Predictive Effects of Online Peer Feedback Types on Performance Quality

Fu-Yun Yu1* and Chun-Ping Wu2

1Institute of Education, National Cheng Kung University, No. 1, University Road, Tainan City 701, Taiwan
2Department of Educational Technology, TamKang University, No.151, Yingzhuan Rd., Danshui Dist., New Taipei City 251, Taiwan

*Corresponding author

(Submitted September 01, 2011; Revised November 08, 2011; Accepted December 06, 2011)

ABSTRACT

This study examined the individual and combined predictive effects of two types of feedback (i.e., quantitative ratings and descriptive comments) in online peer-assessment learning systems on the quality of produced work. A total of 233 students participated in the study for six weeks. An online learning system that allows students to contribute to and benefit from the process of question-generation and peer-assessment was adopted. The regression results indicated that quantitative ratings and descriptive comments significantly predicted the quality of produced work (i.e., question-generation performance) both individually and collectively, and descriptive feedback explained more variance in quality of produced work than did quantitative ratings. The empirical significance of this study and suggestions for online learning system development, instructional implementation and future studies are discussed.

Keywords

Online learning system, Peer-assessment, Performance quality, Types of feedback

Introduction

Several theories and empirical studies have suggested that peer-assessment strategies can enhance the quality of designed products (Olson, 1990; Topping, 1998; Yu, 2011; Yu & Wu, 2011). Existing studies on peer-assessment have mostly focused on investigating: (1) learners’ perceptions of gains from feedback given by peers and attitudes towards peer-assessment (e.g., Venables & Summit, 2003; Wen & Tsai, 2006; Wen, Tsai, & Chang, 2006; Yu & Wu, 2011), (2) the validity and reliability of peer-assessment (e.g., Cho & MacArthur, 2010; Falchikov & Goldfinch, 2000; Hughes & Large, 1993; Smith, 1990; van Zundert, Sluijsmans & van Merriënboer, 2010) and (3) effects of different methods and timings of peer-assessment on learning outcomes (e.g., Gielen, Peeters, Dochy, Onghena, & Struyven, 2010; Strijbos, Narciss, & Dünnebier, 2010; van Steendam, Rijlaarsdam, Sercu, & van den Bergh, 2010; van den Berg, Admiraal, & Pilot, 2006).

In light of many of the distinct features and capabilities of networked technologies (such as high processing speeds, large amounts of storage space, learner control, multimedia, simultaneity, instantaneity, space-, time- and device-independence, flexibility, interactivity, among others), numerous online peer-assessment systems have been developed and have demonstrated their efficacy over the past decade. While great strides have been made in this area, a few research gaps still exist. Among these, effects of different types of feedback on produced work still require more investigation. Considering that feedback can be expressed in both quantitative and descriptive terms, its individual and combined predictive effects are examined in this study. Three research hypotheses are proposed:

1. The averaged quantitative ratings received online from assessors will significantly predict the quality of produced work.
2. The quality of the descriptive comments received online from assessors will significantly predict the quality of produced work.
3. The combined averaged quantitative ratings and the quality of the descriptive comments received online from assessors for the composed questions will significantly predict the quality of produced work.

By assessing the validity of the above questions, we hope to illuminate ways in which peer-assessment activities and environments can be better designed to improve the quality of examined work.
Literature review

Conceptualization of peer-assessment

Peer-assessment can be viewed as a highly effective collaborative learning activity in which learners produce lesson-related artifacts, critically evaluate artifacts produced by their peers and provide them with feedback, and receive feedback from their peers regarding their own produced artifacts (Ammer, 1998; Falchikov & Goldfinch, 2000; Fallows & Chandramohan; 2001; Topping, 1998; van Gennip, Segers, & Tillema, 2010). Through mutual support and assistance, peers with similar learning statuses assist each other to achieve learning goals and to extend both their knowledge bases and skill levels (Topping & Ehly, 2001). Such collaborative relationships not only help to assess the quality of submissions or student performance according to a set of criteria, but also identify areas that may be improved upon and offer suggestions for ways in which the work or performance may be refined (Boud, Cohen, & Sampson, 1999; Topping, 1998; Topping, 2010). Considering that peers tend to be within or near each other’s zone of proximal developments, peers’ comments may be more easily understood by learners than instructors’ (Ammer, 1998; Fallows & Chandramohan, 2001).

