Studies in Educational Evaluation 56 (2018) 94–101

Contents lists available at ScienceDirect

Studies in Educational Evaluation

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

Finding an optimal balance between agreement and performance in an online reciprocal peer evaluation system

Kwangsu Cho[a,*], Christian D. Schunn[b]

[a] Graduate School of Information, Yonsei University, Seoul, South Korea
[b] Learning Research and Development Center, University of Pittsburgh, PA, USA

ARTICLE INFO

Keywords:
Agreement
Performance
Reliability
Reciprocal peer evaluation
Maxima strategy
Inverted U-shaped function

ABSTRACT

Consistent with movements integrating performance evaluation and improvement, the current study examines an assumption that more peer raters will produce better results or the so-called *maxima strategy*. This study examines the maxima strategy from both the agreement and performance perspectives with the intent of examining the role of feedback information in reliability and performance through an optimal number of peer raters per student ratee. It was found that the maxima strategy works consistently from agreement perspectives, whereas the relationship between the maxima strategy and student performance improvement follows an inverted U-shaped function. Accordingly, we recommend that the number of peer raters needs to be decided according to the optimal balance between reliability and performance, which maximizes ratees' performances without sacrificing evaluation reliability.

1. Introduction

Evaluation (or assessment; ratings along various dimensions) and performance improvement (changes in performance over time) are often treated as separate issues. However, many scholars have recognized the important role of evaluation in supporting improvement (Blalock, 1999; Dochy, Segers, & Sluijsmans, 1999). This integration of evaluation and improvement is salient in collective evaluation systems, especially in the form of reciprocal peer evaluation (RPE), which has gained popularity throughout education and training (Holvoet & Renard, 2003; Kulkarni et al., 2015; Magin, 2001; Noroozi, Biemans, & Mulder, 2016). Unlike typical expert-based evaluation systems where participants receive evaluations only from experts, participants in RPE systems maximize resources by playing dual roles: peer rater and *ratee*. As a rater, each participant provides an evaluation that includes both ratings of quality and peer feedback. As a ratee, each receives ratings and feedback from peers. Thus, RPE systems allow participants to construct as well as receive evaluations.

Considering that a primary advantage of RPE systems is providing multiple peer raters, deducing the optimal number of raters warrants examination. Pragmatically, the number of raters in an RPE can vary substantially across settings. In our interdisciplinary experiences, journal, conference, and grant reviewing can often involve two to five reviewers across journals, conferences, and grant programs. Faculty tenure reviews at universities may involve five to more than 16 reviewers. In education, when RPE involves individual document submissions, three to six evaluators per document are common. But when groups of three or four students submit a document and then review individually, each document might be assessed by 12–20 raters. In sum, there can be quite wide variation in what occurs in practice.

In this study, we examine the optimal number of peer raters for effective evaluation in RPE from assessment perspectives and from performance improvement perspectives, testing the hypothesis that the optimal number of peer raters is not the same from both perspectives. Theoretically, there are several factors within both assessment and performance improvement perspectives that influence the optimal number of raters, which we unpack in the sections that follow. For the rest of this section, we briefly review empirical investigations of optimal numbers of raters and introduce factors which will influence both assessment and performance improvement perspectives.

1.1. Differentiating RPE from other evaluation contexts

Few empirical studies have systematically examined the optimal number of reviewers issue. Previous research on cooperative learning suggests the optimal rater number per ratee to be four to five, but not more than eight (Cooper et al., 1990; Feichtner & Davis, 1992; Johnson, Johnson, & Smith, 1998; Nurrenbern, 1995; Slavin, 1995; Smith, 1986). While this research provides valued insight into the issue, we hesitate to apply these findings to RPE systems. These prior studies focused on

* Corresponding author at: #210-1, New Millennium Hall, Yonsei-ro 50, Seodaemun-gu, Seoul 03722, South Korea.
E-mail addresses: kwangsu.cho@gmail.com (K. Cho), schunn@pitt.edu (C.D. Schunn).

https://doi.org/10.1016/j.stueduc.2017.12.001
Received 5 February 2017; Received in revised form 5 October 2017; Accepted 2 December 2017

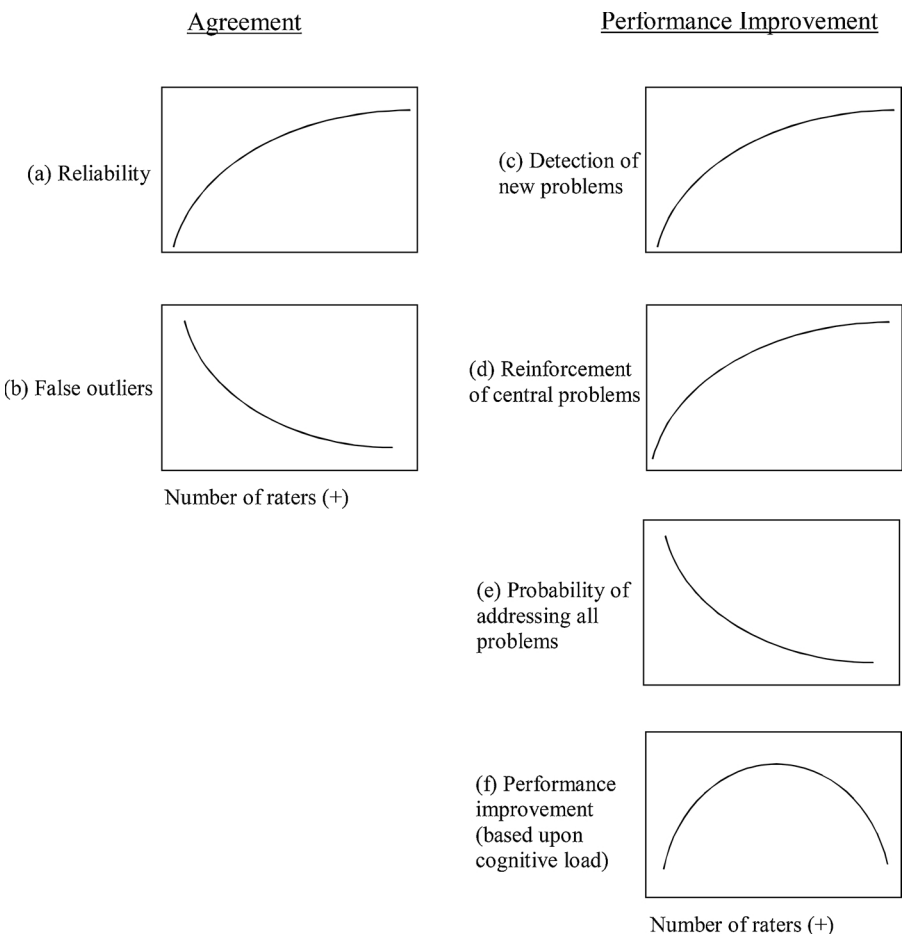

Fig. 1. Predicted effects of the number of raters on different aspects of agreement (in a and b) and performance improvement (in c–f) in RPE contexts.

