World Academy of Science, Engineering and Technology
International Journal of Educational and Pedagogical Sciences
Vol:11, No:10, 2017

An Exploratory Study of Reliability of Ranking vs. Rating in Peer Assessment

Yang Song, Yifan Guo, Edward F. Gehringer

***Abstract*—**Fifty years of research has found great potential for peer assessment as a pedagogical approach. With peer assessment, not only do students receive more copious assessments; they also learn to become assessors. In recent decades, more educational peer assessments have been facilitated by online systems. Those online systems are designed differently to suit different class settings and student groups, but they basically fall into two categories: rating-based and ranking-based. The rating-based systems ask assessors to rate the artifacts one by one following some review rubrics. The ranking-based systems allow assessors to review a set of artifacts and give a rank for each of them. Though there are different systems and a large number of users of each category, there is no comprehensive comparison on which design leads to higher reliability. In this paper, we designed algorithms to evaluate assessors' reliabilities based on their rating/ranking against the global ranks of the artifacts they have reviewed. These algorithms are suitable for data from both rating-based and ranking-based peer assessment systems. The experiments were done based on more than 15,000 peer assessments from multiple peer assessment systems. We found that the assessors in ranking-based peer assessments are at least 10% more reliable than the assessors in rating-based peer assessments. Further analysis also demonstrated that the assessors in ranking-based assessments tend to assess the more differentiable artifacts correctly, but there is no such pattern for rating-based assessors.

***Keywords*—**Peer assessment, peer rating, peer ranking, reliability.

I. Introduction

TRADITIONAL classroom technologies are designed primarily to "push" information from one instructor to many students. Extensive research has shown that much more learning and skill development take place when there is bi-directional interaction in a class: Students work with both students and instructors providing feedback to the students on their artifacts [1]. Peer assessment is an instructional technology designed to support such interactions [2]. Compared to older technologies and traditional instruction, peer assessment enables more effective and timely feedback, increased engagement, and improved conceptual learning, particularly in large classes [3].

The IT revolution has enabled a host of online peer assessment systems, which automate the workflow and remind students to complete their tasks [4]-[7]. The dozens of known educational peer-assessment applications fall into two categories: systems based on rating, and systems based on ranking. Rating-based peer assessment asks assessors to rate each artifact on a Likert scale (or multiple Likert scales for several criteria). Ranking-based peer assessment, on the other hand, asks assessors to rank several artifacts against each other.

Advocates of each kind of system cite advantages (e.g., the ranking-based assessment shows better robustness across rankers [8], while rating-based peer assessment better supports detailed rubrics [9]), but no quantitative comparison has been done of the reliability of these two approaches. There are multiple reasons for this. First, carrying out this comparison requires data collection and transformation from multiple online systems. Second, most of the existing measures of reliability (e.g. ICC, correlation coefficient) are suitable only for rating-based systems, and therefore there is no metric of reliability that can be applied to both rating-based and ranking-based peer assessments.

In our previous work, we defined a Peer-Review Markup Language (PRML), which models the common entities and relations in the educational peer-assessment process [10]. Based on this PRML, we are able to collect data generated by different peer assessment systems and build a data warehouse for the data from multiple systems. This meets the first challenge.

This paper describes our efforts on dealing with the second challenge—to design metrics comparing peer-assessment reliability from data generated by both rating and ranking-based systems. The rest of this paper is organized as follows. Section II presents detailed information about different peer assessment settings from two representative systems. Section III describes the design of our reliability algorithms. Section IV presents our experiment results and our discussions. Section V considers future work.

II. Different Peer Assessment Settings

Peer ranking and peer rating have been long recognized as two main approaches for peer assessment [11]. However, sometimes researchers may confuse peer ranking with peer rating [12]. Though arguably they can be done at the same time, they are built into online educational peer assessment systems with clearly different purposes and emphases.