Several cognitive processes are mobilized when students are engaged in peer-assessment activities. Assessing the relative quality and merits of the examined work encourages students to engage in critical thinking. In addition, both social and argumentation skills as well as substantial knowledge in the applied area are required to enable comments to be accepted by peers. Also, when observing peers’ work, students are likely to be alerted to problems that may exist in their own work and be prompted to make necessary modifications. On the other hand, when students receive feedback from assessors, the comments provided may cause cognitive conflict and direct students to deal with their existing cognitive defects. Knowledge structuring and re-structuring are cultivated through various cognitive and discursive processes (such as deeper elaboration of materials, self-reflection, comparison, clarification, adjustment, and so on). All of the aforementioned processes, based on cognitive conflict theory, social constructivism, social learning theory and metacognition, should promote the development of critical thinking, knowledge integration and cognitive and metacognitive abilities (Falchikov & Goldfinch, 2000; Topping, 1998; van Gennip et al., 2010; Yu, 2011; Yu & Wu, 2011).

Current research state of peer-assessment

Empirical evidence spanning more than two decades has generally demonstrated peer-assessment’s benefits to better learning and increased cognitive abilities, such as critical thinking, metacognitive strategy, cognitive re-structuring as well as improved comprehension of learned material (Brindley & Scoffield 1998; Falchikov & Goldfinch, 2000; Gatfield, 1999; Hanrahan & Isaacs, 2001; Macpherson, 1999; Purchase, 2000; Topping, 1998; Tsai, Lin & Yuan, 2002; van Gennip et al., 2010). In addition to cognitive gains, peer-assessment creates affective benefits for students. The fact that someone other than teachers will view their work tends to induce students to take charge of and invest more in preparation of their initial work and the subsequent revision of the work. These behaviors are, in essence, a manifestation of increased motivation and sense of responsibility toward learning (Humphrey, Greenan, & McIlveen, 1997).

Recently, to help alleviate the time-consuming, tedious and effort-demanding aspects of using a paper-and-pencil format in a large class setting (such as, collection, preparation and assignment of student work, compilation and calculation of feedback provided by peers and the returning of feedback to individual students, and so on), numerous online peer-assessment systems have been developed to manage peer-assessment (for instance, CAP, NetPeas, Vee heuristic, Web-SPA and SWoRD) (Cho & Schunn, 2007; Sung, Chang, Chiou, & Hou, 2005; Tsai et al., 2001). By using computer and network technologies, such developed systems have demonstrated efficacy for carrying out associated tasks efficiently and effectively (Cho & Schunn, 2007; Davies, 2000; Lu & Bol, 2007; Sung et al., 2005; Trautmann, 2009; Tsai et al., 2001).

While great strides have been made in this area and researchers commonly agree upon the importance of clear and pre-specified criteria for objective peer-assessment (Crane & Winterbottom, 2008; Gielen et al., 2010; Ormond, Merry, & Reiling, 1996), some research gaps still exist. Recent research has focused on the effects of different types of feedback (e.g., concise feedback versus elaborated feedback, justified feedback versus non-justified feedback).
In light of the fact that feedback can be expressed in both quantitative and descriptive forms, “if and how these different types of feedback provided lead to the improvement of the quality of submitted work” is a pertinent issue that should be pursued. Since some online peer-assessment systems only require assessors to rate the quality of the examined work along a set of rating scales, while others enable assessors to include elaborate feedback for the author’s consideration, their individual and combined predictive effects on the quality of the produced work will help to establish guidelines for better peer-assessment system design and implementation.

Method

Participants and implementation procedures

Two hundred and thirty-three 5th graders from eight classes participated in the study for six consecutive weeks. In view of the beneficial effect of student question-generation on learning (Brown & Walter, 2005; Rosenshine, Meister, & Chapman, 1996; Yu & Liu, 2005; Yu & Liu, 2009; Yu et al., 2005), it was chosen as the learning activity and thus produced questions would be students’ work. Participants were informed that the introduced online question-generation and peer-assessment activity was intended to augment their science learning.

