2017

A Modified Peer Rating System to Recognise Rating Skill as a Learning Outcome

Violet Cheung
University of San Francisco, vcheung@usfca.edu

Saera R. Khan
University of San Francisco, srkhan@usfca.edu

Recommended Citation
Cheung-Blunden, V., Khan, S.R. A modified peer rating system to recognise rating skill as a learning outcome (2017) Assessment and Evaluation in Higher Education, pp. 1-10. Article in Press. DOI: 10.1080/02602938.2017.1280721

A modified peer rating system to recognise rating skill as a learning outcome

[a] Violet Cheung-Blunden
University of San Francisco, San Francisco, California, USA

[b] Saera R. Khan
University of San Francisco, San Francisco, California, USA

[a] Correspondence regarding this article should be addressed to Dr. Violet Cheung-Blunden, Associate Professor, Department of Psychology, University of San Francisco, 2130 Fulton Street, San Francisco, CA 94117, USA. Email: vcheung@usfca.edu. Tel. (415)422-4373
Violet Cheung-Blunden is an associate professor in the psychology department at the University of San Francisco, a board member and an associate editor for the Society for Terrorism Research, and an officer at the Peace Division of the American Psychological Association. Her primary research interests are mass emotions and public sentiments.

[b] Saera R. Khan, Professor, Department of Psychology, University of San Francisco, 2130 Fulton Street, San Francisco, CA 94117, USA. Email: srkhan@usfca.edu. Tel. (415)422-5794
Saera Khan is a social psychologist who received her PhD at Washington University in St. Louis. Her research explores the social cognitive processes involved in the reliance of stereotypes for judgment and the consequences of these prejudicial judgments for stigmatized groups. She also studies the role of identity and culture in the formation of moral judgments. She teaches Psychology Practicum, Social Psychology, and Advanced Research Methods.

Abstract

The peer rating system used here advances the quantitative literacy goals outlined in social sciences. We instituted a mid-semester intervention to teach rating skills and used an objective index to track longitudinal changes of skill mastery over the course of the semester. Seventy-four students in five advanced research classes followed the procedure of the existing peer rating system by completing reading assignments, writing reflections online, engaging in class discussions, rating their peers' reflections, and receiving feedback of their group effort. Unique to our modified system, peer ratings were compared with each other and also with the instructor ratings to derive individualized indices of reliability and validity. These technical indicators enabled two rounds of assessment before and after a class-wide intervention. An omnibus test across the five classes showed a significant improvement in rating quality due to the intervention. Our courses not only met a quantitative learning outcome but also promised vocational competence.

Keywords: measurement skills; coding open-ended responses; peer ratings; inter-rater reliability

Peer assessments are a necessity in an education environment that is characterized by fiscal constraints and large student/teacher ratios (Spatar, Penna, Mills, Kugija, & Cooke, 2015). Peer assessment has become a pragmatic tool because it saves time for instructors (Sadler & Good, 2006) and may even replace instructor-generated scores if weighed carefully (Nepal, 2012; Spatar et al., 2015). Pragmatism, in the eyes of the instructor, does little to incentivize students to participate in rating exercises (Loddington, Pond, Wilkinson, & Willmot, 2009; Neus, 2011). One way to incentivize students is to grade their rating efforts (Reader, 2007). However, instructors rarely assess the quality of peer ratings. A simple tally of the number of times students missed rating assignments attests to quantity rather than quality. Group-based analyses speak to group responsibility rather than individual accountability (Johnston & Miles, 2004; Zhang & Ohland, 2009).

Pragmatism has obscured the original curriculum objectives of peer assessment. A renewed emphasis on student learning is needed to re-engage students in rating exercises. The origin of students' involvement in assessment is self-assessment. Self-assessment was initially motivated by a curiosity of whether students could assess their own work but eventually evolved into a learning goal – graduates need to assess their own performance without the help of an instructor in vocational settings (Falchikov & Boud, 1989). Students have since been further empowered in the assessment process, who would judge not only their own work but also the work of their peers. In a meta-analysis by Falchikov and Goldfinch (2000), the stated objective of peer assessment was to promote active and collaborative learning (Piaget, 1971). In recent literature, collaborative learning is framed as teamwork with real-world benefits of communication, problem-solving, leadership, and self-management (Rafiq & Fullerton, 1996; Johnston & Miles, 2004; Spatar et al., 2015).