A. Rating-Based Peer Assessment

Rating-based peer assessment systems usually ask assessors (usually student peers, sometimes teaching staff as well [13], [14]) to rate artifacts. The assessors usually review one artifact at a time, therefore detailed review rubrics can easily be applied to peer rating because the reviewed artifact holds the assessor's attention well. In practice, teaching staff may design detailed

Yang Song is with the Department of Computer Science, North Carolina State University, Raleigh, NC 27613 USA (e-mail: ysong8@ncsu.edu).
Yifan Guo is with the Department of Computer Science, North Carolina State University, Raleigh, NC 27613 USA (e-mail: yguo14@ncsu.edu).
Edward F. Gehringer is with the Department of Computer Science, North Carolina State University, Raleigh, NC 27613 USA (e-mail: efg@ncsu.edu).

rubric questions to guide assessors in reviewing different aspects of the artifact, and giving either numerical or textual feedback (or both). The numerical feedback, aggregated together, can be considered as a total score that the assessor assigns to the artifact.

Fig. 1 is a screenshot from the Expertiza online peer assessment system [5]. Each bullet point is one rubric question. Based on each question, assessors are required to give both numerical feedback (on a 5-point Likert scale) and textual feedback. The authors, then, can learn on which aspect they did well or not so well. If the review rubrics are designed properly, longer and more helpful feedback can also be triggered [14].

B. Ranking-Based Peer Assessment

The designers of ranking-based peer assessment systems usually argue that common understanding of the rating standards is hard to reach [8], e.g., on a 5-point Likert scale, how does a "3" differ from a "4"? To reach such a common understanding, a calibration or training phase should be applied [15], [16], which makes the assignment design more complicated.

To avoid the hassle of crafting detailed rubrics and at the same time, improve the reliability of peer assessment, peer ranking is utilized in some systems. The assessors are usually asked to review a fixed number of artifacts. Instead of giving numerical scores to each of the artifacts, assessors need to rank them in order, from strong to weak. Researchers argue that ranking is a more reliable approach to peer assessment because the quantitative feedback (ranking in this case) is given by the comparisons between one artifact and others.

Fig. 2 is part of the user interface from Mobius SLIP [6] which is a ranking-based peer assessment system. Assessors, in this case, are assigned to review four artifacts from their peers and rank them together with their own artifacts. Each assessor can use the scroll bars to rank those five artifacts. Textual feedback is also supported by Mobius SLIP, but it can only be holistic feedback instead of feedback facilitated by detailed rubrics.

III. Tuple Based Reliability Algorithm

A. Global Rank

After enough peer assessments have been done, it is possible to approximate a global rank for an assignment. This can be done with data generated by either rating-based or ranking-based peer-assessment systems. Some systems may display the global rank on leaderboards [17] or simply point students to strong work [8]. The basic assumption of this research is that there is a "ground truth" global ranking of all the artifacts submitted for a specific assignment.

To give a precise definition of the algorithm which generates the global rank for an assignment, let a be an artifact; A be the set of all the artifacts; r be a reviewer; R be the set of all the reviewers; g_a^r be the quantitative grade that r assigned to a; R_r be the set of artifacts reviewed by r; A_a be the set of reviewers who have reviewed a; G_a be the aggregated grade for artifact a based on existing peer assessments; $GlobalRank$ be the global ranking for all the artifacts A in an assignment; $GlobalRank_a$ be the global ranking for artifact a.

G_a can be defined as the average peer-assessment score/rank for artifact a:

$$G_a = \sum_{r \in A_a} g_a^r \Big/ |A_a| \tag{1}$$

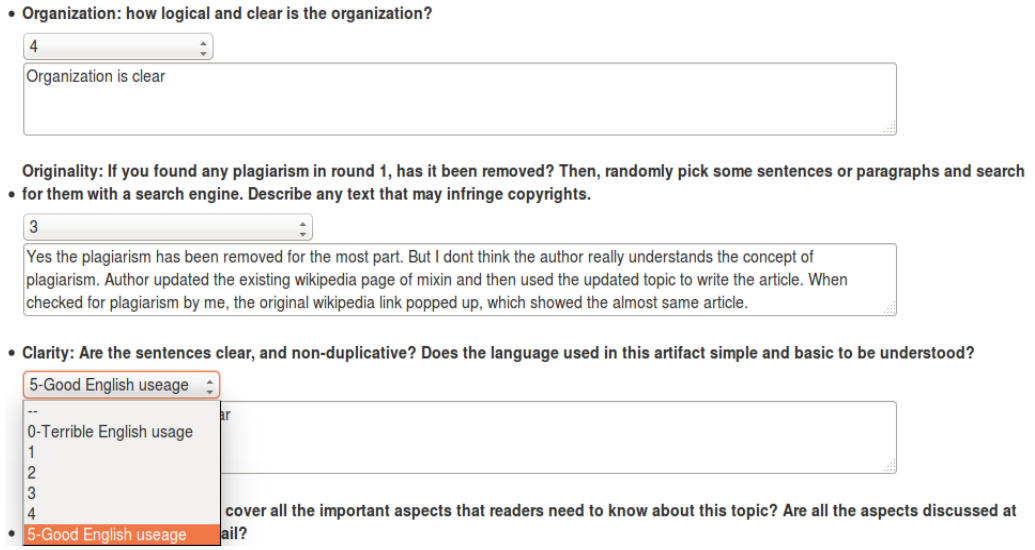

Fig. 1 Screenshot from a rating-based peer assessment system (Expertiza)

Fig. 2 Screenshot from a ranking-based peer assessment system (Mobius SLIP)

In the peer-ranking scenario, G_a is the average rank that artifact a has received; in the peer-rating scenario, G_a is the average peer-review score that artifact a has received.

Based on G_a for all $a \in A$, $GlobalRank_a$ can be defined as:

$$GlobalRank_a = \sum_{a' \in A, a' \neq a} \left| G_{a'} > G_a \right| + 1 \quad (2)$$

Pulsing one is to make the rank start from 1 instead of 0. $GlobalRank$ can be considered as a sequence of artifacts in the order of $GlobalRank_a$.

Please note that 1) it is possible that multiple artifacts have the same rank and 2) it does not matter that the $GlobalRank$ is from best to worst or vice versa as long as the $GlobalRank$ of the artifacts is maintained.

B. Tuple Based Reputation

In peer assessment, the assessments done by an assessor r who has reviewed more than one artifact can be considered to define a $LocalRank_r$. For example, reviewer r has reviewed artifact a_1, a_2, a_3 and a_4, and r's local rank is $\langle a_1, a_2, a_3, a_4 \rangle$ in the same numerical order as $GlobalRank$, for example, from high to low. The $LocalRank_r$ can be further broken into six (there are four assessed artifacts, so $C(4,2)$ possible combinations) 2-tuples. For each 2-tuple, we can define the weight of assessing the two artifacts as $Weight(< a_x, a_y >)$.

Depending on whether an assessor's assessment for a 2-tuple agrees with the $GlobalRank$, we can define the reliability achieved by this assessing this tuple:

$$Achieved(< a_x, a_y >) = \begin{cases} 0, < a_x, a_y > \text{ disagree with } GlobalRank \\ Weight(< a_x, a_y >), otherwise \end{cases} \quad (3)$$

For each assessor r, the reliability can be defined as:

$$\text{Reliability}_r = \frac{\sum_{< a_x, a_y > \in LocalRank_r} Achieved(< a_x, a_y >)}{\sum_{< a_x, a_y > \in LocalRank_r} Weight(< a_x, a_y >)} \quad (4)$$

The range of reliability is $[0,1]$.

If we assume that all the 2-tuples have the same weight (to make it simple, we assume a weight of 1), the reliability becomes the percentage of the 2-tuples where the assessor agrees with the $GlobalRank$:

$$Weight(< a_x, a_y >) = 1 \quad (5)$$

We call this algorithm Tuple-Based Reliability-Unified (TBR-U) in the rest of this paper.

If we emphasize that mistakenly rating/ranking an obviously stronger artifact lower than a much worse one should be punished further, we can modify (5) as below:

$$Weight(< a_x, a_y >) = \left| GlobalRank_{a_x} - GlobalRank_{a_y} \right|^2 \quad (6)$$

In this case, for a tuple $< a_x, a_y >$ which does not match the $GlobalRank$, the greater difference on $GlobalRank_{a_x}$ and $GlobalRank_{a_y}$, the lower reliability the assessor will get. We call this algorithm Tuple-Based Reliability - Differentiated (TBR-D) in the rest of this paper.