For the duration of the study, once each week, after attending three instructional sessions allocated for science, students participated in a 40-minute learning activity in a computer laboratory. To ensure that all participants possessed the fundamental skills required for the introduced activity, a training session on generating questions and assessing peers’ work (the generated questions in this case) was arranged at the beginning of the study. Considering that true/false and multiple-choice questions are among the most frequently encountered question types in primary schools in Taiwan, these two types of question-generation options were chosen. Each week students were directed to individually generate at least one question each type regarding the instructional content covered that week, as well as to assess at least two questions from a pool of peer-generated questions for each question type.

Online learning system

A learning environment (Question-Authoring and Reasoning Knowledge System, QuARKS) that allows students to contribute to and benefit from the process of constructing questions and receiving feedback from their peers about the composed questions was used (Yu, 2009). Essentially, the question-generation sub-system allows multimedia files to be included as parts of questions, as well as the use of various fonts, text sizes and styles (see Figure 1).

The peer-assessment sub-system, on the other hand, allows assessors to give evaluative feedback using an online assessment form. Using the form’s built-in criteria (strengths: concise question-stem and options, important
concepts, well-explained notes; weaknesses: unclear question-stems, overly-complicated question-stems, excessively
verbose options, multiple correct answers, elusive phrasing and implausible distracters, the left-bottom portion of
Figure 2), assessors can assess the overall quality of the generated question on a five-point rating scale (from “far
above average” to “far below average”) and rate their recommendation of the question to be included in a drill-and-
practice item bank (from “Highly recommend” to “Do not recommend at all”). Also, assessors can add descriptive
comments and provide detailed suggestions regarding the question being examined in a designated feedback space
(the bottom right portion of Figure 2).

Figure 2. Assessment form for assessors to provide feedback to question-authors

Variables

For the purposes of this study, three variables were included: quantitative ratings of the examined question, the
quality of the descriptive comments assessors provided regarding the assessed question item and question-generation
performance. Quantitative ratings consisted of two indicators: the overall quality of the produced work (i.e.,
generated questions) and recommendation for inclusion in the drill-and-practice database. The overall quality and
recommendations received from assessors were averaged per question per week throughout the activity.

To ensure objective assessment of students’ work (questions generated and descriptive comments provided), a set
of criteria and procedures were developed. The quality of the descriptive comments received from assessors about the
composed questions and student question-generation performance was defined against a set of criteria. For peer-
assessment, all comments question-authors received with regards to a specific question item were analyzed against a
pre-defined scheme and were averaged. The average scores per question per week were then summed up. With
reference to Nelson and Schunn’s (2009) study on feedback and Yelon’s “open communication” (1996) instructional
principle for providing feedback for students’ performances, the quality of descriptive comments was evaluated on
four discrete levels: general comments, specific comments where strengths and weakness are identified,
identification of areas for improvement and explicit suggestions for further refinement of questions. Definitions and
examples of each of the four assessment levels are listed in Table 1.

Table 1. Four levels of descriptive comments, their definition and examples

<table>
<thead>
<tr>
<th>Levels of descriptive comments</th>
<th>Definition</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>General comments</td>
<td>Comments provided only contained built-in comments directly from the system or were general in nature.</td>
<td>✓ Important concepts. ✓ Good! ✓ Lively.</td>
</tr>
<tr>
<td>Specific comments where strengths or weaknesses are identified</td>
<td>The strengths and/or weaknesses of the posed question were clearly identified and explained in the assessor’s own words.</td>
<td>✓ Each of the options seems to be the correct answer. They are all pretty good distracters!!! ✓ The description of the question item is concise and easy to understand. The picture included in the question item is illustrative. ✓ The question was copied directly from the textbook. ✓ It is good that each option was presented with pictures, but the question item is not interesting.</td>
</tr>
<tr>
<td>Identification of areas</td>
<td>Not only were weaknesses</td>
<td>✓ The way you framed your question and grouped your</td>
</tr>
</tbody>
</table>
for improvement of the posed question identified, areas for improvement were offered.