synchronous or face-to-face bidirectional interactions, whereas RPE systems employ asynchronous and unidirectional interactions. Also, these prior studies focused on malfunctions in interactions due to a variety of difficulties such as scheduling meetings, free riders, and balanced participation. Regarding RPE systems, therefore, the question of the optimal number of raters is still open to examination.

In the RPE situation, what has come to be known as the *maxima strategy* has been prevalently accepted—that is, the implicit assumption that more raters will produce better results for students (e.g. Van der Lans, van de Grift, van Veen, & Fokkens-Bruinsma, 2016). The maxima strategy provides RPE systems with various advantages not afforded by traditional expert-based evaluation systems. For example, in an RPE system, students receive rich feedback without sacrificing expert resources (Noroozi et al., 2016; Rada, Michailidis, & Wang, 1994). Large numbers of peer raters provide more information about peer student ratees' problems (Wittenbaum & Stasser, 1996). Also, students generate as well as receive evaluations, which may help students actively reflect upon their own performance as well as that of others (Cho & Cho, 2013; Cho & MacArthur, 2011; Gentle, 1994; Greenwood & Levin, 1998; Patchan, Hawk, Stevens, & Schunn, 2013; Patton, 1990; Schriver, 1990). They develop crucial evaluation skills applied to their professions (Kwan & Leung, 1996) and in the process dispel negative connotations about evaluation. In addition, their participation motivates them to engage more fully with their tasks (Michaelson & Black, 1994; Patchan et al., 2013; Swaggerty & Broemmel, 2017).

Despite the several advantages, three major concerns discourage using RPE systems in practice: reliability, outliers, and performance. The reliability and outlier concerns focus entirely on ratings and are addressed from the assessment perspective (Asikainena, Virtanena, Postareffb, & Heinoa, 2014; Cho, Schunn, & Wilson, 2006; Shin, Jung, Cho, & Lee, 2012; Van der Lans et al., 2016), whereas the performance concern attends to evaluations overall (ratings and comments) and is addressed from performance improvement (or learning) perspective (Cheng, Liang, & Tsai, 2015). Interestingly, all three concerns are typically addressed by using the maxima strategy in RPE systems.

1.2. Assessment perspectives

Reliability in assessments is often measured for agreement, which concerns the degree to which different peer raters generate consistent evaluations on the same tasks (Van der Lans et al., 2016). Various studies reported medium or low reliability among peer raters (e.g. Cho & Schunn, 2007; D'Augelli, 1973; Mowl & Pain, 1995), whereas other studies report high reliability (e.g. Hughes & Large, 1993; Li et al., 2016). What these studies claim to measure as reliability is actually *mean reliability*, which is defined as expected reliability of an individual rater (Rosenthal & Rosnow, 1991) and is often calculated using a Pearson correlation (for interval scales) or Kappa (for ordinal scales). For example, when the mean reliability between two raters is 0.4, it indicates the expected reliability of either single rater, not that of the combined raters. Intuitively, multiple raters are used to balance out measurement noise from individual raters to produce a more reliable overall evaluation estimate, similar to adding more test items to make an overall test more reliable. Therefore, what this study needs to know is the reliability of the combined score across raters, known as *effective reliability* (Cho et al., 2006; Rosenthal & Rosnow, 1991). To compute this reliability, one can use the following formula adapted from the *Spearman-Brown formula* (Rosenthal & Rosnow, 1991):

$$R = \frac{n\bar{r}}{1 + (n-1)\bar{r}}$$

where R is the effective reliability coefficient, n is the number of raters,

and $\bar{r}$ is the mean reliability among raters; it sometimes also called ICC(c,k) or the one-way random average intraclass correlation. This equation formally captures the relationship between number of raters, single-rater reliability, and effective reliability: effective reliability is an increasing function as the number of raters increases (see Fig. 1a), as well as when the single-rater reliability increases. But note there is an asymptotic relationship embodied in the equation: the additional advantage of each extra rater becomes smaller and smaller.

Another issue discouraging RPE system use is the presence of outliers. The number of evaluators could influence the effect of outliers in two ways: absolute (connected to bias and reliability) and perceived elements (connected to evidence discounting). First outliers can be due to evaluation biases, particularly when participants' identities are disclosed. Biases cause unfair peer evaluations (Bence & Oppenheim, 2004; Michaelson & Black, 1994; Park & Cho, 2016) to which ratees are generally sensitized (Michaelson & Black, 1994). Peer evaluations have been found to be biased by factors such as ratees' gender (Falchikov & Magin, 1997), personal knowledge (Cooper, 1981), and appearances (Oppler, Campbell, Pulakos, & Borman, 1992), as well as their relationship with the rater (Kingstrom & Mainstone, 1985). These concerns can be addressed mainly by practicing anonymity among participants, but the most effective remedy to bias is instituting the maxima strategy (Rosenthal & Rosnow, 1991). In essence, the curve shown in Fig. 1a also addresses the minimizing of individual rater biases/increasing reliability through increasing numbers of raters.