The present paper moves beyond collaborative learning and teamwork, to report on an unexamined curriculum benefit. For social sciences, data processing skills are curricular goals because graduates are expected to know how to distil systematic information from open-ended interviews, field observation notes, or opinion polls (; Lejeune, 2001; Aiken, West, & Millsap, 2009; Bandalos & Kopp, 2012). While each discipline may have its own version of learning goals, they resemble the guideline of the American Psychological Association (APA, 2013), which notes students ought to 'collect, analyse, interpret, and report data' (sec. 2.4 A) and 'design and adopt high quality measurement strategies that enhance reliability and validity' (sec. 2.4 E). Common measurement strategies in psychology include interviews, self-reports, ratings by others, self-observation, direct observation, and archival methods (Aiken et al., 2009).

However, past research has shown a curricular shortfall on the coverage of measurement not only in psychology but also in management, education and a number of other degree programs (Patelis, Kolen, & Parshall, 1997; Aiken et al., 2009; Bandalos & Kopp, 2012; Aguinis & Edwards, 2014; Dahlman & Geisinger, 2015). Deficits in measurement skills can potentially dissuade college graduates from using unfamiliar measurement tools to pursue scientific inquiries or answer practical questions. For example, rating scales are by far the most popular instruments in psychological research (DeVellis, 1991; Van Acker & Theuns, 2010) in Aiken et al.'s (2009) list. In contrast, open-ended responses, observation

notes, interview transcripts, archival documents, are unpopular mostly because coding text data is a relatively rare skillset among college graduates. We believe that peer rating exercises can close the training gap by providing hands-on learning opportunities for students to gain the necessary skills to quantify text data (Hooper & Cowell, 2014). The data we chose were open-ended reading reflections from students in class. The rating skill we targeted was students' ability to validly and reliably code their peers' reading reflections according to a rubric (Landis, Swain, Friehe, & Coufal, 2007).

Once rating skill found an explicit place in the curriculum, its training should be intentional and its outcome should be assessed (James, 2014). We review the existing literature on peer rating system to show that the typical approach to examine the performance of an entire group of raters is an inadequate tool to assess, let alone to provide feedback to individual raters. We modify the existing peer rating system to develop an individualized marker for rating competence. We show that one round of feedback paired with a learning opportunity mid-semester could make rating skills an intentional pursuit among students.

Shortcomings of the Existing Peer Rating System

Previous studies have documented the success of peer rating system based on the convergent validity between novice ratings and instructor ratings. The general practice is to aggregate students' ratings into a student average and then correlate it with the instructor's rating. In one study, Smith (1990) asked students in an Advanced General Psychology class to rate each other's debate performance according to 10 criteria ranging from preparation to delivery. When the peer ratings were compared with the instructor ratings, the correlations were found to be significant in 8 out of the 10 criteria. In another study, Sadler and Good (2006) investigated the peer ratings by seventh-graders on a test with a range of questions from fill-in-the-blank to short answers. With the help of a scoring rubric, a very high correlation in the .90 range was obtained between peer and teacher grading. The students, however, deviated from their teacher by awarding lower scores to the best student work. A limitation of the analytic approach of averaging across peer ratings is its inability to discriminate high-performing raters from under-performing raters. In this case, an instructor may form an erroneous impression that all students were unfairly harsh when rating high quality work and therefore provide inaccurate feedback to student raters.

A meta-analysis by Falchikov and Goldfinch (2000) further illustrates the limitations of treating raters as if they were a homogenous group. Specifically, peer assessments were examined in three settings in higher education: professional practice (e.g. clinical skills, teacher performance), academic products (e.g. essays, examinations) and academic processes (e.g. oral presentation). Even though the mean correlation between peer and faculty assessments for academic products was .75, the overall range of Pearson *r* was between .14 and .99. A correlation as low as .14 has little diagnostic value. A lack of correlation implies an unacceptable level of discrepancy between the average peer rating and the instructor rating, but the instructor cannot assume that every peer rater is equally discrepant unless he/she knows that all peer ratings clustered around the peer mean. At a minimum, an instructor should examine the variability among peer ratings before he/she can start diagnosing the problem in a group of raters.

To approach the topic of variability, Zhang, Johnston and Kilic (2008) shifted their focus from the validity to the reliability of peer ratings. In two studies with a similar design, the authors found that peer ratings had a high Cronbach's alpha in the range of .70s and .80s in one study but a low Cronbach's alpha in the range of .36 to .63 in another study. The way the authors arrived at the alpha values was by calculating a rater error effect – whether a score was a product of the rater (incorrect source) or a product of the ratee (correct source). In the first study, the rater error effect was weak and never accounted for more than 12% of the total variance. In the second study, the rater error effect was strong, accounting for more than 30% of the total variance. The authors conceded that the discrepant rater error effect between the two studies may have resulted from the various backgrounds of the raters or perhaps the various training received before the rating exercise. The shift to reliability by Zhang et al. (2008) confirmed our suspicion that some groups of raters have a high degree of disagreement. However, the reliability analysis in this case was conducted on a group level and thus offered little individualized information to identify under-performing raters (or high-performing raters), let alone to help them improve (or cement) their rating skills.

