# Does Writing Improve as a Function of Number of Reviewers?

**Kwangsu Cho (KWANGSU@Pitt.Edu)** **Christian Schunn (SCHUNN@Pitt.Edu)**
School of Information Science Learning Research & Development Center
University of Missouri, Columbia University of Pittsburgh

**Abstract**

Evaluation management systems, especially reciprocal peer evaluation systems, often have an assumption that more reviewers will produce better results. This tendency is labeled as *the maxima strategy*. This study examines the maxima strategy from both agreement and performance perspectives with the intent of examining the role of information in reliability and performance through an optimal number of peer reviewers per evaluate in writing. It was found that the maxima strategy works consistently with agreement perspectives, whilst the relationship between the maxima strategy and performance improvement follows an inverted U-shaped function. Accordingly, we recommend that the number of reviewers needs to be decided on the optimal balance between reliability and performance, which maximizes writers' performances without sacrificing evaluation reliability.

Consistent with a recent movement of integrating evaluation assessment and training (Dochy, Segers, & Sluijsmans, 1999) with evaluation management systems, reciprocal peer evaluation (RPE) and its distinction of relying upon multiple reviewers has gained popularity throughout education and training (Magin, 2001) and in numerous organizations (Harris & Schubroeck, 1988; Illgen, 1999; Katzenbach & Smith, 1993). According to the American Society for Training and Development (ASTD), for instance, about 33% of firms utilized RPE systems in 1999, an increase from only 10% in 1997 (ASTD, 1999). Unlike typical expert-based evaluation management systems where participants receive evaluations only from experts, participants in RPE systems maximize resources by playing dual roles: reviewer and *writer*. As a reviewer, each participant provides peer feedback. As a writer, each receives feedback from peers. Thus, RPE systems allow participants to construct as well as receive evaluations.

In this study, we examine from assessment perspectives and also from performance perspectives the optimal number of peer reviewers for effective evaluation management in RPE systems. Considering that a primary advantage of RPE systems is providing multiple peer reviewers and hence more feedback, deducing the optimal number of reviewers warrants examination. However, few empirical studies have systematically examined this issue. Therefore, the question of optimal number of reviewers is still open to examination.

In this situation, prevalently accepted is what has come to be known as the *maxima strategy*, meaning the implicit assumption that more reviewers will produce better results. The maxima strategy provides RPE systems with various advantages not afforded by traditional expert-based evaluation management systems. For example, RPE system participants receive rich feedback without sacrificing expert resources (Cho & Schunn, in press). Large numbers of reviewers provide more information about writers' problems (Wittenbaum & Stasser, 1996). Also, participants generate as well as receive evaluations, which may help participants actively reflect upon their own performance as well as that of others (Schriver, 1990). They develop crucial evaluation skills applied to their professions (Oldfield & MacAlphine, 1995) and in the process dispel negative connotations about evaluation. In addition, their participation motivates them to engage more fully with their tasks (Michaelson & Black, 1984).

Despite advantages, three major concerns discourage using RPE systems in practice: reliability, outliers, and performance. The reliability and outlier concerns are addressed from the assessment perspective, while the performance concern is addressed from the learning and performance perspective. Interestingly, all three concerns are typically addressed by using the maxima strategy in RPE systems.

## Reliability Perspective

Reliability is often measured as consistency which concerns the degree to which different peer reviewers generate consistent evaluations on the same tasks. Various studies reported medium or low reliabilities among peer reviewers (e.g. Mowl & Pain, 1995), while some studies report high reliability (e.g. Hughes & Large, 1993). What these studies claim to measure as reliability is actually *mean reliability*, defined as expected reliability of an individual reviewer (Rosenthal & Rosnow, 1991). For example, when the mean reliability between two reviewers is .4, it indicates the expected reliability of either single reviewer, not that of combined reviewers. Therefore, what this study needs to know is the aggregate reliability of the total reviewers, known as *effective reliability* (Rosenthal & Rosnow, 1991). Effective reliability provides the measure of *composite* internal consistency of combined reviewers. To compute this reliability, we 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 reviewers, and $\bar{r}$ is the mean reliability among reviewers. Therefore, effective reliability will demonstrate an increasing asymptote function as the number of reviewers increases (*see* Figure 1a).