High levels of inter-rater conflict between raters is also challenging to producing a clear overall assessment. On the one hand, one rater might be an outlier that requires excising from the overall assessment. On the other hand, the disagreeing rater might have been the only one to capture an important aspect of quality (Shin et al., 2012). Among a small number of raters, there is a reasonable chance that one of the reasonable evaluations will appear (falsely) as an outlier, which will cause the ratee to ignore (erroneously) the evaluation. In general, humans discount evidence that is counter to their beliefs (Koslowski, 1996), and this discounting sharply increases when they see noise in the data (Penner & Klahr, 1996). Consider the case in which a ratee receives five evaluations rated as follows: 4, 4, 5, 6, and 7. Here there are no outliers since the central tendency of 5.2 is well supported by all points and their general variability. Suppose, however, the ratee only received three of those five evaluations: 4, 6, and 7. The mean of 5.7 is still very close to the original mean of 5.2, but the ratee may now consider the score of 4 point as an outlier and thus comes to a overally positive interpretation of the feedback. However, occurrence of false outliers is bound to decrease with an increase of raters (see Fig. 1b).

1.3. Performance improvement perspective

Advocating peer evaluation as a means to improve human performance constitutes the mainstay of the current evaluation system movement to integrate assessment with training (Kulkarni et al., 2015; Patchan et al., 2013). Research indicates that peer evaluation (compared to expert evaluation) may have equivalent or superior effects on peer performance (Asikainena et al., 2014; Cho & Schunn, 2007; Cho et al., 2006; Hinds, Patterson, & Pfeffer, 2001; Hughes & Large, 1993; McIsaac & Sepe, 1997; Stefani, 1994; Wik, Brennan, & Braslow, 1995). For example, when Hinds et al. (2001) asked domain experts and novices to instruct beginners (junior and senior humanities majors) on electronic-circuit activity, the novice-instructed students showed fewer errors than did the expert-instructed students. Peer ratees may benefit from common ground or mutual knowledge (Clark & Brennan, 1991) based upon similar knowledge level between peer raters and ratees (Damon & Phelps, 1989; Rogoff, 1998).

It is commonly assumed that the maxima strategy could augment this favorable effect (Bratton & Gold, 2003). Some experimental investigations comparing one to five rates have found that peer ratees benefit from more feedback (Cho & MacArthur, 2011; Cho & Schunn,2007). Other studies have found a positive correlation between the amount of feedback and performance (MacDonald, Mullin, & Wilder, 2003), peer ratees' perception of more feedback being more helpful (Finn, 1997), benefits of exposure to common ideas across evaluations (Brinko, 1993) and benefits of different perspectives across evaluations (Cohen, 1994; Damon, 1984; Dillenbourg et al., 1996). However, inherent conflict across perspectives are not necessarily beneficial to performance if it leads the ratee to discount the conflicting feedback.

Concerning the impact of the number of raters on performance improvement, different predictions are made by different theories. Herein we discuss predictions based on theories of detection, reinforcement, threshold, and cognitive overload. According to detection theory, more raters tend to find more problems (Borman, 1974; Henderson, 1984). Assuming that ratees improve performance by fixing problems in their task, they need to detect and fix as many problems as possible. Hence, employing the maxima strategy should function most effectively by identifying a greater number of problems to be resolved. However, due to a limit on the number of problems to be found, a linear relationship between the number of raters and the number of detected problems is not expected. Thus, the number of detected problems will increase only up to a certain number of raters, after which diminishing returns will indicate new problems have been exhausted. Therefore, this theory predicts a curve increasing to an asymptote (see Fig. 1c).

Reinforcement theory (Annett, 1969; Deterline, 1962) focuses on the redundancy of problems detected among raters to make a similar prediction. By contrast to the detection theory, this theory emphasizes that ratees improve performance by focusing effort only on problems recurrently mentioned by raters, in part by filtering out incorrect or inappropriate idiosyncratic feedback and in part by shifting attention to the most problematic feedback (Anderson, 1982; Annett, 1969). Therefore, it is expected that the number of problems found in multiple evaluations will grow as a function of the number of raters and that at a certain point, similar to detection theory, the rate of change in the number of recurrent problems could be diminishing or asymptotic (see Fig. 1d).

Threshold theory (e.g. Bernardin & Beatty, 1984) assumes that more raters collectively create evaluations which are more difficult for ratees to satisfy, either by detecting more problems or by some reviewers having more stringent standards. Thus, ratees might improve performance to the degree that they satisfy serious concerns set by all raters. However, it may be too hard to satisfy all important concerns by many raters, and thus there might be a limit to the number of raters that can be successfully satisfied. Overall, this theory predicts that the probability of satisfying all serious rater concerns is a decreasing function of performance across raters (see Fig. 1e). As a result, ratees may give up on making revisions (and thus improving performance) if the probability of satisfactorily addressing all important issues is too small.

Finally, cognitive load theory (Sweller, 1988) suggests that performance follows an inverted-U shape as a function of the number of raters (see Fig. 1f). According to the theory, if ratees attempt to process many received evaluations in working memory, they will struggle because working memory is limited in capacity and duration (Baddeley, 2002). Faced with excessive information, novices or those who have not yet developed complex schemata will experience loss of information from working memory, either parts of the received feedback or their revision plans. In particular, when working memory load from feedback is kept within capacity limits, optimal learning and performance may occur because unused working memory capacity can support the learning processes (Sweller, 1988). By contrast, when mental workload exceeds working memory capacity, learning and performance can be undermined because evaluation information is not properly processed (Baddeley, 2002; Mayer & Moreno, 2003). Therefore, ratees' performance may improve only when assessed by a limited number of raters that, once exceeded, risks cognitive overload and impairs learning and performance. Indeed Patchan, Schunn, and Correnti (2016) found that

feedback was less likely to be implemented in the context of large amounts of other feedback. However, it may be that learners can incrementally process feedback or use external memory aids, and thus overcome work memory limitations. In support of this, Patchan, Schunn, and Correnti, 2016 found no effect of amount of feedback on revision quality.