Rating Skill as a Focus of the Modified Peer Rating System

For a peer rating system to serve the curriculum goal in quantitative training, an instructor needs some ways to know the skill level of individual raters and then provide tailored feedback. Any modification to peer rating system would have to allow each rater (including the instructor) remain as individuals. The granularity lends itself to the analysis of intercoder agreement among a subset of raters. If the subsets are chosen systematically, the incremental change in intercoder agreement would offer information on the quality of contribution by the dispensed rater.

A well-known index of this kind is Cronbach's alpha – a coefficient of reliability commonly used to calculate the degree to which a set of items measures an underlying construct. Our method draws upon the basic concept and the common applications of Cronbach's alpha in social sciences. Students and instructors are accustomed to thinking about the Cronbach's alpha as a useful measure of inter-item reliability for assessing the convergence among multiple items in a scale. In the present case, it is helpful to think of each rater as a single item on a questionnaire. Our method is consistent with the general practice in behavioural coding where intercoder reliability is calculated by treating each coder as if he/she were an item on a questionnaire (Aslan & Cheung-Blunden, 2012). The equivalence between a series of items in a scale and a set of ratings made by different judges was explained in detail by Bravo and Potvin (1991).

Cronbach's alpha is an index of the intercorrelation amongst a set of items and a value over .70 marks an acceptable range of variability among the items that still manage to tap into a single construct (DeVellis, 1991). The same standard can be applied to determine whether the convergence/divergence among a group of raters is at an acceptable level. A participant should leave slightly different impressions on multiple coders in the same way that the participant would respond slightly differently to multiple items on a questionnaire. However, if a coder incorporates too much bias into their observations (i.e., rater error effect), his/her observation would severely depart from the group consensus. Too many coders acting in

this manner would result in a large rater error effect, i.e. a low Cronbach's alpha. If high reliability is achieved, then students' subjective judgments were not wildly inconsistent and they were capable of evaluating the quality of each other's work based on a rubric.

The advantage of our modified peer rating system over the existing system is the flexibility of assessing the competence of student raters on a class level as well as the student's competence as an individual. First, Cronbach's alpha is calculated for the entire group by including the contributions from all raters and the traditional .70 cut-off is used as a standard for group consensus. Second, analogous to using 'Cronbach's alpha if item deleted' as a way to determine the quality of an item on a scale, the quality of a particular rater is determined by calculating Cronbach's alpha after removing all his/her contributions. A boost in the alpha value occurs if the contributions from an under-performing rater were removed; conversely, a drop in the alpha value happens if the contributions from a high-performing rater were removed.

The Present Study

In this paper, we describe a modified peer rating system and test our modification in five small Advanced Research Methods (ARM) classes. Students started by following the convention of the existing peer rating system. For eight reading assignments in the semester, students wrote reflections online, discussed reading material in class and rated their peers' reflections. Using a pre-test and post-test design, the first four reading reflections were considered Time 1 data and the last four reading reflections were considered Time 2 data. At Time 1, the group's overall Cronbach's alpha was analysed in a similar fashion as Zhang et al. (2008) in order to investigate the convergence among raters. Unique to our modified peer rating system, the data were also analysed regarding the degree to which each rater's judgment departed from the group consensus. We hypothesized that providing the performance feedback to individual raters and allowing raters to share their skills in a mid-semester intervention could result in better skill mastery. At Time 2, a similar round of analysis was conducted to investigate the post-intervention benefits.

With Cronbach's alpha as our main index, we evaluated the mastery of rating skills in a single class setting and then across the five classes. The stand-alone reporting of a single class aimed to offer step-by-step instructions for how to implement the modified peer rating system, how to conduct the relevant statistical tests and how to communicate the performance information to a class. We chose our smallest class for this purpose because it was most at risk for falling short of the .70 cut-off of Cronbach's alpha, and also because the instructor's participation provided an opportunity to establish a sense of continuity between the existing and the modified peer rating system in the ways they approach validity. The data from the first class were then combined with the other four classes for an omnibus test. A two-level linear mixed-effects model was used to discern whether the mid-semester intervention was effective at improving rating skills across the span of a semester.