Another issue discouraging RPE system use is outliers due to evaluation biases, particularly when participants' identities are disclosed. Biases cause unfair peer evaluations (Bence & Oppenheim, 2004; Michaelson & Black, 1994) to which writers are generally sensitized (Michaelson & Black,

1994). Peer evaluations are biased by factors such as writers' gender (Falchikov & Magin, 1997), personal knowledge (Cooper, 1981), and appearances (Oppler, Campbell, Pulakos, & Borman, 1992), as well as their relationship with the reviewer (Kingstrom & Mainstone, 1985). These concerns can be addressed mainly by utilizing anonymity among participants; but the most effective remedy to bias is instituting the maxima strategy (Rosenthal & Rosnow, 1991).

A related problem on the boundary between agreement and performance involves false outliers. Evaluation in many tasks has a component that is legitimately subjective: the object being evaluated triggers different problems for different audiences. In any case, some degree of variability in evaluations will occur even among generally accurate reviewers. Among a small number of reviewers, there is a reasonable chance that one of the reasonable evaluations will appear (but falsely) as an outlier, which will cause the writer to ignore (but erroneously) the evaluation. Consider the case in which a writer 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 writer 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 writer may now consider the 4 point as an outlier and thus hasten to a falsely positive interpretation of the feedback. However, occurrence of false outliers is bound to decrease with an increase of reviewers (see Figure 1b).

## Performance Perspective

Advocating peer evaluation as a means to improve peer performance constitutes the mainstay of the current evaluation system movement to integrate assessment with training. Research indicates that peer evaluation (compared to expert evaluation) may equivalently or superiorly influence peer performance (Cho & Schunn, in press; Hinds, Patterson, & Pfeffer, 2001; Hughes & Large, 1993). For example, when Hinds et al. (2001) asked domain experts and novices to instruct novices (junior and senior humanities majors) on electronic-circuit activity, the novice-instructed students showed fewer errors than did the expert-instructed students. It seems that peer writers benefit from common ground or mutual knowledge (Clark & Brennan, 1991) constructed at a similar knowledge level between peer reviewers and writers (Damon & Phelps, 1989; Rogoff, 1998).

It is commonly assumed that the maxima strategy as a proxy of diversity control could augment this favorable effect. Consistently, it was found that peer writers benefit from more feedback (Cho & Schunn, in press) that includes frequent exposure to common ideas across evaluations, different perspectives across evaluations (Brinko, 1993), and more cognitive conflicts across evaluations (Cohen, 1994). In organization contexts, having learned of the positive correlation between amount of feedback and employee performance (MacDonald, Mullin, & Wilder, 2003) and of peer writers' perception of more feedback being the more helpful (Finn, 1997), many organizations have increased the number of reviewers and, hence, the volume of feedback provided to employees (Bratton & Gold, 2003).

Concerning the impact of the maxima strategy on performance, 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 reviewers tend to find more problems (Borman, 1974). Assuming that writers 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 the potentially limited number of problems to be found, a linear relationship between the number of reviewers and the number of detected problems is not expected. Thus, the number of detected problems will increase only up to a certain number of reviewers, after which diminishing returns will indicate an expiration of newfound problems. Therefore, this theory predicts a curve increasing to an asymptote (see Figure 1c).

Reinforcement theory (Annett, 1969; Deterline, 1964) focuses on the redundancy of problems detected among reviewers. By contrast to the detection theory, this theory emphasizes that writers improve performance by focusing efforts only on problems recurrently mentioned by reviewers, thus in part by filtering out incorrect or inappropriate idiosyncratic feedback and in part by shifting attention to the most problematic feedback (Annett, 1969; Anderson, 1982). Therefore, it is expected that the number of problems found in multiple evaluations grows as a function of the number of reviewers and that at a certain point, the rate of change in the number of recurrent problems could be diminishing or asymptotic (see Figure 1d).