In sum, there are opposing predictions about how the number of raters will influence ratee performance based on different aspects that underlie reliable evaluation and performance improvement. The goal of this study is to determine the optimal number of peer raters to address measurement and performance concerns in RPE systems. As shown in Fig. 1, agreement theories support the maxima strategy: higher effective reliability and lower rate of false outliers as number of raters increases. By contrast, performance theories predict different patterns of performance. The detection and reinforcement theories support maxima strategy use, whereas the threshold and the cognitive overload theories caution against maxima strategy abuse in that ratees may give up on revisions or have no spare cognitive capacity for learning processes when the amount of feedback received is too large. Therefore, this study examines the impact of the number of peer raters on agreement and performance concerns to find the optimal number of peer raters in RPE systems. Agreement is expected to be asymptotically maximized with increasing number of raters, whereas performance may show actual reductions after a peak performance level is reached with an intermediate number of raters. However, since agreement may also affect performance, the point of diminishing returns may in fact be at a very large number of raters, larger than would typically occur in most situations. On a related point, it is important to note that the optimal number may vary with setting. This study focuses on the case of peers with relatively low domain knowledge and medium task ability because that case is very common in training situations.

2. Method

2.1. Participants

Participants included 248 undergraduate students from three cognitive psychology courses for nonmajors at a research university in the US. As non-majors, they came from diverse backgrounds and were relative novices to the domain of psychology in general and cognitive psychology in particular. They participated in RPE activities for course credit. Each participant played the dual role of rater and ratee. As a ratee, each participant wrote a document, received evaluation feedback (in the form of both ratings and comments), and revised the document based on the feedback. As a rater, each participant evaluated six documents in both their first and final versions. A total of 2490 evaluations by 248 students of 496 documents were analyzed, with 16% attrition reflecting some raters failing to complete some or all of their assigned rating tasks. Because we controlled only the number of ratees per rater but the number of raters per ratee was randomly assigned, participants received evaluations from different numbers of peer raters, either because a reviewer failed to complete an assigned task (producing cases of fewer than 6 reviews per document) or because some participants failed to submit a document but did complete reviews (producing cases with more than 6 reviews per document). Among the 248 participants, three participants received peer evaluations from two peers, 25 from three, 90 from four, 51 from five, 27 from six, 29 from seven, and 23 from eight—thus the range from three to eight raters is sufficiently well sampled. Rater-ratee pairings were assigned randomly and blindly. All participants used SWoRD (Scaffolded Writing and Rewriting in the Discipline), a web-based RPE system which used the mechanism of RPEs (Cho & Schunn, 2007), described in the Interface and Procedure section.

2.2. Document task

The task assigned to participants was to write a document within one content area of cognitive psychology. Participants received various writing topics, which they tailored to their content areas. The mean document size was 5.9 pages (SD = 1.6), 18.5 paragraphs (SD = 12.3), and 1447 words (SD = 407). No differences in document size were found across content areas.

The participants evaluated each draft along three dimensions: *flow*, *logic*, and *insight*. Consequently, performance was defined as the average of the three dimensions, and performance improvement was defined as the improvement from first to second draft. For guidance, participants received instruction on important features of effective evaluation in each of the dimensions. The *flow* dimension, the most basic level, considered the extent to which a document involved a lack of faults or problems in prose flow (i.e., being able to follow the arguments and general flow of the paper). The *logic* dimension examined the extent to which a document presented a strong argument with supporting facts (regardless of whether presented in a confusing order). The *insight* dimension accounted for the extent to which a document contributed new knowledge to the content area. Raters assessed each document both qualitatively and quantitatively in each of the three dimensions. They generated both written comments and numeric ratings on a 7-point scale from *disastrous (1)* to *excellent (7)*.

2.3. Interface and procedure

In the evaluation process, all participants used the SWoRD system. An overview of the system was presented in class, but no special training was provided on the evaluation task. Ratees electronically submitted their documents to SWoRD, which distributed the documents to randomly selected sets of peer raters. Raters reviewed the documents, generated written comments and numeric ratings, and submitted results to the system. Having received the results from the system, ratees revised their documents accordingly, submitted revisions, and in turn *back-evaluated* the raters' feedback in terms of the feedback's effectiveness according to a 5-point scale from *not helpful at all (1)* to *very helpful (5)*. The process involved a second round: the same set of raters reviewed the revisions and submitting another set of comments and ratings, and ratees receiving and back-evaluating the second round of feedback. All proceedings were conducted anonymously.

2.4. Analyses

Analyses examined the predicted relationships with number of evaluators on assessment and performance dimensions. Since number of evaluators varied naturally rather than through experimental variation, follow-up analyses verified that other aspects of the evaluation process did not systematically vary by number of evaluators. In particular, the number of evaluators was assumed to not be influenced by the quality of the documents or the characteristics of the evaluators. This assumption should have been met given the random assignment of raters to ratees.

For assessment purposes, measures of effective reliability and outlier frequency are computed, and then statistically analyzed as a function of number of raters. A follow-up analysis verifies that standard deviations (actual variability among raters) is not varying as a function of number of raters.

For performance purposes, improvements from first to second drafts were analyzed as a function of number of raters. A secondary analysis tested whether perceived helpfulness of provided feedback decreased as a function of number of raters.

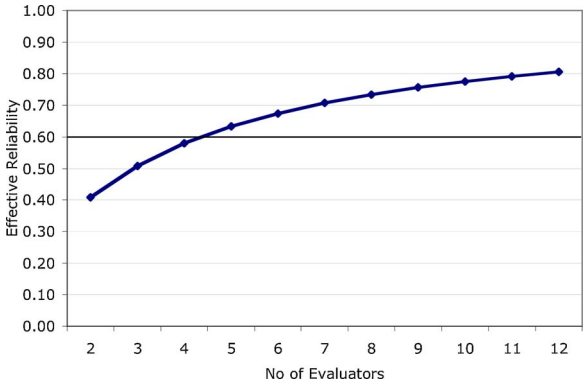

Fig. 2. Estimated effective reliability as a function of numbers of raters.

3. Results & discussion

3.1. Reliability

To determine the optimal number of raters in terms of reliability, the mean reliability among individual raters was first computed to be r = 0.26, using an estimation procedure developed by Cho and Schunn (2007) that is needed when relatively few entries in the documents by raters matrix is filled. This reliability value is similar to observed journal reviewer reliability (Marsh & Ball, 1989), and within the range and only slightly below mean values found for 12 undergraduate and 4 graduate courses (Cho, Schunn, & Wilson, 2006). Using the Spearman-Brown formula, effective reliabilities as a function of the number of raters were estimated as shown in Fig. 2. Based on 0.60 as a generally accepted value for consistency (Rosenthal & Rosnow, 1991), it can be concluded that a group of raters should consist of at least four or more raters in this context to achieve the standard of 0.60.