Method

Participants

Seventy-four traditional college students from five advanced research classes participated in the present study. All students were either juniors or seniors in psychology. The gender composition was 81% women and 19% men. Nearly half of the participants were Caucasian (49%), followed by Asian (31%), Hispanic (16%) and African American (4%).

Measures

Validity. Since the instructor participated in a subset of the rating exercises in the first class, the convergence between student and expert ratings was used as an index of validity. We analysed validity by 1) following the convention of existing peer rating system, which is to correlate the average peer rating with the instructor rating and 2) following the convention of modified peer rating system, which is to examine the instructor's impact on the group's inter-rater agreement.

Group reliability. Cronbach's alpha was calculated to examine the reliability of the ratings at Time 1 and then again at Time 2. An improvement in the group's reliability from Time 1 to Time 2 would signal that the class became versed at rating their peers' reflections.

Impact on reliability by individual raters. Analogous to 'Cronbach's alpha if item deleted,' Cronbach's alpha was re-calculated by excluding the contributions by each rater. In theory, one fewer rater (analogous to one fewer item in a questionnaire) would reduce reliability. However, the amount of reduction would vary depending on whether the contributions from a high-performing rater or those from an average rater were excluded. Each student received an individualized Cronbach's alpha which indicated how the group fared in reliability without him/her. The analysis was run twice in a semester, at Time 1 and Time 2 respectively. The Time 1 analysis offered tailored feedback to each rater before they receive an intervention from a class period devoted to rating skills. Both Time 1 and Time 2 analyses were used to assess student rating skills.

Procedure

The syllabus stated rating skills as a learning outcome and informed students of their dual roles in the class. As writers, they would reflect on the reading materials and post their reflections online before class. As raters, they would rate each other's reflections and submit the ratings to the instructor after class.

Early in the semester, the instructor posed the question of how to evaluate the quality of reflections. After brainstorming with the class, the instructor guided the discussion towards two criteria: good and well-written. Through this guided discussion, the classes typically come to a consensus on the operationalization of the criteria as follows. Well-written entailed a) correct English, b) sound structure and flow, c) clearly articulated main idea, and d) a concrete connection to the reading. Good was characterized as a) correct understanding of the reading, b) thoughtful analyses, and c) a novel point inspired by the reading. Without any rating experience, it was only natural for students to regard these two criteria as separate constructs. It would take actual rating experience before students could think more deeply about the relationship between the criteria. The authors suspected that well-written was a prerequisite or a subjugate construct of good but took the conservative approach of treating the criteria as separate before statistically proven otherwise. The instructor documented the operationalization of these criteria in a grading rubric and posted it online for all students to consult when they wrote and rated reflections.

Before each class meeting, students submitted their reflections online (to be viewed by the instructor only). The instructor compiled all the reflections, removed names, and replaced names with

codes. The instructor then posted this viewable document online. During class, the instructor determined whether the criterion of the day was good or well-written with a flip of a coin. After class, students read and rated their peers' reflections using the criterion of the day (without rating their own reflections) and returned their ratings to the instructor electronically. The instructor compiled all ratings, including his/her own, in an Excel file. The file's organization follows the usual file format for recording grades, where students' names are row headings and the various sources of grades are column headings. In the present case, the rows headings of the Excel sheet contained the codes of the writers. The columns headings contained the source of the ratings, by the name of the rater, the reading assignment and the rating criterion used.

At Time 1, the instructor analysed the ratings and conveyed the results to the class. Some class time was allocated to not only inform each rater about their individual contributions to the group, but also encourage raters to share their grading practices with the class. The instructor had the freedom to structure the discussion to improve rating quality. Common discussion topics were:

1. Indiscriminant scores: Some raters gave similar scores to all ratees due to a variety of reasons but a useful starting point was to discuss the overarching function of measurement. Effective measures are supposed to illuminate the differences among ratees. Thus, a rater ought to try to discriminate the quality of their peers' work by taking advantage of the spectrum of the rating scale.
2. Insufficient knowledge of the rubric: The class discussed the difficulties of using an unfamiliar rubric and having to keep the entire rubric in mind (Landis, Swain, Friehe, & Coufal, 2007). Useful solutions included spending some time to familiarize with the rubric and having a copy of the rubric at hand during rating exercises.
3. Wide interpretation of the rubric: When certain parts of the rubric enjoyed a wide interpretation, it was used as a teachable moment for test construction, item development and item evaluation (Bandalos & Kopp, 2012). The instructor explained that a rubric has a variety of interpretations just as a questionnaire has a variety of items. In practice, items become a part of the questionnaire when they represent the core construct or add meaningful variability. Many items are absent from the questionnaire because they do not have the necessary construct validity. An obscure interpretation of the rubric could very well be valid, but for a different construct. The proper way to pursue a particular interpretation in the future is to reflect on its underlying construct, draft a set of new criteria to exemplify the construct, and embark on a new round of ratings.