Threshold theory (e.g. Bernardin & Beatty, 1984) assumes that more reviewers impose more difficult evaluations for writers to satisfy. Thus, writers may improve performance to the degree that they satisfy concerns set by all reviewers. In this sense, the maxima strategy plays a role of validity. For example, tasks examined by a greater number of reviewers are considered of higher quality than those examined by a lesser number of reviewers. Yet it is simply harder to satisfy all of the reviewers, and there might be a limit to the number of reviewers that could be successfully satisfied. Hence, it is predicted that the maxima strategy will show a decreasing function of performance across reviewers (see Figure 1e).

Finally, cognitive overload theory (Sweller, 1998) suggests that performance follows an inverted-U shape as a function of the number of reviewers (see Figure 1f). According to the theory, writers should process given evaluations in working memory. But because working memory is limited in capacity and duration (Baddley, 2002), only successfully processed evaluation feedback will be retained in writers' long-term memory, which is unlimited

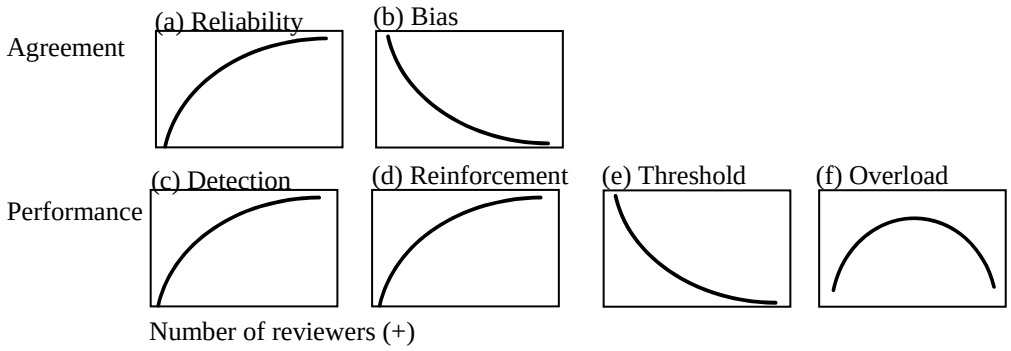


Figure 1. Theoretical predictions

in capacity and very organized using schemata. Faced with excessive information, novices or those who did not yet develop schemata will experience controlled processing of information in the working memory (Schneider & Schiffrin, 1977; Schiffrin & Schneider, 1977). Thus when working memory load is kept minimal, optimal learning and performance may occur because unused working memory capacity can support the learning process in long-term memory. 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, writers' performance will improve only when assessed by a limited number of reviewers that once exceeded, risks cognitive overload and impairs learning and performance.

In sum, there are inconsistent predictions about how the number of reviewers will influence writer performance. The goal of this study is to determine the optimal number of peer reviewers to address reliability and performance concerns in RPE systems. As shown in Figure 1, reliability theories support the maxima strategy. Thus, it is expected that more reviewers improve reliability and outlier concerns. 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. Therefore, this study examines the impact of the number of peer reviewers on agreement and performance concerns to find an optimal number of peer reviewers in RPE systems. 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 core is very common in training situations.

## Method

**Participants**. Participants included 248 undergraduate students from three cognitive psychology courses for non-majors at the University of Pittsburgh. They participated in RPE activities for their course credits. Each participant played the dual role of reviewer and writer. As an writer, each participant wrote a document, received evaluation feedback, and revised the document based on the feedback. As a reviewer, each participant evaluated six documents in both their first and final versions. A total of 2,490 evaluations by 248 students on 496 documents were analyzed. Because we controlled only the number of writers per reviewer but the number of reviewers per writer was randomly assigned, participants received evaluations from different numbers of peer reviewers. 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. Reviewer-writer pairings were assigned randomly and blindly. All participants used the SWoRD system, a web-based reciprocal peer evaluation system (Cho & Schunn, in press), described in the Interface and Procedure section.

**Document task**. The task assigned to participants was to write a document within a content area of cognitive psychology. Participants received various writing topics, which they tailored to their content areas. Average document size was 5.89 pages (SD = 1.6), 18.5 paragraphs (SD = 12.3), and 1,446.8 words (SD = 406.9). No differences were found across content areas.

