ELSEVIER

Learning and Instruction 20 (2010) 265–269

Learning and Instruction

www.elsevier.com/locate/learninstruc

Guest Editorial

Unravelling peer assessment: Methodological, functional, and conceptual developments

Jan-Willem Strijbos [a,*], Dominique Sluijsmans [b,c,d]

[a] Centre for the Study of Learning and Instruction, Institute of Education and Child Studies, Faculty of Social and Behavioral Sciences, Leiden University, P.O. Box 9555, 2300 RB, Leiden, The Netherlands

[b] Faculty of Education, HAN University of Applied Sciences, P.O. Box 30011, 6503 HN, Nijmegen, The Netherlands

[c] Centre of Learning Sciences and Technologies, Open University of the Netherlands, P.O. Box 2960, 6401 HL Heerlen, The Netherlands

[d] Department of Educational Development and Research, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, The Netherlands

Abstract

Peer assessment is an educational arrangement where students judge a peer's performance quantitatively and/or qualitatively and which stimulates students to reflect, discuss and collaborate. However, empirical evidence for peer assessment effects on learning is scarce, mostly based on student self-reports or involving comparison of peers' and teachers' ratings or anecdotal evidence from case studies. Systematic investigation of learning effects necessitates methodological, functional, and conceptual development in peer assessment research. This implies sound (quasi-)experimental studies, the definition of specific peer assessment mechanisms, and affiliations with other research domains. The articles in this special issue address these three needs and offer new directions for research.

Keywords: Peer assessment; Peer feedback; Quasi-experimental research; Collaborative learning; Formative assessment

1. Introduction

In the past two decades a conceptual shift has occurred in the practice of assessment, from a teacher-directed perspective to one that involves students in the assessment process (Boud, 1995), or in other words the shift from a testing culture to an assessment culture (Birenbaum, 2003). However, the effectiveness of any assessment depends on the quality of assessment and how it is incorporated by students in subsequent performance, or more specifically: why, what, when, how and who should (be) assess(ed) (Segers, Dochy, & Cascallar, 2003).

In a testing culture the main purpose of an assessment is to make evaluative decisions for summative purposes. Shortcomings of summative assessment are that it is decontextualised and individualistic, isolated from the learning process; moreover, it takes place only at the end of a course to judge how well a student performed. Summative assessment focuses strongly on the cognitive aspects of learning, often applies a single performance score, and it is designed and conducted by the instructor. In contrast, the features of an assessment culture are that an assessment does not only serve summative but also, and to a large extent, formative purposes. Formative assessment is contextualised and aims to build a comprehensive picture of learners' characteristics. It is an integral part of a learning process, and it takes place several times during a course rather than only at the end. Formative assessment focuses on cognitive, social, affective, and meta-cognitive aspects of learning, often applies a multi-method approach and it leads to a profile instead of a single score. Most notably, the students are actively involved in the assessment process, for example through negotiation of the criteria, the design of the assessment and/or the interpretation and value of the assessment for performance improvement.

Despite the increase in formative assessment, the important role of summative assessment should be acknowledged through a unified approach using both traditional (summative)

* Corresponding author. Tel.: +31 71 5274048; fax: +31 71 5273619.
E-mail address: jwstrijbos@fsw.leidenuniv.nl (J.W. Strijbos).

0959-4752/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.learninstruc.2009.08.002

and progressive (formative) perspectives (Shute, 2007). Moreover, any assessment involves the use of feedback information and whether this use is more summative or formative, is an issue of interpretation rather than one of absolutes (Hattie, 2003). However, in contrast to the large body of research on summative assessment methods (educational as well as psychometric) research on formative assessment practices, although it is accumulating, is still developing. This special issue presents six studies that address current developments in the context of one such progressive mode of assessment, namely peer assessment.

2. Peer assessment practices

Peer assessment is an educational arrangement where students judge a peer's performance quantitatively, by providing a peer with scores or grades, and/or qualitatively, by providing the peer with written or oral feedback (Topping, 1998). Peer assessment stimulates students to share responsibility, reflect, discuss and collaborate (Birenbaum, 1996; Boud, 1990; Orsmond, Merry, & Callaghan, 2004; Sambell & McDowell, 1998).

