a lower validity. Having lower DQV in later drafts is perhaps not surprising, as students address the issues their peers are able to identify. The additional direct relationship to IRR may reflect that the remaining issues not addressed in the revisions are ones that are more challenging for the class and thus partially beyond what they are able to consistently identify. An alternative explanation could be that revisions might have addressed comments of some reviewers, but not of all reviewers. When the revised drafts went back to the same set of reviewers, the disagreement among reviewers on the same piece of document potentially turned larger due to the uneven addressing of reviewers' comments. At the same time, large disagreement among raters for later drafts could be confusing to the ratee, potentially harming their trust in the peer assessment process and motivation to build upon the feedback. Training may produce larger consensus among raters. In addition, instructor intervention (e.g., providing feedbacks/ guidance) seems to be necessary for the documents with large disagreement among peers. Our results suggest this could happen more often with later drafts than with the first one. In contexts where instructor or teaching assistant grading is possible, strategically focusing on grading in later drafts is recommended based on these results.

Third, different from the hypothesised negative effect, Class Size had a positive effect on both IRR and Validity. The effect was direct for IRR and indirect via IRR for Validity. There might be specific features of small classes that contributed to the smaller IRRs; it is plausible that the small classes were less resourced (e.g., less teaching assistant help), which might have resulted in less structured implementation of peer assessment activities. This positive effect supports the use of interactive instructional models like peer assessment especially in large classes; such support is important because strategies like peer assessment enable large classes to assign complex tasks that normally could not be assigned due to limited assessment resources. That the relationship was positive and not mediated via DQV is interesting. Follow-up research is required to understand this pattern. For example, are larger classes more likely to involve explicit training on peer review or do smaller classes tend to have more complex or heterogeneous document topics (e.g., as would be the case in research papers) and less functional anonymity (i.e., reviewers can guess author identity because of classroom discussions of paper topics)?

Fourth, there was the expected positive relationship between IRR and Validity. RQ3 is strongly supported by the results. This relationship was sufficiently strong and consistent that it supports a useful rule-of-thumb for instructors: in order to have a high likelihood of having an acceptable validity, it seems necessary to have at least a medium-high IRR (e.g., >0.5 based in Figure 10a). However, it was also interesting to note that having near zero IRR still generally involved positive validity correlations. Thus, these data call into question the general wisdom that reliability is a necessary pre-condition of validity. Instead, we observed that moderate levels of validity are possible despite low levels of IRR, but high levels of validity are most consistently obtained with at least moderate reliability levels. This pattern can be reconciled with a view of peer assessment as individually incomplete: each peer tends to identify and comment upon legitimate problems in the document (Wu & Schunn, 2020c) but they rarely notice all the problems in the document (Gao et al., 2019). That is, because different students see different issues, potentially of different importance, IRR can be expected to be low. But the problems they notice are real and thus contribute to a valid assessment, particularly when the average of multiple peers are combined.

To sum up, the major effects on IRR were consistent to a large extent in both the Reliability and Validity models. Three variables, Average Ability of Peer Raters, Draft # and Class Size, drove both IRR and Validity in a similar way but with different magnitudes. The ceiling effect of Average Ability of Peer Raters is somewhat consistent with the central tendency effect in Matsuno (2009) or the exaggeration effect in Piech et al. (2013), which both indicated more errors at extreme quality levels. Interestingly, neither the Average Ability of Peer Raters effect nor the Class Size effect nor the Draft # effect has been studied in meta-analyses of reliability or validity. The currently observed draft # effect is somewhat contradicting the positive rater involvement effect reported in meta-analysis (Li et al., 2016). However, repeated use of peer assessment in later drafts is importantly different from direct rater training in that it confounds rating experience effect with other variables like potentially more homogeneous documents, raters' expectations from a previous draft, or document improvements based upon feedback that do not translate into better understanding about the rubrics and improved performance in later assignments. Raters' motivation might also explain the negative correlation between Draft # and IRR/Validity (Li et al., 2020; Meek et al., 2017). If students are less motivated to participate in multiple drafts, more experiences might reduce peer assessment reliability/validity. As expected, more significant moderators including small effect size moderators were found in this study, as compared to most meta-analyses, due to the larger statistical power.

5.3 | Unique effects that drive reliability and validity

Although there was much agreement in the findings of the Reliability and Validity models, there were several differences between the two models. Some of the differences were in what was a statistically significant predictor of DQV or IRR, and some of the differences are between what are the effects on IRR versus effects on Validity.