The participants evaluated each draft along three dimensions: *flow*, *logic*, and *insight*. Consequently, writing quality was defined as the average of the three dimensions. 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. The *logic* dimension examined the extent to which a document was structurally coherent in terms of the organization of facts and arguments. The *insight* dimension accounted for the extent to which a document contributed new knowledge to the content area. Reviewers assessed each document and both qualitatively and quantitatively in each of the three dimensions. They generated both written comments and numeric ratings on a seven-point scale from *disastrous (1)* to *excellent (7)*.

**Interface and Procedure**. In the evaluation process, all participants used the SWoRD system (Cho & Schunn, in press). Writers electronically submitted their documents to SWoRD, which distributed the documents to randomly selected sets of peer reviewers. Reviewers reviewed the documents, generated written comments and numeric ratings, and submitted the results to the system. Having

received the results from the system, writers revised their documents accordingly, submitted revisions, and in turn *back-evaluated* the reviewers' 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: reviewers reviewing the revisions and submitting another set of comments and ratings, writers receiving and back-evaluating the second round of feedback. All proceedings were conducted anonymously.

## Results

Reliability

In order to determine the optimal number of reviewers in terms of reliability, a mean reliability among individual reviewers was first computed, which registered as .26, p. < .05. Using the Spearman-Brown formula, effective reliabilities as a function of the number of reviewers were estimated as shown in Figure 2, supporting the prediction.

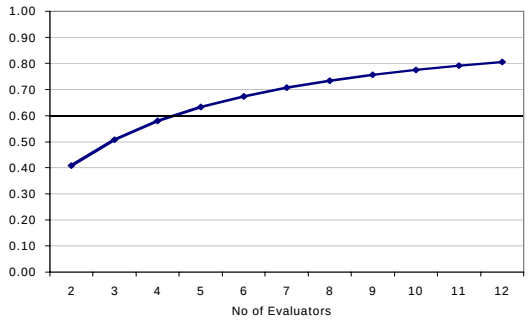


Figure 2. Effective reliabilities

The rate of true outlier reviewers, defined as outlier found in the maxima data case (n=8 reviewers), is relatively low (M = .25, SD = .50) or approximately 3% of reviewers. Although it is statistically challenging to detect among smaller numbers of evaluations, the frequency of true outliers (t) will be a linear positive function of the number of reviewers (n), $t = .3n$ . Yet the number of false outliers is expected to decrease with the number of reviewers.

We conducted the Grubbs test (Grubbs & Beck, 1972) on the data to see how many outliers were found as a function of reviewers (see Figure 3). Because the number of detected outliers is smaller except for n=7 and 8, we can assume the majority of theses cases are false outliers. The Grubbs test calculates the standardized difference between each evaluation and the mean of all evaluations. Thus, it is computed based on the distance between each rating and all ratings divided by their standard deviation. Figure 3 shows the average number of outliers according to each number of reviewers. An estimated exponent function computed was $O = 1.97e^{-.39n}$ where $O$ is the number of outliers and n is the number of reviewers. The function shows an excellent fit, $R^2$ = .90. The differentiation of the function proves a decreasing asymptotic function, $\frac{\Delta O}{\Delta n} = -.77e^{-.39n}$. As expected, Figure 3 shows that the occurrences of outliers decrease as the number of reviewers increases.