The difference between TBR-U and TBR-D algorithm is that TBR-D uses higher weights for tuple $< a_x, a_y >$ if the difference of $GlobalRank_{a_x}$ and $GlobalRank_{a_y}$ is larger. For example, consider an assignment with 20 artifacts where a_1 is the strongest and a_{20} is the weakest. If an assessor mistakenly assesses the tuple $< a_1, a_{20} >$, the decrease in the reliability should be higher than if (s)he mistakenly assesses the tuple $< a_{10}, a_{11} >$.

IV. Experiments

A. Dataset

The experiment was done on data from the PeerLogic data warehouse [18]. The data is taken from multiple educational peer assessment systems and transformed into the same schema. There are 15,498 assessments (assessments give either a rank or *comprehensive* rating, comprising responses to all criteria in a single rubric) from 466 assignments used in this experiment. Table I provides more details of our dataset.

TABLE I
Details about the Dataset Used in the Experiment

	Rating-based	Ranking-based
Num. of assignments	260	206
Num. of participants	3323	2163
Num. of assessments	6626	8870
Avg. participants per assignment	31.4	43.6

Due to different class settings, the courses which used rating-based peer assessment required students to do fewer assessments (two on average) than the courses which used ranking-based assessment (four on average).

B. Experiment Results and Discussion

We calculated the individual reliabilities for all the assessors based on TBR-U and TBR-D algorithms. The distributions of students' reliabilities are plotted in Figs. 3 and 4.

The bars on the right side of Figs. 3 and 4 show that more student assessors have reliabilities between 80% to 100% than have lower reliabilities. Comparing the reliability from rating-based assessors, we found that the average reliability from TBR-D (0.690) is almost the same to the average reliability from TBR-U (0.692). For the reliability from ranking-based assessors, we found that the average reliability from TBR-D (0.890) is higher than the average reliability from TBR-U (0.821). Table II gives more details about the reliability of those

two algorithms.

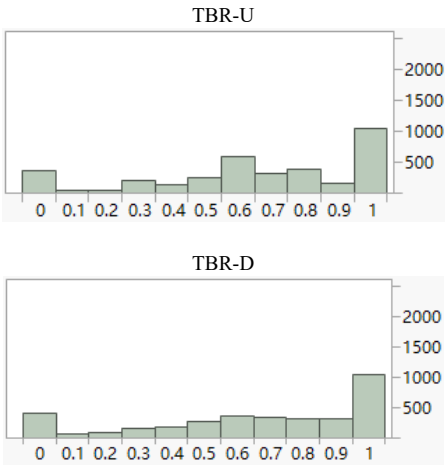

Fig. 3 Distribution of assessors' reliability in rating-based assignments

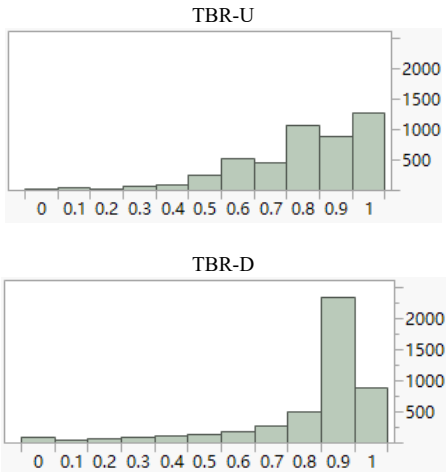

Fig. 4 Distribution of assessors' reliability in ranking-based assignments

TABLE II
STATISTICS OF ASSESSOR RELIABILITIES

	Rating-based	Ranking-based
Avg. TBR-U	0.692	0.821
Std. TBR-U	0.314	0.169
Avg. TBR-D	0.690	0.891
Std. TBR-D	0.334	0.190

The average reliabilities from rating-based assessors and ranking-based assessors show that the assessors tend to be more reliable in ranking-based assessment, no matter which algorithm is used. On average, the ranking-based assessors can rank 12.9% more of the artifacts correctly (or at least agree with the majority of other assessors) compared with the rating-based assessors. This difference can be largely explained by the fact that ranking-based assessors need to compare different artifacts before they give the rank, and therefore the ranks are likely to be more reliable. The rating-based assessors, on the other hand, review only one artifact at one time, and therefore, there is a higher chance that they may rate a weaker artifact higher than a stronger one.