To begin with, while Average # of Raters per Document consistently showed negative effects on IRR across the two models, its effect on Validity was positive and direct. As mentioned in the previous section regarding a potential workload effect on IRR, the average rating drawn from a larger number of peer ratings appeared to produce a more unbiased evaluation and therefore higher validity. One explanation is that the different biases possibly cancelled out each other with a larger number of peer raters (Brennan, 2005). Thus, practitioners will need to find a good tradeoff between (1) balancing biases via obtaining more diverse evaluations and (2) selecting a peer assessment workload that is appropriate for students and the assessment task.

Second, Assignment # was positively related to Validity, but not IRR. This pattern of results for Assignment # (including the negative relationship to DQV) is consistent with a practice effect (improvements in understanding of rubrics and task performance), and it supports the use of peer assessment activities at multiple times for different assignments throughout the semester to achieve better consensus between peers and experts. In theory, getting authentic peer assessment experience seems to be an effective approach for peer raters to understand the rubrics as suggested by the positive relation between Assignment # and Validity. Practically speaking however, further research is required to understand whether validity continues to improve after the second assignment. The current dataset had too few courses using more than two assignments to robustly examine this issue beyond the second assignment.

Third, several contextual variables significantly associated with IRR (i.e., Language, Education Level, and Discipline) did not show statistical significance with IRR or Validity in the Validity model. These contextual variables may have been primarily related to IRR only, with effects that were not strong enough to drive Validity. In addition, with the smaller dataset used in the Validity model, those effects may have disappeared due to the reduced statistical power. Alternatively, there may have been systematic differences in the courses that included validity data. Peer assessment may have different reliability and validity results when implemented in different contexts (Falchikov & Goldfinch, 2000; Li et al., 2016). Additional research will be needed to understand what is underlying the effects of those three contextual variables on IRR to further unpack why they might not have effects on Validity. Practically, the lack of any large effects of those contextual variables on Validity supports the broad use of peer assessment across those contexts of language, Educational Level, and Discipline.

Fourth, only one specific rubric-related characteristic (Length of Rubric Description) was found to be associated with IRR, but this effect was rather small and it did not show up as a significant predictor of IRR or Validity in the Validity model, where the sample size decreased. On the one hand, long rubric descriptions may include multiple problems, which may distract rater attention from the most significant problem. On the other hand, this finding may suggest that rubrics may play less of a central role in validity than was hypothesised. For example, perhaps training with exemplars may be sufficient to counter the negative effects of weak rubrics. That the specific rubric-related variables that were tested here did not show strong effects may simply suggest more qualitative and fine-grained characteristics need to be investigated to specifically target rubric sensitivity and specificity.

5.4 | Practical implications

There are several important implications for practice. First, when instructors want to increase IRR and validity of ratings, they can include more students with different proficiency levels in the subject area in peer assessment (e.g., by including students across sections in a common peer assessment task) or they should avoid using rubric scales that will mainly have students cluster at the top (or bottom) of the scale. Second, the appropriate number of peer raters usually lies within 3-to-5 raters per document. Depending on the purpose of the evaluation, whether it is more important to have agreement among peer raters or to have agreement between peer raters and the instructor, a slightly smaller number or larger number should be selected. Instructors should avoid going beyond 5 raters, as this large workload could negatively motivate student engagement. Third, teachers should use peer assessment activities primarily for first drafts rather than for second or later drafts of an assignment to avoid confusion to students caused by larger disagreements among peers or between peers and instructors.

5.5 | Limitations and future research

The study bears several limitations. First, while we have included a broad range of variables in the model across context, assessment task, and rubric levels, the list of measured variables is neither perfect nor exhaustive. More fine-grained or qualitative characteristics that were not available in the currently studied dataset could be investigated in future research (e.g., the presence of student-friendly language; whether the students were included in the rubric development; whether teacher explained the rubric in class). In addition, inclusion of other characteristics of the actual peer assessment tasks or characteristics of the measured document in the model might produce extra insights.

Second, while we have included a large-scale cross-context sample in this study, the majority of the sample was first draft and first assignment because most of the courses only implemented peer assessment once throughout the course. This imbalance in the sample might have introduced some biases regarding to the effects of Draft # and Assignment #. Therefore, those effects are worthy of further investigation.

Third, this study collected data from a specific online peer assessment platform, with its specific functionalities. Many of the features used in this platform (e.g., anonymous, multi-peer, analytic reviewing with support comments) are relatively common. However, the online peer assessment system at the time only supported written or visual documents to be peer reviewed, not video or audio documents. Generalising the findings to an even broader context (e.g., paper-based reviewing, reviewing on multimedia documents or non-anonymous reviewing) will require collecting data from those contexts.