- It’s good that the tested concept is important (i.e., noise and its measuring unit). However, it tested trivia facts that no one is expected to memorize (what is the decibel level of noise observed at flight take-off).

Explicit suggestions for further refinement of questions

- Explicit suggestions for further refinement of the question were provided.

- It is nice that you used a character in your question, but your question is not clear enough. Might change it to: Mickey feels the construction site is too noisy. What is most likely the decibel level of noise at the construction site? Also, you may consider adding a picture depicting the surroundings of the construction site.

- The scenario and characters in the question stem appeals to the reader. The tested concept (wind instrument) is an important topic to be mastered. However, the options are too simple. People can easily find the answer right. Try to include borderline examples or other less seen instruments (e.g., wooden fish or tambourine) to increase its item difficulty.

In reference to the Torrance creativity index (1974) and King’s question cognitive levels (1992) and questions generated by students, the following criteria were adopted to assess students’ performances in question-generation: fluency, flexibility, elaboration, originality, cognitive level and importance. Each of the indexes was operationally defined to ensure objective assessment. Specifically, each question was graded along the following six dimensions:

- **Fluency (0-3)** assesses the correctness of wording and punctuation, clarity of meaning and logic, and the relevancy of the constructed question.
- **Flexibility (0-2)** gauges the interconnectedness between the currently covered topic/unit and prior topics/units and any self-derived examples.
- **Elaboration (0-2)** assesses the refinement of the questions in terms of creating scenarios for the question, using multi-media files to enhance understanding of the question, the discrimination of alternatives, and so on.
- **Originality (0-2)** examines the uniqueness of a specific question as compared to those of peers (i.e., innovative methods of formulating questions or content/ideas).
- **Cognitive Level (0-3)** evaluates the cognitive level demanded of the question author: fact, comprehension, or integration (King, 1992). Comprehension indicates that students used their own words to define or describe learned content, whereas fact indicates that wording in the question is very similar to that used in the instructional content. Integration indicates that a link has been built across topics/units and that explanations have been provided to build connections.
- **Importance (0-1)** evaluates the importance of the concepts assessed in the constructed question.

To establish inter-rater reliability, ten pieces of student work (questions generated and comments received about the question) created during two out of the six implementation weeks were randomly selected from eight participating classes and evaluated by another independent rater (N = 160). The results of the inter-rater reliability were $r = 0.73$, $p < 0.01$ and $r = 0.94$, $p < 0.01$ for question-generation and peer-assessment, respectively. They proved to be satisfactory.

### Results

**Descriptive statistics and relationships among examined variables**

The means and standard deviations of the quality of feedback received on the composed questions (including quantitative peer-ratings and descriptive comments) and students’ question-generation performance are listed in Table 2.
Table 2. Descriptive statistics and correlations between variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Quantitative ratings</th>
<th>Descriptive comments</th>
<th>Produced Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative ratings</td>
<td>1</td>
<td>0.1</td>
<td>0.28**</td>
</tr>
<tr>
<td>Descriptive comments</td>
<td>0.1</td>
<td>1</td>
<td>0.37**</td>
</tr>
<tr>
<td>Produced work</td>
<td>0.28**</td>
<td>0.37**</td>
<td>1</td>
</tr>
</tbody>
</table>

Mean (SD) 3.45 (0.68) 6.70 (3.60) 32.11 (13.56)

Note. N = 233, * p < 0.05, ** p < 0.01

The predictive effect of quantitative ratings on the quality of produced work

The peer-assessment activity following the question-generation activity allowed question-assessors to rate its quality and to make a recommendation on the assessed question. A preliminary analysis using Pearson correlations was conducted to determine the relatedness of the quantitative ratings and question-generation performance. The quantitative ratings were significantly correlated with question-generation performance ($r = 0.28, p < 0.01$); therefore, this variable was included in the regression analysis. The regression results presented in Table 3 indicates that the quantitative ratings significantly predict question-generation performance ($\beta = 0.28, p < 0.01$).