Results

Single Class Setting

We demonstrate how to apply the modified peer rating system in a single class setting by analysing the data from the first class. Since the instructor of this class participated in the rating exercise at Time 1, validity was analysed following the convention of the existing peer rating system (Falchikov & Goldfinch, 2000). We aggregated peer ratings into average scores for each writer and found a significant correlation between the student average

ratings and instructor ratings ($r(10)=.82$, $p=.004$) across the writers. This convergence between expert and novice raters is our attempt to connect with how peer rating system in the past approached validity.

Our second approach followed the modified peer rating system by analysing Cronbach's alpha. With the peer and instructor ratings taken together, the group reached a Cronbach's alpha of .705. Further analyses in the vein of 'Cronbach's alpha if item deleted' showed that the exclusion of the instructor ratings was the most detrimental to the group's reliability because the alpha dropped to an unacceptable value of .588 without her. None of the student raters were as influential to the group's reliability as the instructor. The most any excluded student raters could impact the alpha was to drop the value from .705 to .632. The role of the instructor in this class is similar to the role of a quintessential item in a measurement scale where the item most centrally located in a construct tends to cause the most remarkable drop in Cronbach's alpha when it is deleted. Our finding that instructor ratings were centrally located among the student ratings is an alternative approach to validity. This approach allows validity and reliability to be analysed the same way such that the instructor has a choice to participate in rating exercises without affecting the analytic method.

Having connected with the analytic method in the existing peer rating system, we focused on the peer raters in the rest of the analyses (Table 1). Under the row heading 'None' in Table 1, it can be seen that when no one was excluded, the class had a Cronbach's alpha of .588 at Time 1. We recalculated the Cronbach's alpha after removing the contributions of one rater at a time in order to gauge the quality of each rater. For example, when Rater 1 was excluded, the Cronbach's alpha at Time 1 dropped to .541. Rater 1 is considered a prudent rater because removing his/her contributions caused a drop in Cronbach's alpha. Conversely, excluding the contributions from Rater 2 boosted the group's reliability from .588 to .635. Such raters were considered low-performing because removing their contributions resulted in an increase in Cronbach's alpha. Repeating the same analysis at Time 2, the class reached a higher inter-rater reliability of .707 (Table 1**).**

Omnibus Test Across Five Classes

In order to investigate the longitudinal change of intercoder agreement from Time 1 to Time 2 across the five classes, we used the Cronbach's alpha for each rating exercise as raw data. Each Cronbach's alpha was then tagged by two attributes – whether it came from Time 1 or Time 2 and the class that rendered it. The nested data were analysed using a two-level linear mixed-effects model where *i* stands for peer rating exercise occasion, *j* for time point and *k* for class (Rabe-Hesketh, Skrondal, & Pickles, 2004).

$$y_{ijk} = \beta_1 + \beta_2 x_j + \zeta_k + \varepsilon_{ijk}$$

The aforementioned model was a starting place because it included one fixed effect and one random effect (Muth et al., 2016). Our primary interest lied in the estimate of fixed effect β_2, which in this case is the change in intercoder agreement from Time 1 to Time 2. ζ_k is the random intercept and its inclusion is necessitated by the possibility that each class has its own proclivity toward a specific agreement level. Our experience with each class confirmed that classes operated at their own level of agreement perhaps due to unmeasured class characteristics, such as the instructor or the type of reading assignments.

Our results showed that β_2 was significant at $p = .016$ ($\beta = 0.097$, $SE = 0.040$). Therefore, the average class significantly improved intercoder agreement from Time 1 to Time 2. Our results also showed that the between-class standard deviation was 0.139 ($SE = 0.049$) whereas the within-class standard deviation was 0.127 ($SE = 0.015$). The ratio of the between-cluster variance to the total variance, 0.55 in this case, is the Intraclass Correlation (ICC). In mixed models, ICC is used to not only justify for clustering but also demonstrate the effect of clustering (Rabe-Hesketh et al., 2004).