Finally, it is important to recognise that the current SEM analyses, similar to moderator analyses in meta-analysis, are fundamentally correlational in nature. Many important correlates were included in the analyses, and because of the improved power in the current approach, unique statistical relationships were better teased apart. However, there could be other confounds at play, and future experimental manipulations should be conducted to verify the causal status of the variables that were identified in the current study.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/jcal.12861.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in IRR and validity at https://osf.io/rfvcg/.

ORCID

Christian D. Schunn https://orcid.org/0000-0003-3589-297X

Yong Wu https://orcid.org/0000-0001-9224-8561

REFERENCES

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Psychological Association.

Babcock, P., & Betts, J. R. (2009). Reduced-class distinctions: Effort, ability, and the education production function. *Journal of Urban Economics*, *65*, 314–322. https://doi.org/10.1016/j.jue.2009.02.001

Birjandi, P., & Hadidi Tamjid, N. (2012). The role of self-, peer and teacher assessment in promoting Iranian EFL learners' writing performance. *Assessment & Evaluation in Higher Education*, *37*(5), 513–533. https://doi.org/10.1080/02602938.2010.549204

Bouzidi, L., & Jaillet, A. (2009). Can online peer assessment be trusted? *Educational Technology & Society*, *12*(4), 257–268. https://www.learntechlib.org/p/74983/

Brennan, R. L. (2005). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27–34. https://doi.org/10.1111/j.1745-3992.1992.tb00260.x

Chang, C.-C., Tseng, K.-H., Chou, P.-N., & Chen, Y.-H. (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

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education*, *48*(3), 409–426. https://doi.org/10.1016/j.compedu.2005.02.004

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. W. (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

Cortina, J. M. (2016). Apples and oranges (and pears, oh my!): The search for moderators in meta-analysis. *Organizational Research Methods*, *6*(4), 415–439. https://doi.org/10.1177/1094428103257358

de Alfaro, L., & Shavlovsky, M. (2016). Dynamics of peer grading: An empirical study. In *The 9th International Conference on Educational Data Mining* (pp. 62–69). International Educational Data Mining Society.

Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2019). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review*, *32*(2), 481–509. https://doi.org/10.1007/s10648-019-09510-3

Dynarski, S., Hyman, J., & Schanzenbach, D. W. (2013). Experimental evidence on the effect of childhood investment on postsecondary attainment and degree completion. *Journal of Policy Analysis and Management*, *32*(4), 692–717. https://doi.org/10.1002/pam.21715

Ericsson, K. A. (2008). Deliberate practice and acquisition of expert performance: A general overview. *Academic Emergency Medicine*, *15*(11), 988–994. https://doi.org/10.1111/j.1553-2712.2008.00227.x

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

Foreign Service Institute. (2017). Language difficulty ranking: Effective language learning. http://www.effectivelanguagelearning.com/language-guide/language-difficulty

Gao, Y., Schunn, C. D., & Yu, Q. (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. https://doi.org/10.1080/02602938.2018.1499075

Garcia-Loro, F., Martin, S., Ruiperez-Valiente, J. A., San, C. E., & Castro, M. (2020). Reviewing and analyzing peer review inter-rater reliability in a MOOC platform. *Computers & Education*, *154*(4), 1–35. https://doi.org/10.1016/j.compedu.2020.103894

Gielen, M., & De Wever, B. (2015). Structuring peer assessment: Comparing the impact of the degree of structure on peer feedback content. *Computers in Human Behavior*, *52*, 315–325. https://doi.org/10.1016/j.chb.2015.06.019

Gielen, S., Dochy, F., & Onghena, P. (2011). An inventory of peer assessment diversity. *Assessment & Evaluation in Higher Education*, *36*(2), 137–155. https://doi.org/10.1080/02602930903221444

Glass, G. V. (2000). Meta-analysis at 25. http://www.gvglass.info/papers/meta25.html

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jeffery, D., Krassimir, Y., Alison, C., & Kerry, R. (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

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., … Othman, A. R. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education*, *11*(2), 121–137. https://doi.org/10.1207/s15324818ame1102_1

Li, H., Bialo, J. A., Xiong, Y., Hunter, C. V., & Guo, X. (2021). The effect of peer assessment on non-cognitive outcomes: A meta-analysis. *Applied Measurement in Education*, *34*(3), 179–203. https://doi.org/10.1080/08957347.2021.1933980

Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education*, *45*(2), 193–211. https://doi.org/10.1080/02602938.2019.1620679

Li, H., Xiong, Y., Zang, X., Kornhaber, M. L., Lyu, Y., Chung, 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, H., Zhao, C., Long, T., Huang, Y., & Shu, F. (2021). Exploring the reliability and its influencing factors of peer assessment in massive open online courses. *British Journal of Educational Technology*, *52*, 2263–2277. https://doi.org/10.1111/bjet.13143

Luxton-Reilly, A. (2009). A systematic review of tools that support peer assessment. *Computer Science Education*, *19*(4), 209–232. https://doi.org/10.1080/08993400903384844

Magin, D., & Helmore, P. (2001). Peer and teacher assessments of oral presentation skills: How reliable are they? *Studies in Higher Education*, *26*(3), 287–298. https://doi.org/10.1080/03075070120076264

Markland, D. (2007). The golden rule is that there are no golden rules: A commentary on Paul Barrett's recommendations for reporting model fit in structural equation modelling. *Personality and Individual Differences*, *42*(5), 851–858. https://doi.org/10.1016/j.paid.2006.09.023

Marty, M. C., Henning, J. M., & Willse, J. T. (2010). Accuracy and reliability of peer assessment of athletic training psychomotor laboratory skills. *Journal of Athletic Training*, *45*(6), 609–614. https://doi.org/10.4085/1062-6050-45.6.609

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing*, *26*(1), 75–100. https://doi.org/10.1177/0265532208097337

Meek, S. E. M., Blakemore, L., & Marks, L. (2017). Is peer review an appropriate form of assessment in a MOOC? Student participation and performance in formative peer review. *Assessment and Evaluation in Higher Education*, *42*(6), 1000–1013. https://doi.org/10.1080/02602938.2016.1221052

Miller, P. J. (2003). The effect of scoring criteria specificity on peer and self-assessment. *Assessment & Evaluation in Higher Education*, *28*(4), 383–394. https://doi.org/10.1080/0260293032000066218

Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide* (7th ed.). Muthén & Muthén.

Panadero, E., Romero, M., & Strijbos, J. W. (2013). The impact of a rubric and friendship on construct validity of peer assessment, perceived fairness and comfort, and performance. *Studies in Educational Evaluation*, *39*(4), 195–203. https://doi.org/10.1016/j.stueduc.2013.10.005

Patchan, M. M., Schunn, C. D., & Clark, R. J. (2017). Accountability in peer assessment: Examining the effects of reviewing grades on peer ratings and peer feedback. *Studies in Higher Education*, *43*, 2263–2278. https://doi.org/10.1080/03075079.2017.1320374

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (2013). Tuned models of peer assessment in MOOCs. In *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 153–160). International Educational Data Mining Society.

Rasouli, S., & Esfandiari, R. (2022). Severity differences across proficiency levels among peer-assessors. *Journal of Modern Research in English Language Studies*, *9*(2), 173–196.

Russell, J., Van Horne, S., Ward, A. S., Bettis, E. A., III, & Gikonyo, J. (2017). Variability in students' evaluating processes in peer assessment with calibrated peer review. *Journal of Computer Assisted Learning*, *33*, 178–190. https://doi.org/10.1111/jcal.12176

Sadler, P., & 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

Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology*, *109*(8), 1049–1066. https://doi.org/10.1037/edu0000190

Schunn, C. D. (2016). Writing to learn and learning to write through SWoRD. In S. A. Crossley & D. S. McNamara (Eds.), *Adaptive educational technologies for literacy instruction*. Taylor & Francis.

Schunn, C. D., Godley, A., & DeMartino, S. (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. https://doi.org/10.1002/jaal.525

Sung, Y.-T., Chang, K.-E., Chang, T.-H., & Yu, W.-C. (2010). How many heads are better than one? The reliability and validity of teenagers' self- and peer assessments. *Journal of Adolescence*, *33*(1), 135–145. https://doi.org/10.1016/j.adolescence.2009.04.004

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

Tseng, S. C., & Tsai, C. C. (2007). On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education*, *49*(4), 1161–1174. https://doi.org/10.1016/j.compedu.2006.01.007

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

Weiner, I. B., Freedheim, D. K., Schinka, J. A., & Velicer, W. F. (2003). *Handbook of psychology, research methods in psychology*. Wiley.