Table 3. Regression analysis for quantitative ratings predicting the quality of produced work

<table>
<thead>
<tr>
<th>Model</th>
<th>B</th>
<th>SEB</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>12.72</td>
<td>4.46</td>
<td></td>
</tr>
<tr>
<td>Quantitative ratings</td>
<td>5.60</td>
<td>1.27</td>
<td>0.28**</td>
</tr>
<tr>
<td>R-square</td>
<td></td>
<td></td>
<td>0.08</td>
</tr>
<tr>
<td>F</td>
<td></td>
<td></td>
<td>19.59**</td>
</tr>
</tbody>
</table>

Note. a. Predictor: (Constant), Quantitative ratings  
  b. Dependent variable: Question-generation performance  
  c. * p < 0.05, **p < 0.01

The predictive effect of the quality of descriptive comments on the quality of produced work

The same preliminary analysis using Pearson correlations was conducted to determine the relatedness of the quality of descriptive comments and question-generation performance. The quality of descriptive comments was found to be significantly correlated with question-generation performance ($r = 0.37, p < 0.01$); therefore, this variable was included in the regression analysis. The regression results presented in Table 4 indicate that the quality of descriptive comments significantly predicts question-generation performance ($\beta = 0.37, p < 0.01$).

Table 4. Regression analyses for the quality of descriptive comments ability to predict the quality of produced work

<table>
<thead>
<tr>
<th>Model</th>
<th>B</th>
<th>SEB</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>22.80</td>
<td>1.74</td>
<td></td>
</tr>
<tr>
<td>Quality of descriptive comments</td>
<td>1.39</td>
<td>0.23</td>
<td>0.37**</td>
</tr>
<tr>
<td>R square</td>
<td></td>
<td></td>
<td>0.14</td>
</tr>
<tr>
<td>F</td>
<td></td>
<td></td>
<td>36.48**</td>
</tr>
</tbody>
</table>

Note. a. Predictor: (Constant), Quality of descriptive comments  
  b. Dependent variable: Question-generation performance  
  c. * p < 0.05, **p < 0.01

The combined predictive effect of quantitative ratings and quality of descriptive comments received on the quality of produced work

To avoid multicollinearity, a preliminary analysis using Pearson correlations was conducted to determine the relatedness of the quantitative ratings and the quality of descriptive comments. The quantitative ratings were not...
correlated with the quality of descriptive comments ($r = 0.1, p = 0.13$); therefore, these two variables could both be included in the multiple regression analysis. As shown in Table 5, the quality of descriptive comments significantly predicted a significant proportion of variance on students’ question-generation performance ($R^2 = 0.14, F = 36.48, p < 0.01$). Adding the variable of quantitative ratings significantly enhanced the R-square ($R^2$ change = 0.06, $F = 16.98, p < 0.01$); therefore, it was determined that together the qualities of descriptive comments and quantitative ratings can significantly predict question-generation performance ($\beta_{\text{qual}} = 0.35, p < 0.01; \beta_{\text{quan}} = 0.25, p < 0.01$, respectively).

**Table 5. Multiple regression analyses of the quality of feedback’s ability to predict the quality of produced work**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th></th>
<th></th>
<th>Model 2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>SEB</td>
<td>$\beta$</td>
<td>B</td>
<td>SEB</td>
<td>$\beta$</td>
</tr>
<tr>
<td>Constant</td>
<td>22.80</td>
<td>1.75</td>
<td>6.44</td>
<td>4.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quality of descriptive comments</td>
<td>1.39</td>
<td>0.23</td>
<td>0.37**</td>
<td>1.30</td>
<td>0.22</td>
<td>0.35**</td>
</tr>
<tr>
<td>Quantitative ratings</td>
<td></td>
<td></td>
<td></td>
<td>4.91</td>
<td>1.19</td>
<td>0.25**</td>
</tr>
<tr>
<td>R-square</td>
<td>0.14</td>
<td></td>
<td></td>
<td>0.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F for change in R-square</td>
<td>36.48**</td>
<td></td>
<td></td>
<td>16.98**</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note. a. Dependent variable: Question-generation performance

$b.* p <0.05, **p <0.01$

**Discussion and conclusions**

Numerous online learning systems have been developed for learners to interact with their peers online to improve the produced work through peer-assessment. This study explored whether different types of feedback received from peers contributed to the quality of produced work (in this specific case, question-generation performance). Moving beyond the frequently adopted research paradigm that examines assesses’ subjective attitudes toward peer-assessment (Topping, 2010), objective measures were used to examine whether any predicative relationships existed between examined variables.