3.2. Outliers

We estimated an upper bound on the true rate of outliers using a Grubbs test for outliers (Grubbs & Beck, 1972) in the maximal data case (n = 8 raters). The Grubbs test is applied to a set of ratings on a document and calculates an outlier present/absent binary decision based on the standardized difference between each evaluation of a document and the mean of all evaluations of that document. Thus, it is computed based on the distance between each rating and all ratings for a document divided by their standard deviation. An estimated true outlier rate across documents was found to be relatively low (M = 0.25 out of 8 ratings, SD = 0.50) or approximately 3% of ratings; of course with many ratings, ratees can expect a higher rate, such as 1 in 4 ratees who received 8 ratings having at least one outlier. We then conducted the Grubbs test on all the documents to see how many outliers were found as a function of raters (see Fig. 3). Because the number of detected outliers is relatively high except for n = 7 and 8, we can assume the majority of these cases are false outliers. For example, based on 3% true outliers, there should have only been a mean of 0.09 outliers for N = 3, 0.12 outliers for N = 4, and 0.15 outliers for N = 5 (i.e., a low linear increase of expected number of true outliers = # of ratings x 0.03 outliers/rating). Thus, the mean of 1.4 outliers out of N = 3 raters implies a false outlier rate of 1.3. As expected, Fig. 3 shows that the occurrence of false outliers decreases as the number of raters increases, generally following an exponential decline (i.e., asymptotically approaching true outlier rates), rather than the small linear increase that true outliers should show.

The prior analysis assumes that the observed variability in ratings was not systematically different as the number of raters increased. Since raters are randomly assigned to ratees, there should be no bias. However, perhaps quality of documents influenced whether raters

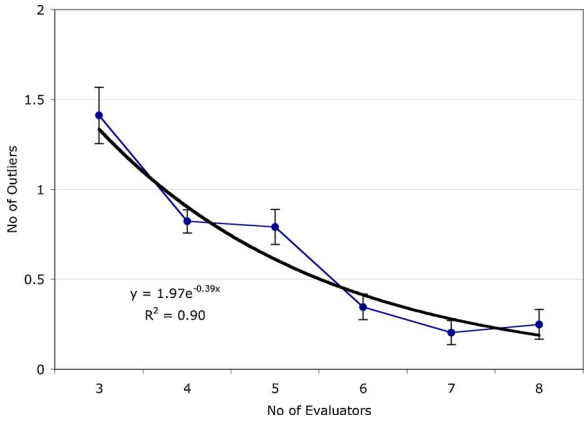

Fig. 3. Average numbers of outliers with standard errors of mean.

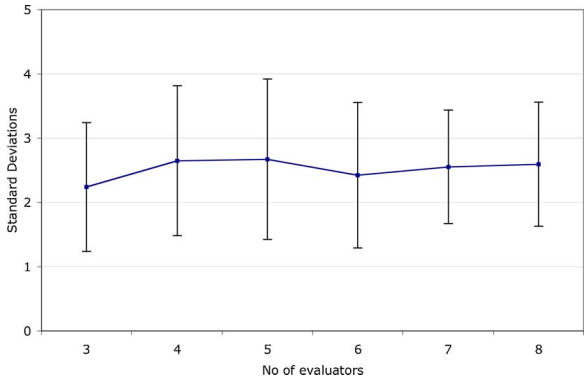

Fig. 4. Average standard deviations with standard deviation bars as a function of number of evaluators.

completed their rating tasks, which would then change the variability. To test this assumption of equivalent variance, a one-way ANOVA with number of raters as a between-subjects variable was performed on standard deviations across raters as a dependent variable (see Fig. 4). No significant difference was found on the size of standard deviations, $F(5, 295) = 0.84$, $p = 0.52$. Therefore, the observed high rate of statistical outliers for lower numbers of raters was not due to the documents receiving fewer ratings being systematically noiser in evaluations.

3.3. Performance

Concerning the performance predictions, task quality improvement was computed and analyzed as a function of the number of raters. Task quality improvement was defined as the difference between the quality of the first draft and the final draft. Fig. 5 shows that the mean performance improvement follows an inverted U-shape as a function of number of raters. Initially, as the number of raters increases, performance likewise increases, peaks around six raters, and then decreases, a trend that is consistent with cognitive load theory. Indeed, the performance levels at N = 7 and N = 8 were significantly below the performance level at N = 5 ($ps < 0.01$). A second-degree polynomial provided a good fit to the data ($R^2 = 0.83$). The peak performance level for this function is 5.8 raters. Although extrapolation is always risky, the function suggests that with 11 or more raters, evaluatees' 2nd draft performance would suffer rather than improve.

However, the performance decrement after 5.8 raters in the inverted-U shaped performance could be the result of ratees perceiving the quality of feedback as lower when being presented by evaluations from more raters. Thus, unlike the assumption that larger numbers of raters provide more information, the actual amount of evaluation information perceived to be useful could be low (Marwell & Oliver,

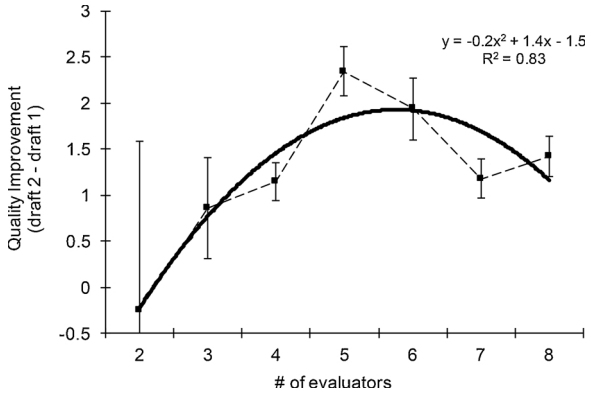

Fig. 5. Mean (with SE bars) first to final draft performance improvements as a function of number of evaluators along with a best-fitting polynomial.