One covariate we could add to the model was practice effect. We therefore tagged each Cronbach's alpha by a reading assignment number. Our results showed that the inclusion of reading number did not change the value of the estimated residual from the previous model. Furthermore, the coefficient for reading number was not significant ($\beta = -0.002$, $SE = 0.018$, $p = .930$). Therefore, practice effect could not explain the gain in intercoder agreement.

Another covariate we ought to add to the model was rating criterion in the case that good and well-written had a different impact on the intercoder agreement. We tagged each Cronbach's alpha by the criterion used, i.e. whether the reflections were judged based on good or well-written. Our results showed that the coefficient for criterion was not significant ($\beta = -0.001$, $SE = 0.045$, $p = .991$). Therefore, the criteria of good and well-written did not have a significant impact on the intercoder agreement.

Table 1. Recalculated Cronbach's Alphas after excluding the contributions from each rater.

Rater Excluded	Time 1	Time 2
None	0.588	0.707
Rater 1	0.541^{+}	0.719^{-}
Rater 2	0.635^{-}	0.678^{+}
Rater 3	0.489^{+}	0.666^{+}
Rater 4	0.600^{-}	0.715^{-}
Rater 5	0.600^{-}	0.706^{+}
Rater 6	0.477^{+}	0.651^{+}
Rater 7	0.566^{+}	0.709^{-}
Rater 8	0.579^{+}	0.679^{+}
Rater 9	0.550^{+}	0.648^{+}
Rater 10	0.583^{+}	0.630^{+}

Note. $^{+}$ a high-performing rater who boosted group reliability, ^{n} an average rater who did not impact group reliability, $^{-}$ a under-performing rater who undermined group reliability.

Discussion

The peer rating system used here advances the quantitative literacy goals outlined in social sciences. From this perspective, peer-rating exercises are educational pursuits that are inherently meaningful to the students. Students not only meet a quantitative training requirement during university studies but also reap further benefits after graduation (Boud & Falchikov, 2006). Students in occupation-oriented majors can easily appreciate rating skills as they prepare for the workplace (U.S. Department of Education, 2012). For academic-oriented majors, who by definition have limited job prospects in their disciplines, a case for career preparation is harder to make. However, the era of big data heralds a need to process a large quantity of data in order to inform treatment decisions or service choices (Bisel, Barge, Dougherty, Lucas, & Tracy, 2014).

With a learning goal in place, we used one of the most challenging data types in our classes. Open-ended comments and reflections may be intimidating to novice judges. To examine whether the students gained rating skill overtime, we focused on the first class as a stand-alone case and then combined the data from five classes. The first class was the smallest and thus most at risk for falling short of the Cronbach's alpha cut-off. With this in mind, the instructor participated in a subset of the rating exercises to shore up reliability. The instructor's participation also presented an opportunity to demonstrate validity. Convergence between the instructor and student ratings is regarded as an index of validity. If a non-significant correlation were found between the instructor and student ratings, the case may be that the expert rater (i.e., instructor) and the novices (the student aggregate) were operationalizing the criteria differently. Therefore, a review of the grading rubric and its operationalization would be needed. When the data from the five classes were combined, the analysis showed that our students reached a greater consensus in judging their peers' work in the second half of the semester than in the first half. The longitudinal improvement in rating skill was due to the mid-semester intervention rather than practice effect.

While our findings point to an improvement in the quality of ratings due to an intervention, Zhu (1995) found an increase in the quantity of feedback to peers' writing assignments due to training. The intervention was an experience-sharing session in our longitudinal research design whereas the manipulation was a set of teacher-student conferences in Zhu's (1995) experimental design. If quantitative literacy gains momentum as an outcome of peer rating system, future studies are needed to uncover the details of how peer raters learn and what they learn. Students may respond differently to teaching modalities, such as teacher-student conference versus experience sharing, small-group versus large-group intervention. Students may gain different component skills, form a better grasp of the rating rubric to a sensitivity to the cues in the data (Cathey, 2007). The typical discussion topics were listed in the procedure section but each student could have walked away with a personal take-away message. In retrospect, we could have taken notes of the discussions and provided written qualitative feedback to each student.

Feedback and Assessment Tool

While a learning outcome is typically measured at the end of the semester, multiple points of assessment may be instituted along the way to provide feedback to the students. Whether as a feedback or an assessment tool, it is the

most effective if its derivation and meaning are straightforward to everyone involved. We chose Cronbach's alpha as our main index because it is an assessable analytical method in psychology. The method of 'Cronbach's alpha if item deleted' is particularly familiar to (aspiring) psychologists who rely on questionnaires as a measurement tool. The well-accepted .70 rule of thumb was a convenient cut-off to judge class success. One of our future ambitions is to involve students in the actual numerical analysis so they can practice their statistical training in class. Other disciplines and course formats may require a different reliability index, but the same principle applies. For example, ICC may be considered for large online classes where a flexible match between raters and ratees is desired (Luo, Robinson, & Park, 2014).

