Abstract Title Page
Not included in page count.

Title:
In the pen of the author or eye of the beholder? A measurement framework for understanding peer evaluation in writing assignments.

Authors and Affiliations:
Lee Branum-Martin, Georgia State University
Melissa M. Patchan, Georgia State University

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Peer learning is often used in classrooms to help and support knowledge and skill acquisition. One form of peer learning, peer assessment, involves the quantitative (i.e., peer ratings) or qualitative (i.e., peer feedback) evaluation of a learner's performance by another learner among students. While we might be concerned about the quality of the author's writing in the assignment, we might also be concerned about the quality of the evaluation given by peers. Importantly, the effectiveness of peer assessment could be diminished or obscured by how well peers can evaluate their classmates' work. In addition, instructional or intervention effects might also be affected by how well peer feedback is given and received. Each evaluation in such a design represents two sources of variability: that of the author and of the peer. A multilevel structural equation framework can clarify these sources of variability as well as other important questions. This presentation will give an example of how this modeling framework can clarify the substantive questions important to effective instruction and intervention in writing assignments, with implications for peer assessment designs more generally.

Purpose / Objective / Research Question / Focus of Study:

In peer assessment, there are three closely interrelated issues, which threaten valid inference for instructional or educational evaluation. First, there is an issue of measurement: ratings of multiple features might fail to represent a single dimension of writing quality. If multiple ratings do not correlate homogeneously, then there will be inconsistent, error-laden representation of the construct (e.g., writing quality or intervention impact). Second, all quality ratings as evaluated by peers (i.e., *peer ratings*) represent two sources of variability: the writing ability of the author as well as potential bias in the reader giving that score. Therefore, a model to separate these sources of variance can be clarifying and useful. Third, there is the potential for validation when external or additional criterion variables are included, such as quality ratings as evaluated by oneself (i.e., *self ratings*), instructor ratings, or performance tests, but only insofar as those criterion variables are *included at the appropriate level*. For example, while ratings representing the author's perception of the helpfulness of the comments provided by the critics (i.e., helpfulness ratings) should be included at the response level, self-ratings should be included at the author level, not repeatedly at the response level. The purpose of the current study is to illustrate these concepts in a single measurement modeling framework.

Setting:

This study was conducted in an Introduction to Cognitive Psychology course at a top-tier, public research university located in a Midwestern state in the United States.

Population / Participants / Subjects:

Of the 313 students enrolled in the course, 17 students opted out of participating in this research study. Therefore, only data from the 296 students were included for analysis. This sample represented students (73% female) at all undergraduate levels (1% freshman, 9%

sophomore, 71% junior, 17% senior, 2% other) and a variety of majors (59% social science, 17% natural science, 12% multiple disciplines, 3% business, 3% undeclared, 2% humanities, 2% education, 1% mathematics). Although the majority of the students were native English speakers, 7% of the students were not.

Intervention / Program / Practice:

Students were required to write a 5-7 page essay on a cognitive psychology topic of their choice. The assignment involved synthesizing relevant class materials with a recent research article. As students submitted their first draft, they also completed a self-evaluation. This evaluation involved rating the quality of their paper using a seven-point scale and providing an explanation for each rating on three dimensions: *content* (i.e., whether the paper clearly, thoroughly, and accurately discussed the class material and a recent article), *style* (i.e., whether the paper was appropriately organized with no grammatical errors), and *conformity* (i.e., whether the paper addressed all of the assignment details). After the first draft deadline, five peers' papers were randomly assigned to each student. Students had four weeks to review their peers' papers. These evaluations involved rating the quality of the peers' paper using the same seven-point scale that was used to evaluate one's own paper and providing constructive comments on the three dimensions. After the reviewing deadline, the reviews were automatically released to the students. They had one week to revise their draft using their peers' feedback. The students also rated the helpfulness of the comments they received using a seven-point scale. The final draft was graded by the instructor (results not reported here).

Research Design:

As in other peer evaluation settings, the dependent variables were not independent observations. There were 295 authors who received 1394 ratings from 296 critics. In this design, the data were nested within both authors and critics, but there were no reciprocal ratings (i.e., no student also reviewed a paper written by one of his or her critics), which is also known as a block design (Kenny, 1994).

Data Collection and Analysis:

Data collection: The peer review process was facilitated by SWoRD (Scaffolded Writing and Rewriting in the Discipline), an anonymous web-based peer-review system (Cho & Schunn, 2007). Using pseudonyms to ensure anonymity, students logged into this system to upload their papers and submit their evaluations.