This study found that quantitative ratings and quality of descriptive comments that question-authors received from peers contributed to question-generation performance both individually and together. Specifically, question-authors who received higher quantitative ratings tended to demonstrate better performance composing questions. Also, better quality descriptive feedback received for composed questions lead to better performance on question-generation tasks. Furthermore, question-authors who received higher quality descriptive feedback together with higher ratings on their questions tended to have better question-generation performances.

The obtained findings have important empirical significance and implications for online system development and instructional implementation. Quantitative ratings and descriptive comments are both forms of feedback that can be easily offered for the author’s consideration during online peer-assessment; however, their effects on the product quality have not yet been substantiated. The present study, to the best of the researchers’ knowledge, evidenced the individual and combined effects of both types of feedback for the first time.

Based on the findings of this study, several suggestions are provided. First, in view of the significant combined predicative effects of quantitative ratings and descriptive feedback, developers of online peer-assessment systems are advised to include both designs in the system to achieve optimum performance. This advice is particularly crucial for online systems that currently incorporate only quantitative ratings for feedback to authors.

Furthermore, this study found that the descriptive feedback variable explained more variance in question-generation performance than quantitative rating, so the importance of providing question-authors with descriptive feedback should not be ignored, especially with regard to those who have received relatively lower quantitative ratings from their peers. As indicated, question-authors who received higher quantitative ratings were found in this study to demonstrate better performance composing questions. Nonetheless, looking at the flip side of the result leads to the conclusion that question-authors who received lower quantitative ratings tended to demonstrate lower question composing performance. As the degree of explicitness and specifics are different between quantitative ratings and descriptive comments, it will understandably be hard for authors who receive lower ratings to improve their performances without elaborate explanations or detailed recommendations from assessors. Providing descriptive...
comments to authors should be emphasized to assessors when a lower than average quality of produced work is observed.

Moreover, as has been found, providing a summary of the performance with the locations of specific problems and solutions or explanations is associated with increased understanding (Nelson & Schunn, 2009). Training sessions by instructors about providing quality descriptive comments are essential to maximize the effect of peer-assessment.

Suggested topics for future studies are as follows. As mentioned, the effects of different types of feedback have been the focus of recent research on peer-assessment. While studies along these lines help to provide empirical evidence and build guidelines for constructive peer-assessment learning environments, some areas require further investigation. First, while this study supported the individual and collective predictive effects of quantitative ratings and descriptive comments on the quality of produced work, since that the two different types of feedback demand different amounts of effort from participants (quantitative ratings presumably require less mental effort and time than descriptive comments) and differ greatly in terms of the extent of explicitness of provided information (with descriptive comments providing more specific suggestions and thus presumably offering more support than quantitative ratings), an investigation into the effects of different types of feedback on participation/engagement and interaction cycles, students preference and overall satisfaction, and so on, using an experimental research design method, would have pedagogical and research significance.

Finally, previous research has indicated that the impacts of different types of descriptive feedback (e.g., elaborate and specific versus concise and general) on performance and produced work varied with the competency levels of students involved in the peer-assessment activities. For instance, Strijbos and his colleagues’ study (2010) indicated that concise general feedback from the low-competence assessors resulted in positive learning gains for feedback receivers. On the other hand, Gielen et al (2010) found that high-competency assesses seem to rely less on elaborate specific feedback. With these research findings in mind, further investigation into how the competency level of assessors and assessees interact with types of feedback and influence performance would be practically and empirically relevant.

Acknowledgement

This paper was funded by research grants from the National Science Council, Taiwan, ROC (NSC 96-2520-S-006-002-MY3; NSC 99-2511-S-006-015-MY3).

References