Cronbach's alpha also has strategic advantages for the present study, which involves multiple small classes. Our small class sizes (designed to enrol 10-18 students) were most at risk for falling short of the .70 cut-off. However, with the exception of one time point in the first class, our experience with each class showed that a satisfactory level of interrater reliability was well within reach. Having succeeded in small samples, our modified system ought to be applicable to large samples. Rather than large class size, the modified peer rating system should be tested with other class characteristics, such as those including online courses and non-traditional students (who fall outside of the preconceived norms for college students primarily in terms of age and work experience).

Recommendations

A few details we have gleaned from our experience with the modified system may be helpful to future applications. The instructor has the flexibility to manage his/her workload but the decision needs to take class size into consideration. The results of Study 1 showed that in a class of 10 or fewer students, the instructor may have to participate in the peer rating exercise for the class to reach an acceptable Cronbach's alpha. The results from our other classes showed that a reliability of .70 or higher should be well within reach in a larger class of 15 or more students. Even then, the instructor may choose to participate in a couple of rating exercises to establish the credibility of high-performing student raters in the class.

The instructor is in a position to manage student workload, again by considering class size. Our analyses were conducted after four reading reflections for a reasonable Cronbach's alpha. Smaller classes may compensate by including more reflections per analysis whereas larger classes can afford to conduct analyses after a single set of reflections. The instructor may reduce the number of ratings to half by deciding in class, with a flip of a coin, whether a particular set of reading reflection is rated and which rating criterion is used. Instructors should pay close attention to the level of disagreement in class discussions and avoid rating the reflections from controversial reading assignments.

References

Aguinis, H., & Edwards, J. R. (2014). Methodological wishes for the next decade and how to make wishes come true. *Journal Of Management Studies*, *51*(1), 143-174. doi:10.1111/joms.12058

Aiken, L. S., West, S. G., & Millsap, R. E. (2009). Improving training in methodology enriches the science of psychology. *American Psychologist*, *64*(1), 51-52. doi:10.1037/a0014161

American Psychological Association. (2013). *APA guidelines for the undergraduate psychology major: Version 2.0.* Washington, DC: Retrieved from http://www.apa.org/ed/precollege/undergrad/index.aspx

Aslan, S., & Cheung-Blunden, V. (2012). Where does self-control fit in the Five-Factor Model? Examining personality structure in children and adults. *Personality And Individual Differences*, *53*(5), 670-674. doi:10.1016/j.paid.2012.05.006

Bandalos, D. L., & Kopp, J. P. (2012). Teaching introductory measurement: Suggestions for what to include and how to motivate students. *Educational Measurement: Issues And Practice*, *31*(2), 8-13. doi:10.1111/j.1745-3992.2012.00229.x

Bisel, R. S., Barge, J. K., Dougherty, D. S., Lucas, K., & Tracy, S. J. (2014). A round-table discussion of 'big' data in qualitative organizational communication research. *Management Communication Quarterly*, *28*(4), 625-649. doi:10.1177/0893318914549952

Boud, D., & Falchikov, N. (2006). Aligning Assessment with Long-Term Learning. Special Issue: Learning-Oriented Assessment: Principles and Practice. *Assessment & Evaluation In Higher Education*, *31*(4), 399-413.

Bravo, G., & Potvin, L. (1991). Estimating the reliability of continuous measures with Cronbach's alpha or the intraclass correlation coefficient: Toward the integration of two traditions. *Journal of Clinical Epidemiology*, *44*(4/5), 381-390.

Cathey, C. (2007). Power of peer review: An online collaborative learning assignment in social psychology. *Teaching of Psychology*, *34*, 97–99.

Dahlman, K. A., & Geisinger, K. F. (2015). The prevalence of measurement in undergraduate psychology curricula across the United States. *Scholarship Of Teaching And Learning In Psychology*, *1*(3), 189-199. doi:10.1037/stl0000030

DeVellis, R. F. (1991). *Scale development: Theory and applications*. Newbury Park, CA: Sage Publications.

Falchikov, N., & Boud, D. (1989). Student Self-Assessment in Higher Education: A Meta-Analysis. *Review Of Educational Research*, *59*(4), 395-430.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review Of Educational Research*, *70*(3), 287-322. doi:10.2307/1170785

Hooper, J., & Cowell, R. (2014). Standards-Based Grading: History Adjusted True Score. *Educational Assessment*, *19*(1), 58-76.