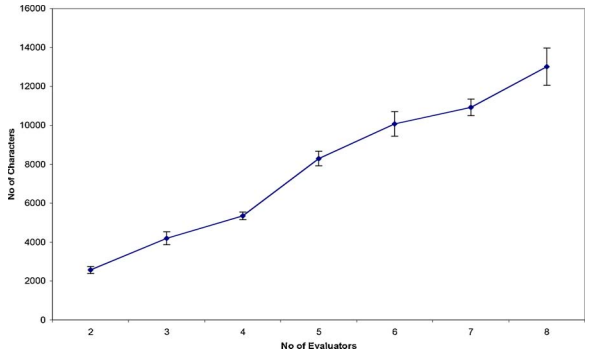

Fig. 6. Mean amount of accepted feedback length (with standard error bars) as a function of number of evaluators.

1993). To test this under-productivity possibility, we computed the amount of feedback that the participants accepted and processed, which was measured by the number of characters in the written comments that the ratees rated as *helpful* (3 or higher) for revising their documents. Excluded was feedback rated *less helpful* (2 or lower). Fig. 6 shows that accepted feedback (F) increases as a function of number of raters (n), $F = 1769.3n - 1072.5$, $R^2 = 0.98$, $p < 0.001$. Consequently, the hypothesis involving decreasing amount of feedback perceived to be useful after 5.8 raters is rejected.

3.4. Evaluation time on task

The amount of evaluation time spent by each rater contextualized our argument. Although task evaluation time is not a focus in this paper, it should be noted that the number of raters actually used in evaluation is frequently based on the time on task each evaluator is able to invest. Instructors in a classroom have to make decisions about how learning tasks should be designed, which takes into account the amount of time each learning task takes. In particular, whether a peer evaluation task will take 30 min or 10 h will influence how many reviews are assigned to each reviewer. In our study, the raters were asked to report how much time they spent doing the evaluations. It was found that raters spent an average of 34.0 min (SD = 17.8) reading each document and 29.6 min (SD = 19.0) generating each evaluation. Therefore, evaluation time averaged about one hour per document, or about 12 h total evaluating six first and final peer documents. Because this involves a substantial level of resource commitment, the dimension level of returns with larger numbers of raters is especially important.

4. General discussion

In this study, we considered the optimal number of raters from both agreement and performance perspectives by taking into account reliability, outliers, and performance issues. The reliability data show that at least four raters were necessary in this context to produce acceptable effective reliabilities (i.e., at least 0.6), which is interestingly similar to the estimate of ideal numbers in cooperative learning (e.g., Johnson et al., 1998; Nurrenbern, 1995; Slavin, 1995). Analyses of outliers also showed four or more raters have less than one falsely identified outlying rater on average, and six or more raters produced outlier detection rates than were similar to true outlier rates (of 3% of ratings). Finally, performance improvement data indicate six raters to be optimal. Looking across these results, it can be concluded that five or six raters produces near optimal performance overall for agreement and performance in an RPE system, and little value would be gained from increasing the number of raters beyond that point.

Consistent with the integration of evaluation and performance (Blalock, 1999) the current research tries to find a balance between agreement and performance in terms of the number of raters in the RPE context. The multiple rater approach as a key characteristic in RPE systems features the maxima strategy. However, this strategy was considered only from an agreement perspective and not from a performance improvement perspective.

How would the current results generalize to other evaluation contexts? Given the sampled population and task (relative novices coming from diverse backgrounds and given detailed rating and commenting guidelines), the current results are more relevant to trainees with varying levels of training in the domain, diverse backgrounds, and primarily just rubrics provided as support rather than extensive training. Peer evaluations will often have this flavor, as well 180 and 360 performance evaluations in companies, in which evaluators not usually given specialized training in evaluation.

This paper contributes to the field by addressing peer evaluation not just from an agreement perspective, but also from a performance improvement perspective. According to previous research, when using a summative or average approach, the agreement between peer raters and expert raters was satisfactory (Freeman, 1995; Hughes & Large, 1993; Stefani, 1994). It is consistent with the agreement assumption that with more raters, an estimate will be closer to the true value of what is being estimated. However, this study showed there are trade-offs between agreement and performance when both perspectives are jointly considered. As predicted, the maxima strategy works consistently with agreement theories; thus, more raters improve reliability and remove negative effects such as false outliers, although asymptotically, with strongly diminishing returns after six reviewers. By contrast, performance improvement follows an inverted U-shaped trajectory as a function of the number of raters, consistent with the cognitive load theory (Sweller, 1988). Unlike a common assumption that more evaluations augment a favorable peer evaluation effect, we found that too much feedback may hamper ratees' performance. Similarly, Cheng et al. (2015) and Goodman and Wood (2004) found that increasing the amount of specific feedback might actually hurt learning.

The results of this study have implications for RPE. The maxima strategy using multiple raters is considered to generate more accurate or acceptable evaluations (Latham & Wexley, 1982), an assumption grounded in assessment theory. At the same time, it is regarded as critical to provide performance feedback. Therefore, it seems the maxima strategy is implicitly accepted as the means of increasing both the agreement of evaluations and the performance of receivers. In addition, the advances in available information technology greatly enhance RPE efficiency not only by providing ratees with rich feedback on their performance but also by facilitating the involvement of greater numbers of raters. The results of this study, however, caution about this very abuse of the maxima strategy promoted by agreement and performance theories as well as technology: too many raters might simply overload ratees with information (Jones, Ravid, & Rafaeli, 2004). Therefore, when deciding the number of raters, especially for performance, the optimal number of raters should be considered. In addition,

the results also redefine the role of information systems to be one of metering the amount of evaluation information delivered to ratees by selecting information of high quality while removing that of low quality.