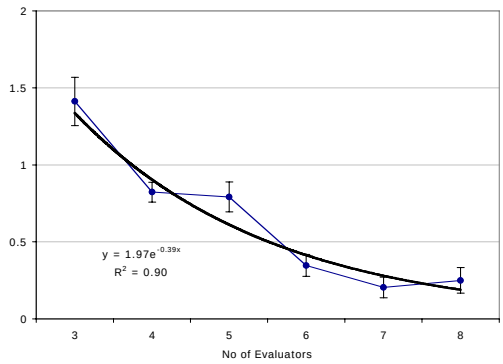


Figure 3. Outliers

Results indicate that peer evaluations locate within a certain range of deviations regardless of numbers of reviewers. In other words, outliers among a small number of evaluations would not register as such in a larger context. For example, with three reviewers, the differences among evaluations would be highly contrasted; while with eight reviewers the differences among reviewers would blur. To examine this interpretation, a one-way ANOVA with numbers of reviewers as a between-subject variable was performed on standard deviations as a dependent variable: 3 reviewers (M = 2.24, SD = 1.00), 4 reviewers (M=2.65, SD = 1.17), 5 reviewers (M = 2.67, SD = 1.25), 6 reviewers (M = 2.42, SD = 1.13), 7 reviewers (M = 2.55, SD = .88), and 8 reviewers (M = 2.59, SD = .96). No significant difference was found on the size of standard deviations, $F$ (5, 295) = .84, $p$ = .52. Therefore, this result supports the hypothesis in that perceived outliers in a smaller number of peer reviewers may not be perceived as outliers in larger numbers of peer reviewers because they tend to appear within a certain range of legal variations.

***Performance***

Concerning the performance predictions, task quality improvement as a function of the number of reviewers was computed. The task quality improvement was defined as the difference between the quality of first documents and that of final documents. Figure 4 shows that the performance improvement is an inverted U-shape. As the number of reviewers increases, performance likewise increases, peaks around six reviewers, and then decreases. Because the trend appears consistent with the cognitive overload theory, a polynomial of 2nd degree in the number of reviewers, *n*, was estimated as follows:

$$I = -.16n^2 + 1.84n - 3.43 \text{ ------ (1)}$$

where $I$ is the performance improvement and $n$ is the number of reviewers. Thus, there is either a negative or an absence of impact on performance when the amount of feedback is too little or too much. The estimated polynomial

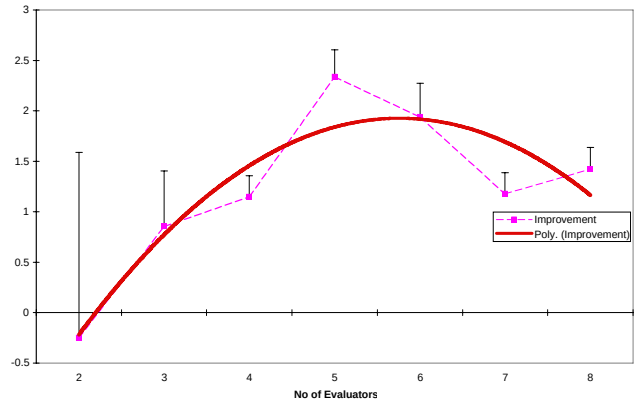


Figure 4. Performance improvement

function shows a good fit, $R^2$ = .83. To find the maximum improvement point, (1) was differentiated to calculate the rate of changes:

$$\frac{\Delta I}{\Delta x} = -.32n + 1.84 \text{ ------ (2)}$$

Thus the rate of performance improvement increases until the number of reviewers is 5.8 reviewers (1.84/.32). With more than 5.8 reviewers, the rate of performance improvement decreases. In addition, the function explains that with 11 or more reviewers, evalautees' performance would suffer rather than improve. Therefore, the function supports the cognitive overload hypothesis.

However, the performance decrement after 5.8 reviewers in the inverted-U shaped performance could be the result of writers receiving less feedback even with more reviewers (e.g., if reviewers compensated for being in a setting with more reviews by giving less feedback per review). Thus, unlike the assumption that larger numbers of reviewers provide more information, the actual amount of evaluation information could be low (Marwell & Oliver, 1993). To test this undersupply 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 writers rated as *helpful* (3 or higher) for revising their documents. Excluded was feedback rated *less helpful (*2 or lower). Accepted feedback (F) increases as a function of number of reviewers (*n*), $F = 1769.3n - 1072.5$, $R^2$ = .98, p < .001. Consequently, the hypothesis involving decreasing amount of feedback after 5.8 reviewers is rejected.