James, D. (2014). Investigating the Curriculum through Assessment Practice in Higher Education: The Value of a 'Learning Cultures' Approach. *Higher Education: The International Journal Of Higher Education And Educational Planning*, *67*(2), 155-169.

Johnston, L., & Miles, L. (2004). Assessing contributions to group assignments. *Assessment & Evaluation In Higher Education*, *29*(6), 751-768. doi:10.1080/0260293042000227272

Landis, M., Swain, K. D., Friehe, M. J., & Coufal, K. L. (2007). Evaluating Critical Thinking in Class and Online: Comparison of the Newman Method and the Facione Rubric. *Communication Disorders Quarterly*, *28*(3), 135-143.

Lejeune, M. M. (2001). Measuring the Impact of Data Mining on Churn Management. *Internet Research*, *11*(5), 375-87.

Loddington, S., Pond, K., Wilkinson, N., & Willmot, P. (2009). A case study of the development of WebPA: An online peer-moderated marking tool. *British Journal Of Educational Technology*, *40*(2), 329-341. doi:10.1111/j.1467-8535.2008.00922.x

Luo, H., Robinson, A. C., & Park, J. (2014). Peer Grading in a MOOC: Reliability, Validity, and Perceived Effects. *Journal Of Asynchronous Learning Networks*, *18*(2), 5-18.

Muth, C., Bales, K. L., Hinde, K., Maninger, N., Mendoza, S. P., & Ferrer, E. (2016). Alternative Models for Small Samples in Psychological Research: Applying Linear Mixed Effects Models and Generalized Estimating Equations to Repeated Measures Data. *Educational And Psychological Measurement*, *76*(1), 64-87.

Nepal, K. P. (2012). An Approach to Assign Individual Marks from a Team Mark: The Case of Australian Grading System at Universities. *Assessment & Evaluation in Higher Education. 37*(5), 555–562.

Neus, J. L. (2011). Peer Assessment Accounting for Student Agreement. *Assessment & Evaluation in Higher Education. 36*(3), 301-314

Patelis, T., Kolen, M. J., & Parshall, C. G. (1997). Surveys of programs and employment in educational measurement. *Educational Measurement: Issues & Practice*, *16*(3), 25-27. doi:10.1111/j.1745-3992.1997.tb00597.x

Piaget, J. (1971). *Biology and knowledge: An essay on the relations between organic regulations and cognitive processes*. Oxford, England: U. Chicago Press.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized Multilevel Structural Equation Modeling. *Psychometrika*, *69*(2), 167-190.

Rafiq, Y., & Fullerton, H. (1996). Peer Assessment of Group Projects in Civil Engineering. *Assessment & Evaluation in Higher Education. 21*(1), 69-81.

Reader, W. (2007). Non-participation in seminars: Free rider avoidance and value maximisation. *Psychology Learning & Teaching*, *6*(2), 121-129. doi:10.2304/plat.2007.6.2.121

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

Smith, R. A. (1990). Are peer ratings of student debates valid?. *Teaching Of Psychology*, *17*(3), 188-189. doi:10.1207/s15328023top1703_13

Spatar, C., Penna, N., Mills, H., Kutija, V., & Cooke, M. (2015). A robust approach for mapping group marks to individual marks using peer assessment. *Assessment & Evaluation In Higher Education*, *40*(3), 371-389. doi:10.1080/02602938.2014.917270

U.S. Department of Education. (2012). Occupational and academic majors in postsecondary education. Retrieved from http://nces.ed.gov/pubs2012/2012256.pdf

Van Acker, F., & Theuns, P. (2010). On possibilities and limitations of using self-anchoring scales in web surveys. *Quality & Quantity: International Journal Of Methodology*, *44*(6), 1129-1137. doi:10.1007/s11135-009-9265-4

Zhang, B., Johnston, L., & Kilic, G. B. (2008). Assessing the reliability of self- and peer rating in student group work. *Assessment & Evaluation In Higher Education, 33*(3), 329-340. doi:10.1080/02602930701293181

Zhang, B., & Ohland, M. W. (2009). How to assign individualized scores on a group project: An empirical evaluation. *Applied Measurement In Education*, *22*(3), 290-308. doi:10.1080/08957340902984075

Zhu, W. (1995). Effects of training for peer response on students' comments and interaction. *Written Communication, 12*, 492–528.