Finally, although this study did not provide much analysis on the cost of using multiple raters, its importance should not be overlooked—raters spent a large amount of time on completing these reviews and the asymptotic gains from additional reviewers should be seriously considered even when attending to only overall assessment reliability. In addition, this study did not focus on the difficulty or complexity of tasks either, which substantially influences the efforts required to complete reviews as well as potentially influencing the other factors in agreement and performance improvement. We agree that task characteristics may define various aspects of evaluation and its effectiveness. However, how specific tasks influence evaluation effectiveness is not yet well understood (Hattie & Timperley 2007; Kluger & DeNisi, 1996) and is recommended for further study. In reporting empirical findings that six raters constitute an optimal number in an RPE system, this study contributes critical knowledge to research practices on evaluation feedback. Although many studies show that evaluation feedback improves task performance (e.g. Earley, Northcraft, & Lituchy, 1990; Ilgen 1999), a considerable amount of research also reports that feedback does not automatically improve performance and in fact deteriorates it (e.g., Cheng et al., 2015; Cho & Cho, 2011; Goodman & Wood 2004; Hattie & Timperley 2007). Given mixed results concerning the impact of evaluation feedback, the number of raters or the amount of feedback could play the role of an independent or mediating control variable, making it possible to refine existing theories and develop new ones.

References

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*(4), 369.

Annett, J. (1969). *Feedback and human behavior.* Harmondsworth, England: Penguin.

Asikainena, H., Virtanena, V., Postareffb, L., & Heinoa, P. (2014). The validity and students' experiences of peer assessment in a large introductory class of gene technology. *Studies in Educational Evaluation, 43*, 197–205.

Baddeley, A. D. (2002). Is working memory still working? *European Psychologist, 7*(2), 85.

Bence, V., & Oppenheim, C. (2004). The influence of peer review on the research assessment exercise. *Journal of Information Science, 30*(4), 347–368.

Bernardin, H. J., & Beatty, R. W. (1984). *Performance appraisal: Assessing human behavior at work.* Kent Pub. Co.

Blalock, A. B. (1999). Evaluation research and the performance management movement from estrangement to useful integration? *Evaluation, 5*(2), 117–149.

Borman, W. C. (1974). The rating of individuals in organizations: An alternate approach. *Organizational Behavior and Human Performance, 12*(1), 105–124.

Bratton, J., & Gold, J. (2003). *Human resource management* (3rd ed.). Palgrave Macmillan.

Brinko, K. T. (1993). The practice of giving feedback to improve teaching: What is effective. *Journal of Higher Education,* 574–593.

Cheng, K. H., Liang, J. C., & Tsai, C. C. (2015). Examining the role of feedback messages in undergraduate students' writing performance during an online peer assessment activity. *The Internet and Higher Education, 25*, 78–84.

Cho, Y. H., & Cho, K. (2011). Peer reviewers learn from giving comments. *Instructional Science, 39*(5), 629–643.

Cho, K., & Cho, M. (2013). Self-regulation training on a social network system. *Social Psychology of Education, 16*, 617–634.

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of Educational Psychology, 103*(1), 73–84.

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.

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.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.). *Perspectives on socially shared cognition* (pp. 127–149). Washington, DC: American Psychological Association.

Cohen, E. (1994). *Designing group work: Strategies for homogeneous classrooms.* New York: Teachers College Press.

Cooper, J. L., Prescott, S., Cook, L., Smith, L., Mueck, R., & Cuseo, J. (1990). *Cooperative learning and college instruction: Effective use of student learning teams.* Long Beach, CA: California State University Foundation.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90*(2), 218.

D'Augelli, A. R. (1973). The assessment of interpersonal skills: A comparison of observer, peer, and self ratings. *Journal of Community Psychology,* 177–179.

Damon, W., & Phelps, E. (1989). Critical distinctions among three approaches to peer education. *International Journal of Educational Research, 13*(1), 9–19.

Damon, W. (1984). Peer education: The untapped potential. *Journal of Applied Developmental Psychology, 5*(4), 331–343.

Deterline, W. A. (1962). *An introduction to programmed instruction.* New York, NY: Prentice-Hall.

Dillenbourg, P., Baker, M. J., Blaye, A., & O'Malley, C. (1996). The evolution of research on collaborative learning. In P. Reinmann, & H. Spada (Eds.). *Learning in humans and machines* (pp. 189–205). Oxford, England: Elsevier.

Dochy, F. J. R. C., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education, 24*(3), 331–350.

Earley, P. C., Northcraft, G. B., Lee, C., & Lituchy, T. R. (1990). Impact of process and outcome feedback on the relation of goal setting to task performance. *Academy of Management Journal, 33*(1), 87–105.

Falchikov, N., & Magin, D. (1997). Detecting gender bias in peer marking of students' group process work. *Assessment & Evaluation in Higher Education, 22*(4), 385–396.

Feichtner, S. B., & Davis, E. A. (1992). Why some groups fail: A survey of students' experiences with learning groups. In A. S. Goodsell, M. R. Maher, & V. Tinto (Eds.). *Collaborative learning: A sourcebook for higher education.* National Center on Postsecondary Teaching, Learning, & Assessment, Syracuse University.

Finn, D. M. (1997). Perceptions of performance feedback received by female and male managers: A field study. *Dissertation Abstracts International Section A: Humanities & Social Sciences, 57*(January (7-A)), 3119 [US: Univ Microfilms International].

Freeman, M. (1995). Peer assessment by groups of group work. *Assessment & Evaluation in Higher Education, 20*(3), 289–300.

Gentle, C. R. (1994). Thesys: An expert system for assessing undergraduate projects. In M. Thomas, T. Sechrest, & N. Estest (Eds.). *Deciding our future: Technological imperatives for education* (pp. 1158–1160). Austin, TX: University of Texas.

Goodman, J. S., & Wood, R. E. (2004). Feedback specificity, learning opportunities, and learning. *Journal of Applied Psychology, 89*(5), 809–821.

Greenwood, D. J., & Levin, M. (1998). *Introduction to action research: Social research for social change.* Thousand Oaks, CA: SAGE.

Grubbs, F. E., & Beck, G. (1972). Extension of sample sizes and percentage points for significance tests of outlying observations. *Technometrics, 14*(4), 847–854.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Henderson, R. I. (1984). *Practical guide to performance appraisal.* Reston,Va: Reston Publishing Company.

Hinds, P. J., Patterson, M., & Pfeffer, J. (2001). Bothered by abstraction: The effect of expertise on knowledge transfer and subsequent novice performance. *Journal of Applied Psychology, 86*(6), 1232–1243.

Holvoet, N., & Renard, R. (2003). Desk screening of development projects: Is it effective? *Evaluation, 9*(2), 173–191.