**Evaluation Time on Task**. The amount of evaluation time spent by each reviewer contextualized our argument. Although task evaluation time is not a focus in this paper, it should be noted that the number of reviewers is frequently based on the estimated time on task each participant needs or can invest. In our study, the reviewers were asked to report how much time they spent to do the evaluations. It was found that reviewers spent averaged 34.0 minutes (SD= 17.8) reading each document and 29.6 minutes (SD= 19.0) generating each evaluation. Therefore, evaluation time averaged about one hour per document and about 12 hours total evaluating first and final peer documents.

## Discussion

The maxima strategy as a key characteristic in RPE was considered from the reliability and also from the performance perspective. This study shows that there are trade-offs between reliability and performance when perspectives are jointly considered. As predicted, the maxima strategy works consistently with agreement theories. Thus, more reviewers improve reliability and evaluation biases. By contrast, performance improvement follows an inverted U-shaped trajectory as a function of the number of reviewers, as is consistent with the cognitive overload theory (Sweller, 1998). Unlike a common assumption that more evaluations augment a favorable peer evaluation effect, we found that too much feedback may hamper writers' performance. Consistently, Goodman and Wood (2004) found that increasing the amount of specific feedback may hurt learning.

The results of this study have an implication for evaluation management and training. The maxima strategy using multiple reviewers 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 reliability of evaluations and the performance of receivers. In addition, the advances of available information technology greatly enhance RPE efficiency by providing writers with rich feedback on their performances but also by facilitating the involvement of greater numbers of reviewers. The results of this study, however, caution about this very abuse of the maxima strategy promoted by reliability and performance theories as well as technology: too many reviewers may simply overload writers with information (Jones, Ravidi, & Rafaeli, 2004).

Of course, there remains much to be improved. First, this study did not focus on the difficulty or complexity of tasks. 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 (Kluger & DeNisi, 1996) and is recommend for further study.

In reporting empirical findings that six reviewers constitute an optimal number in a RPE system, this study contributes critical knowledge to research practice on evaluation feedback. Although many studies show that evaluation feedback improves task performance, a considerable amount of research reports that evaluation feedback does not improve performance and in fact deteriorates it (see Kluger & DeNisi, 1996). Given mixed results concerning the impact of evaluation feedback, the number of reviewers 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

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

Anderson, J .R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*, 369-406.

Baddeley, A. D. (2002). Is working memory still working? *European Psychologist, 7*, 85-97.

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

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

Brinko, K. T. (1993). The practice of giving feedback to improve teaching: What is effective? *Journal of Higher Education, 64* (5).

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

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

Borman, W. C. (1974). The rating of indivisuals in organizations: An alternative approach. *Organizational Behavior and Human Performance, 12*, 105-124.

Cho, K., & Schunn, C. D. (in press). Scaffolded writing and rewriting in the discipline. *Computers & Education*.

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*. Washington, DC: American Psychological Association.

Cohen, E. (1994). *Designing groupwork: Strategies for the heterogeneous classroom*. NY: Teachers Collage Press.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90*, 218-244.

Damon, W., & Phelps, E. (1989). Critical distinctions among three approaches to peer education. *International Journal of Educational Research, 13*, 9-20.

Deterline, W.A. (1962). An Introduction to Programmed Instruction. New York: Prentice-Hall.

Dochy, F., 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.

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

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(7-A), Jan 1997, 3119. US: Univ Microfilms International.

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

Harris, M. M. & Schubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41*, 43-62.

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.

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

Jones, Q., Ravidi, 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.

Katzenbach, J. R. & Smith, D. K. (1993). *The wisdom of teams*. Cambridge, MA: Harvard Business School Press.

Kluger & DeNisi, 1996

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.

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

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

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

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

Michaelson, L. K. & Black, R.H. (1994). *The key to harnessing the power of small groups I: Higher education building learning teams*. Growth Partners p. 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*, 324-335.

Rogoff, B. (1998). Cognition as a collaborative process. In W. Damon (Series Ed.) & D. Kuhn & R. S. Siegler (Vol. Eds.), *Handbook of child psychology: Vol. 2: Cognition, perception & language*. New York: John Wiley & Sons.

Rosenthal, R. & Rosnow, R. L. (1991). *Essentials of behavioral research, 2nd*. New York: 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.

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.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257-285

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*. CA: Sage.