Wu, Y., & Schunn, C. D. (2020a). From feedback to revisions: Effects of feedback features and perceptions. *Contemporary Educational Psychology*, *60*, 101826. https://doi.org/10.1016/j.cedpsych.2019.101826

Wu, Y., & Schunn, C. D. (2020b). The effects of providing and receiving peer feedback on writing performance and learning of secondary school students. *American Educational Research Journal*, *58*(3), 492–526. https://doi.org/10.3102/0002831220945266

Wu, Y., & Schunn, C. D. (2020c). When peers agree, do students listen? The central role of feedback quality and feedback frequency in determining uptake of feedback. *Contemporary Educational Psychology*, *62*, 101897. https://doi.org/10.1016/j.cedpsych.2020.101897

Wu, Y., & Schunn, C. D. (2021). From plans to actions: A process model for why feedback features influence feedback implementation.

Instructional Science, *49*, 365–394. https://doi.org/10.1007/s11251-021-09546-5

Xiao, Y., & Lucking, R. (2008). The impact of two types of peer assessment on students' performance and satisfaction within a Wiki environment. *Internet and Higher Education*, *11*, 186–193. https://doi.org/10.1016/j.iheduc.2008.06.005

Zhang, F., Schunn, C., Li, W., & Long, M. (2020). Changes in the reliability and validity of peer assessment across the college years. *Assessment & Evaluation in Higher Education*, *45*(8), 1073–1087. https://doi.org/10.1080/02602938.2020.1724260

Zheng, L., Zhang, X., & Cui, P. (2019). The role of technology-facilitated peer assessment and supporting strategies: A meta-analysis. *Assessment & Evaluation in Higher Education*, *45*(3), 372–386. https://doi.org/10.1080/02602938.2019.1644603

Zong, Z., Schunn, C. D., & Wang, Y. (2021). What aspects of online peer feedback robustly predict growth in students' task performance? *Computers in Human Behavior*, *124*, 106924. https://doi.org/10.1016/j.chb.2021.106924

How to cite this article: Xiong, Y., Schunn, C. D., & Wu, Y. (2023). What predicts variation in reliability and validity of online peer assessment? A large-scale cross-context study. *Journal of Computer Assisted Learning*, 1–21. https://doi.org/10.1111/jcal.12861

APPENDIX A

ORIGINAL MODEL FOR INTER-RATER RELIABILITY

	Std. coefficient	*z*-Value	*p*-Value
Document Quality Variability ON			
Education Level: University	−0.08	−1.67	0.10
Language: Foreign Difficult	−0.26	−4.72	<0.01
Discipline: Professions	0.17	5.49	<0.01
Discipline: Science	0.17	3.68	<0.01
Assignment #	−0.05	−1.54	0.13
Draft #	−0.09	−3.38	<0.01
Class Size	0.04	1.14	0.25
Average # of Raters per Document	0.01	0.31	0.76
Total # of Dimensions	0.03	0.62	0.53
Length of Rubric Description	0.00	0.14	0.89
Number of Substantive Anchor Statements in the Rubrics	−0.01	−0.17	0.87
Average Ability of Peer Raters	−0.58	−16.35	<0.01
Inter-Rater Reliability ON			
Education Level: University	−0.05	−1.63	0.10
Language: Foreign Difficult	0.07	1.73	0.08
Discipline: Professions	0.09	2.95	<0.01
Discipline: Science	0.02	0.49	0.63
Assignment #	−0.01	−0.47	0.64
Draft #	−0.07	−1.60	0.11
Class Size	0.12	4.97	<0.01
Average # of Raters per Document	−0.15	−6.06	<0.01
Total # of Dimensions	−0.04	−1.54	0.12
Length of Rubric Description	0.07	3.29	<0.01
Number of Substantive Anchor Statements in the Rubrics	0.01	0.29	0.77
Average Ability of Peer Raters	0.05	1.63	0.10
Document Quality Variability	0.63	24.67	<0.01

(Continues)