Hughes, I. E., & Large, B. J. (1993). Staff and peer-group assessment of oral communication skills. *Studies in Higher Education, 18*(3), 379–385.

Ilgen, D. R. (1999). Teams embedded in organizations: Some implications. *American Psychologist, 54*(2), 129–139.

Johnson, D. W., Johnson, R. T., & Smith, K. A. (1998). *Active learning: Cooperation in the college classroom.* Edina. MN: Interaction Book Company.

Jones, Q., Ravid, G., & Rafaeli, S. (2004). Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration. *Information Systems Research, 15*(2), 194–210.

Kingstrom, P. O., & Mainstone, L. E. (1985). An investigation of the rater-ratee acquaintance and rater bias. *Academy of Management Journal, 28*(3), 641–653.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning.* MIT Press.

Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., ... Klemmer, S. R. (2015). *Peer and self- assessment in massive online classes. In Design Thinking Research.* Springer International Publishing131–168.

Kwan, K. P., & Leung, R. W. (1996). Tutor versus peer group assessment of student performance in a simulation training exercise. *Assessment & Evaluation in Higher Education, 21*(3), 205–214.

Latham, G. P., & Wexley, K. N. (1982). *Increasing productivity through performance appraisal.* Reading, MA: Addison-Wesley.

Li, H., Xiong, Y., Zang, X. L., Kornhaber, M., 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.

MacDonald, J. E., Mullin, J. E., & Wilder, D. A. (2003). Weekly feedback vs. daily feedback: An application in retail. *Journal of Organizational Behavior Management, 23*(2–3), 21–43.

Magin, D. (2001). Reciprocity as a source of bias in multiple peer assessment of group work. *Studies in Higher Education, 26*(1), 53–63.

Marsh, H. W., & Ball, S. (1989). The peer review process used to evaluate manuscripts submitted to academic journals: Interjudgmental reliability. *Journal of Experimental Education, 57*, 151–169.

Marwell, G., & Oliver, P. (1993). *The critical mass in collective action: A micro social theory.* New York, NY: Cambridge University Press.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 43–52.

McIsaac, C. M., & Sepe, J. F. (1997). Improving the writing of accounting students: A cooperative venture. *Journal of Accounting Education, 14*(4), 515–533.

Michaelson, L. K., & Black, R. H. (1994). The key to harnessing the power of small groups I: Higher education building learning teams. *Growth Partners,* 14.

Mowl, G., & Pain, R. (1995). Using self and peer assessment to improve students' essay

writing: A case study from geography. *Innovations in Education and Training International, 32*(4), 324–335.
Noroozi, O., Biemans, H., & Mulder, M. (2016). Relations between scripted online peer feedback processes and quality of written argumentative essay. *The Internet and Higher Education, 31*, 20–31.
Nurrenbern, S. (1995). *Experiences in cooperative learning: A collection for chemistry teachers*. Madison, WI: Institute for Chemical Education, University of Wisconsin Board of Regents.
Oppler, S. H., Campbell, J. P., Pulakos, E. D., & Borman, W. C. (1992). Three approaches to the investigation of subgroup bias in performance measurement: Review, results, and conclusions. *Journal of Applied Psychology, 77*(2), 201–217.
Park, J., & Cho, K. (2016). Toward the integration of peer reviewing and computational linguistics approaches. *Journal of Educational Computing Research.*. http://dx.doi.org/10.1177/0735633116656454.
Patchan, M. M., Schunn, C. D., & Correnti, R. J. (2016). The nature of feedback: How peer feedback features affect students' implementation rate and quality of revisions. *Journal of Educational Psychology, 108*(8), 1098–1120.
PPatchan, M. M., Hawk, B., Stevens, C. A., & Schunn, C. D. (2013). The effects of skill diversity on commenting and revisions. *Instructional Science, 41*(2), 381–405.
Patton, M. (1990). *Qualitative evaluation methods* (2nd ed.). Thousand Oaks, CA: SAGE.
Penner, D. E., & Klahr, D. (1996). When to trust the data: Further investigations of system error in a scientific reasoning task. *Memory & Cognition, 24*(5), 655–668.
Rada, R., Michailidis, A., & Wang, W. (1994). Collaborative hypermedia in a classroom setting. *Journal of Educational Multimedia and Hypermedia, 3*(1), 21–36.
Rogoff, B. (1998). Cognition as a collaborative process. In W. Damon, D. Kuhn, & R. S. Siegler (Eds.). *Handbook of child psychology: Vol. 2: cognition, perception & language*. New York, NY: John Wiley & Sons Inc.
Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York, NY: McGraw Hill.
Schriver, K. A. (1990). *Evaluating text quality: The continuum from text-focused to reader-focused methods. technical report No. 41*. National Center for the Study of Writing and Literacy.
Shin, H. J., Jung, H. W., Cho, K. S., & Lee, J. H. (2012). A prediction method of learning outcomes based on regression model for effective peer review learning. *Journal of Korean Institute of Intelligent Systems, 22*(5), 624–630.
Slavin, R. E. (1995). *Cooperative learning: Theory, research, and practice* (2nd ed.). Boston, MA: Allyn & Bacon.
Smith, K. A. (1986). Cooperative learning groups. In S. F. Schmoberg (Ed.). *Strategies for active teaching and learning in university classrooms*. Minneapolis, MN: Office of Educational Development Programs, University of Minnesota.
Stefani, L. A. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education, 19*(1), 69–75.
Swaggerty, E. A., & Broemmel, A. D. (2017). Authenticity, relevance, and connectedness: Graduate students' learning preferences and experiences in an online reading education course. *Internet and Higher Education, 32*, 80–86.
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.
Van der Lans, R. M., van de Grift, W. J. C. M., van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation, 50*, 88–95.
Wik, L., Brennan, R. T., & Braslow, A. (1995). A peer-training model for instruction of basic cardiac life support. *Resuscitation, 29*(2), 119–128.
Wittenbaum, G. M., & Stasser, G. (1996). Management of information in small groups. In J. L. Nye, & A. M. Brower (Eds.). *What's social about social cognition? research on socially shared cognition in small groups*. Thousand Oaks, CA: SAGE.