Analysis: The data represent three scores of self-ratings, three scores of peer ratings, and three scores of helpfulness ratings. Models were fit in xxM (Mehta, 2013), a package for R software. Models were also fit in Mplus 7.2 (Muthén & Muthén, 2012). The current analysis treats the seven-point outcome scores as continuous data (categorical versions of the same models fit in Mplus resulted in findings that were essentially the same), with full information estimation for missing data (less than 1% were missing).

Findings / Results:

The data are both multivariate and cross-classified. If the multilevel structure was ignored, zero order correlations among the peer ratings were moderate and homogeneous (.49 to .67). Similarly, the correlations among the helpfulness ratings were moderate and homogeneous (.44 to .65), and the correlations among the self ratings were high and homogeneous (.67 to .78).

However, the above correlations ignore the cross-classified nesting structure and might be biased (upward or downward). In order to understand the correlation structure for the peer ratings and helpfulness ratings, we fit unconditional models for the peer ratings and helpfulness ratings in xxM. Because self-ratings involved only three variables at one level, and they correlated highly and homogeneously, these variables will be considered later in the full, joint model.

Figure 1 shows the correlations among the three peer ratings at the author, critic, and response (residual/error) levels. At the author level, the correlations were very high (.90, .83, .75). At the critic level the correlations were high (.71, .69, .86). The residuals were still moderately related (.58, .43, .31). The intraclass correlations ranged from 27% to 33% at the author level and 8% to 13% at the critic level, with the response level having 54% to 65% of the variance. Results were similar for the helpfulness ratings (Figure 2). Because the correlations were so high and homogeneous at the author and critic levels, a confirmatory factor model was fit for both quality measures (i.e., self ratings and peer ratings) and helpfulness ratings.

Figure 3 shows the full, joint measurement model of ratings by authors and critics, including the peer ratings (i.e., qual), the self-ratings (i.e., self), and the helpfulness ratings (i.e., help). Each of these three levels is shown in a rectangle. The circles represent the hypothesized latent factors to capture how the sets of three scores each measure a single construct. The numbers on the straight arrow paths are fully standardized factor loadings, indicating the correlation between the indicator and its factor.

At the author level, authors who scored high on one dimension tended to score high on all dimensions (loadings .83 to .98 for the Qual factor). Similarly, authors who thought the comments they received were helpful found comments for all dimensions helpful (loadings .69 to .98 on the Help factor). Lastly, authors who thought their own writing was of good quality tended to rate themselves well on all dimensions (loadings .77 to .90 on the Self factor).

At the critic level, critics who rated papers high on one dimension tended to rate them high on all dimensions (loadings .77 to .94 for the Qual factor). Likewise, critics whose comments were considered to be helpful on one dimension tended to provide helpful comments on all dimensions (.80 to .94 for the Help factor).

The numbers on the curved paths are correlations. At the author level, the relations among the three latent factors were: .41 between the self-ratings and the peer ratings and .20 between the peer ratings and the helpfulness ratings. However, the relation between the author's quality ratings and the author's ratings of helpfulness was essentially zero (r = -.03). At the critic level, the relation between the critics' quality ratings and the authors' ratings of helpfulness was also essentially zero (r = -.15). Residual correlations are shown at the response level and were low to moderate among the peer ratings (r = .32 to .58) and helpfulness ratings (r = .26 to .43), and essentially zero across the two types of ratings (r = -.04 to .09).

Conclusions:

The cross-classified results show a correlation structure dramatically different than would otherwise be observed. On average, approximately 30% of the variance in peer ratings was due

to authors and 10% due to critics. This split in variance was similar for the helpfulness ratings. The three dimensions of ratings were highly consistent with each other, both for quality and for helpfulness, as evaluated by the critic as well as the author. This measurement consistency (indexed by high loadings) suggests that writing quality is perceived as a reasonably coherent construct by critics as well as authors and reflects the author's written work as a single construct. The consistency among helpfulness ratings also suggests that critics who give good feedback generally do so in all three dimensions, and that authors also have a consistent way of seeing evaluations (variance due to authors was high). These relations suggest that both authors and critics have systematic biases to see writing and ratings in ways that are typical of themselves as well as typical of the target they see (e.g., the essay or the comments).