The TBR-D algorithm punishes the assessors more if they mistakenly rate artifacts which are clearly differentiable. In other words, if an assessor rates the first best artifact lower than the worst artifact in an assignment (which is more unlikely to happen), the reliability will drop more in TBR-D than in TBR-U. Therefore, we expect the average reliability from TBR-D to be higher than the average reliability from TBR-U. We found this to be true among ranking-based assessors: the average reliability increases almost 7% from TBR-U to TBR-D. This means that though it is harder for assessors to rank two artifacts of roughly the same quality, assessors tend to rank more differentiable artifacts correctly. However, the rating-based assessors do not show this pattern: the average reliability from rating-based assessors barely changes from TBR-U to TBR-D. This means that the likelihood for rating-based assessors to make mistakes (generating 2- tuples which do not agree with the global rank) is almost the same regardless of the global ranks of two random artifacts. We believe this is mainly caused by the fact that the assessors do not re-visit the artifacts they have reviewed frequently, especially before they assess new artifacts.

V. CONCLUSION AND FUTURE WORK

We have introduced tuple-based reliability for measuring assessors' reliability in educational peer assessment. This concept can measure reliabilities for both rating-based and ranking-based assessments, and therefore, can be used to make a comprehensive comparison on assessors' reliabilities for both peer-assessment settings. Our experiments on the Peerlogic data warehouse [10] show that the reliabilities achieved by ranking-based assessors are on average higher than reliabilities of rating-based assessors'. This finding corroborates Shah *et al.* [19] and suggests that if reliability is the higher priority for the teaching staff, ranking-based peer assessment system may be a better choice. However, rating also has its advantages over ranking. It is easier to use a detailed review rubric in a rating system, because a reviewer can rate an artifact on a set of criteria without flipping back and forth among the artifacts for each criterion on which the artifacts are ranked. In a system with detailed rubrics, authors can receive more formative feedback, which helps them to improve their artifacts [14].

There is also more space for rating-based peer assessment system designers to improve their tools, and facilitating assessors to re-visit their previous assessments should be one of them. The visualization in Fig. 2 is from a ranking-based system; however, the designers can also build this kind of visualization into rating-based systems to remind assessors of the artifacts they have already assessed and the scores they assigned to them. We believe features like this can improve the reliability of rating.

In this work, the average peer-review scores/ranks for each artifact were used to generate the global rank. There are some approaches which generate more credible aggregated scores, e.g. with reputation algorithms [20], [21]. We did not use this approach because no reputation algorithm can be applied to the data generated by both ranking-based and rating-based systems

yet. Such a reputation algorithm is still a missing piece for carrying out more comprehensive comparisons between rating-based and ranking-based assessment approaches.

ACKNOWLEDGMENT

This research is part of the PeerLogic project, which is funded by the National Science Foundation under grants 1432347.

REFERENCES

[1] S. M. Brookhart, The Art and Science of Classroom Assessment. The Missing Part of Pedagogy. ASHE-ERIC Higher Education Report, Volume 27, Number 1. ERIC Clearinghouse on Higher Education, One Dupont Circle, Suite 630, Washington, DC 20036-1183 ($24)., 1999.

[2] K. Topping, "Peer Assessment Between Students in Colleges and Universities," Rev. Educ. Res., vol. 68, no. 3, pp. 249–276, Sep. 1998.

[3] F. Dochy, M. Segers, and D. Sluijsmans, "The use of self-, peer and co-assessment in higher education: A review," Stud. High. Educ., vol. 24, no. 3, pp. 331–350, Jan. 1999.

[4] D. Babik, E. F. Gehringer, J. Kidd, P. Ferry, and T. David, "Probing the Landscape: A Systematic Meta-review of Online Peer Assessment Systems in Education," in CSPRED 2016: Workshop on Computer-Supported Peer Review in Education, 9th International Conference on Educational Data Mining (EDM 2016), Raleigh, N.C, 2016.

[5] E. Gehringer, "Expertiza: information management for collaborative learning," Monit. Assess. Online Collab. Environ. Emergent Comput. Technol. E-Learn. Support, pp. 143–159, 2009.