APPENDIX B

FINAL MODEL FOR INTER-RATER RELIABILITY

	Std. coefficient	*z*-Value	*p*-Value
Document Quality Variability ON			
Education Level: University	−0.09	−2.03	<0.05
Language: Foreign Difficult	−0.25	−5.27	<0.01
Discipline: Professions	0.17	5.37	<0.01
Discipline: Science	0.18	3.73	<0.01
Assignment #	−0.05	−1.52	0.13
Draft #	−0.09	−3.43	<0.01
Class Size	0.04	1.27	0.20
Average # of Raters per Document	0.00	0.17	0.87
Length of Rubric Description	0.01	0.16	0.87
Average Ability of Peer Raters	−0.58	−16.53	<0.01
Inter-Rater Reliability ON			
Education Level: University	−0.04	−1.45	0.15
Language: Foreign Difficult	0.07	1.76	0.08
Discipline: Professions	0.09	3.06	<0.01
Discipline: Science	0.01	0.30	0.76
Assignment #	−0.01	−0.49	0.63
Draft #	−0.07	−1.53	0.13
Class Size	0.12	4.83	<0.01
Average # of Raters per Document	−0.14	−5.77	<0.01
Length of Rubric Description	0.07	3.31	<0.01
Average Ability of Peer Raters	0.04	1.52	0.13
Document Quality Variability	0.63	24.56	<0.01

APPENDIX C

ORIGINAL MODEL FOR VALIDITY

	Std. coefficient	z-Value	p-Value
Document Quality Variability ON			
Education Level: University	−0.02	−0.26	0.80
Language: Foreign Difficult	−0.40	−3.87	<0.01
Discipline: Professions	0.21	3.05	<0.01
Discipline: Science	0.17	3.41	<0.01
Assignment #	−0.02	−0.30	0.76
Draft #	−0.16	−3.79	<0.01
Class Size	0.03	0.41	0.68
Average # of Raters per Document	−0.01	−0.17	0.86
Total # of Dimensions	0.10	1.59	0.11
Length of Rubric Description	−0.06	−1.45	0.15
Number of Substantive Anchor Statements in the Rubrics	−0.06	−0.72	0.47
Average Ability of Peer Raters	−0.60	−8.38	<0.01
Inter-rater Reliability ON			
Education Level: University	−0.04	−0.67	0.50
Language: Foreign Difficult	−0.01	−0.17	0.87
Discipline: Professions	0.05	1.15	0.25
Discipline: Science	−0.05	−0.93	0.35
Assignment #	−0.05	−1.27	0.21
Draft #	−0.30	−5.12	<0.01
Class Size	0.11	2.67	<0.01
Average # of Raters per Document	−0.21	−4.28	<0.01
Total # of Dimensions	0.00	0.05	0.96
Length of Rubric Description	0.07	2.07	<0.05
Number of Substantive Anchor Statements in the Rubrics	0.04	1.02	0.31
Average Ability of Peer Raters	0.08	1.62	0.11
Document Quality Variability	0.55	12.38	<0.01
Validity ON			
Education Level: University	−0.04	−0.59	0.56
Language: Foreign Difficult	−0.33	−2.22	<0.05
Discipline: Professions	−0.07	−1.55	0.12
Discipline: Science	0.02	0.23	0.82
Assignment #	0.21	4.70	<0.01
Draft #	0.08	1.60	0.11
Class Size	−0.20	−1.07	0.29
Average # of Raters per Document	0.10	1.60	0.11
Total # of Dimensions	0.02	0.51	0.61
Length of Rubric Description	0.00	−0.06	0.95
Number of Substantive Anchor Statements in the Rubrics	0.24	1.88	0.06
Average Ability of Peer Raters	−0.03	−0.43	0.67
Document Quality Variability	0.17	2.62	<0.01
Inter-Rater Reliability	0.23	4.20	<0.01

APPENDIX D

FINAL MODEL FOR VALIDITY

	Std. coefficient	z-Value	p-Value
Document Quality Variability ON			
Assignment #	−0.12	−2.33	<0.05
Draft #	−0.34	−8.23	<0.01
Class Size	0.15	2.30	<0.05
Average # of Raters per Document	0.03	0.47	0.64
Average Ability of Peer Raters	−0.33	−5.27	<0.01
Inter-Rater Reliability ON			
Assignment #	−0.07	−1.64	0.10
Draft #	−0.30	−4.98	<0.01
Class Size	0.11	3.05	<0.01
Average # of Raters per Document	−0.20	−4.96	<0.01
Average Ability of Peer Raters	0.07	1.59	0.11
Document Quality Variability	0.56	16.54	<0.01
Validity ON			
Assignment #	0.16	4.23	<0.01
Draft #	0.05	1.08	0.28
Class Size	−0.15	−0.78	0.43
Average # of Raters per Document	0.12	2.00	<0.05
Average Ability of Peer Raters	0.02	0.39	0.70
Document Quality Variability	0.21	4.26	<0.01
Inter-Rater Reliability	0.24	4.59	<0.01