The relations between these constructs have implications as well. Among critics, the way a critic views a paper is generally not related to how their advice is perceived for helpfulness. Among authors, there is a tendency for authors who write well to give higher ratings of helpfulness to the critics (r = .41). There was only a weak relation between self-ratings and peer ratings (r = .20), suggesting that these students do not have strong agreement on the quality of writing. It is interesting that for both quality ratings and helpfulness ratings, approximately 30% of the variance was due to the author, suggesting that quality ratings are driven mostly by the author and that helpfulness ratings reflect the perception of the author (i.e., the quality of criticism might not be perceived in a stable or reliable way).

One limitation to the current study is that it included only three outcomes per construct. Additional outcomes could allow for differentiation among facets of writing quality or helpfulness. With only three tests per factor (zero degrees of freedom), the factors cannot be evaluated individually for fit (though the standardized loadings can be interpreted as validity coefficients). Another limitation to the current design is the set of three criterion measures from only one source: self-ratings. Additional criterion measures from achievement tests or expert/instructor ratings could further clarify questions of external validity. A third limitation, not testable in current software, is a person-level correlation across author and critic: to what extent does a good author tend to give good feedback?

In conclusion, the strong measurement properties suggest that writing quality and helpfulness are measured well as constructs by these items. Undergraduate students may have only weak agreement on what constitutes good writing. The intraclass correlations reflect reliability: differences across types of critics (i.e., self versus peer) indicate disagreement. The strength of the current framework is not only the measurement models permitted, but also the separation of sources of perception to their appropriate levels: authors versus critics.

Future research could focus on several extensions of this measurement framework. First, intervention and longitudinal effects can be examined as additional outcomes and grouping variables. Interventions could be tested as mean differences (e.g., increases in quality) or as modifications to variances and covariances. For example, an effective intervention might reduce bias by reducing variance across critics (i.e., they agree more on what constitutes quality writing). Second, measures of external validation could be added, including tests of written ability at the author level or expert reviews by trained critics. Third, peer interaction could be evaluated by allowing reciprocal ratings, creating a level for dyads. Such dyadic relations result in a version of the Social Relations Model (Kenny, 1993). A reciprocal design would allow the estimation of the association between author and critic when they have both seen each other's work. Finally, classroom differences could also be modeled as an additional level of clustering, along with measures of instructional effectiveness (e.g., observational data).

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

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

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York, NY, US: Guilford Press.

Mehta, P. D. (2013). n-level structural equation modeling. In Y. Petscher, C. Schatschneider, & D. L. Compton (Eds.), *Applied quantitative analysis in the social sciences* (pp. 329-362). New York: Routledge.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide. Seventh edition.* Los Angeles, CA: Muthén & Muthén.

Appendix B. Tables and Figures

Not included in page count.

Figure 1: Quality ratings: Unconditional multilevel cross-classified correlations and variance percentages (fully standardized results)

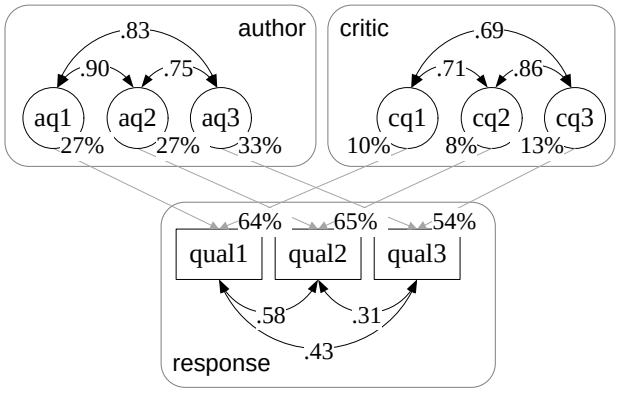

Figure 2: Helpfulness ratings: Unconditional multilevel cross-classified correlations and variance percentages (fully standardized results)

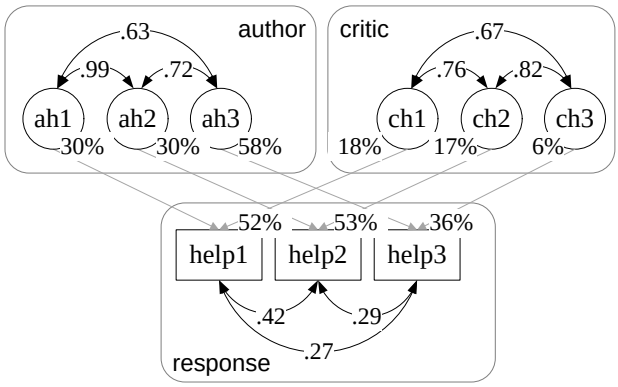

Figure 3: Full measurement model of ratings by authors and critics (fully standardized results; mean structure not shown)