[6] "Mobius SLIP: UNCG develops a new online learning tool," Research & Economic Development, 06-Nov-2013. (Online). Available: http://research.uncg.edu/spotlight/mobius-slip-uncg-develops-a-new-online-learning-tool/. (Accessed: 08-Jul-2016).

[7] L. De Alfaro and M. Shavlovsky, "CrowdGrader: A Tool for Crowdsourcing the Evaluation of Homework Assignments," in Proceedings of the 45th ACM Technical Symposium on Computer Science Education, New York, NY, USA, 2014, pp. 415–420.

[8] D. Tinapple, L. Olson, and J. Sadauskas, "CritViz: Web-based software supporting peer critique in large creative classrooms," Bull. Tech. Comm. Learn. Technol., vol. 15, no. 1, 2013.

[9] Y. Song, Z. Hu, and E. F. Gehringer, "Closing the Circle: Use of Students' Responses for Peer-Assessment Rubric Improvement," in Advances in Web-Based Learning -- ICWL 2015, F. W. B. Li, R. Klamma, M. Laanpere, J. Zhang, B. F. Manjón, and R. W. H. Lau, Eds. Springer International Publishing, 2015, pp. 27–36.

[10] Y. Song, F. Pramudianto, and E. F. Gehringer, "A markup language for building a data warehouse for educational peer-assessment research," in 2016 IEEE Frontiers in Education Conference (FIE), 2016, pp. 1–5.

[11] J. S. Kane and E. E. Lawler, "Methods of peer assessment," Psychol. Bull., vol. 85, no. 3, pp. 555–586, 1978.

[12] M. van Zundert, D. Sluijsmans, and J. van Merriënboer, "Effective peer assessment processes: Research findings and future directions," Learn. Instr., vol. 20, no. 4, pp. 270–279, Aug. 2010.

[13] J. Hamer, K. T. K. Ma, and H. H. F. Kwong, "A Method of Automatic Grade Calibration in Peer Assessment," in Proceedings of the 7th Australasian Conference on Computing Education - Volume 42, Darlinghurst, Australia, Australia, 2005, pp. 67–72.

[14] Y. Song, Z. Hu, Y. Guo, and E. F. Gehringer, "An experiment with separate formative and summative rubrics in educational peer assessment," in 2016 IEEE Frontiers in Education Conference (FIE), 2016, pp. 1–7.

[15] Z. Hu, Y. Song, and E. Gehringer, "The Role of Initial Input in Reputation Systems to Generate Accurate Aggregated Grades from Peer Assessment - Semantic Scholar," in CSPRED 2016: Workshop on Computer-Supported Peer Review in Education, 9th International Conference on Educational Data Mining (EDM 2016), 2016.

[16] Y. Song, E. F. Gehringer, J. Morris, J. Kid, and S. Ringleb, "Toward Better Training in Peer Assessment: Does Calibration Help?," in Computer-Supported Peer Review in Education (CSPRED-2016), 2016.

[17] P. Denny, A. Luxton-Reilly, and J. Hamer, "The PeerWise System of Student Contributed Assessment Questions," in Proceedings of the Tenth Conference on Australasian Computing Education - Volume 78, Darlinghurst, Australia, Australia, 2008, pp. 69–74.

[18] F. Pramudianto et al., "Peer Review Data Warehouse: Insights From Different Systems," in CSPRED 2016: Workshop on Computer-Supported Peer Review in Education, 9th International Conference on Educational Data Mining (EDM 2016), 2016.

[19] Shah, N. B., Bradley, J. K., Parekh, A., Wainwright, M., & Ramchandran, K. (2013, December). A case for ordinal peer-evaluation in MOOCs. In NIPS Workshop on Data Driven Education.

[20] Y. Song, Z. Hu, and G. Gehringer, "Who Took Peer Review Seriously: Another Perspective on Student-Generated Quizzes," in CSPRED 2016: Workshop on Computer-Supported Peer Review in Education, 9th International Conference on Educational Data Mining (EDM 2016), Raleigh, N.C, 2016.

[21] Y. Song, Z. Hu, and E. F. Gehringer, "Pluggable reputation systems for peer review: A web-service approach," in IEEE Frontiers in Education Conference (FIE), 2015. 32614 2015, 2015, pp. 1–5.