Peer assessment practices vary along a wide array of characteristics. Topping (1998) derived 17 characteristics from a literature review, which were subsequently ordered in four clusters by Van den Berg, Admiraal, and Pilot (2006), and further expanded by Gielen (2007) and Strijbos, Ochoa, Sluijsmans, Segers, and Tillema (2009). The variety in characteristics of peer assessment is reflected in peer assessment reviews which reveal a high level of diversity and ambiguity in peer assessment practices, making it very difficult to understand how peer assessment contributes to learning (Dochy, Segers, & Sluijsmans, 1999; Sluijsmans, Dochy, & Moerkerke, 1999; Topping, 1998, 2003; Van Gennip, Segers, & Tillema, 2009) and poses problems for wider generalisation. In other words, we need to address the gap between what we know about peer assessment and what we claim in general about the benefits of peer assessment for learners.

3. Expanding peer assessment research: new lenses and a pair of shades

For peer assessment research to advance and systematically unravel the mechanisms that foster student learning there is a need to include a wider variety of studies. First, there is a need for *methodological* development, that is, an increase of (quasi-)experimental studies that investigate the effects of specific peer assessment mechanisms on learning. Second, more rigour is required regarding the operationalisation and purpose of peer assessment (*functional* development). Third, *conceptual* development is needed through the affiliation with related research domains.

3.1. Need for methodological development

So far, the majority of peer assessment studies have collected students' self-reports of their learning in peer assessment practices. Although these studies provide a valuable description of students' learning experience and insights into peer assessment practices (Papinczak, Young, & Groves, 2007; Sivan, 2000; Smith, Cooper, & Lancaster, 2002), the specificity of these practices prevents the generalisation for learning from peer assessment.

Despite Topping's (1998) call, peer assessment research that applies a control group or an (quasi-)experimental design – enabling investigation of the relation between peer assessment methods, mechanisms and outcomes – is still very small. It should be noted that ecologically valid research settings complicate the inclusion of a genuine control group, and that a mixed-method approach or triangulation of multiple and diverse studies is proposed (Kember, 2003). Nevertheless, (quasi-)experimental studies enable the investigation of specific mechanisms in relation to specific outcomes (i.e., hypothesis testing) compared to descriptive studies (hypothesis generation) and offer a complementary perspective (Strijbos & Fischer, 2007).

Methodological developments in peer assessment research can include, but are not limited to, greater variety in (a) research designs, (b) research instruments, and (c) analytic techniques. In addition, establishing quality criteria for peer assessment research, for example criteria for transparency and interpretation of research findings, could foster generalisability.

3.2. Need for functional development

The summative application of peer assessment, through a comparison of peer and teacher ratings, has been – and still is – a strong focus in research on peer assessment. Falchikov and Goldfinch (2000) reviewed 48 studies and concluded that peer ratings were highly correlated with teacher ratings ($r = .69$). More recently, Cho, Schunn, and Wilson (2006) found that the aggregate of at least four peer assessments were as reliable and valid as teacher assessments, whereas the reliability and validity of single peer assessments were much lower – presumably because students evaluate a subset of all teacher assessments and as a consequence develop different evaluative perspectives that are reflected in rating variability. Irrespective of the findings on the reliability of peer ratings versus teacher ratings (Cho et al., 2006; Falchikov & Goldfinch, 2000; Magin, 2001; Stefani, 1994; Topping, 2003; Zhang, Johnston, & Kilic, 2008), similarity in peer and teacher ratings provides no information as to whether the ratings affect students' subsequent performance. It is implicitly assumed that the high degree of similarity between peer and teacher ratings reflects rating fairness and that student responses to peer ratings will be similar to their responses to teacher ratings.

Peer assessment practices – as reviewed in Falchikov and Goldfinch (2000) – have hardly evolved beyond a summative and quantitative view of peer assessment with a strong reliance on scoring and grading. Moreover, within these practices peer assessment is disconnected from the instructional setting and results in a lack of 'constructive alignment' (Biggs, 1996). Hence, a wider variety of operationalisations of

peer assessment, that is, clear definitions of the purpose of a specific peer assessment practice in relation to the anticipated student learning, is needed. A point in case is the increased focus on the content of peer feedback in the context of peer assessment (Nelson & Schunn, 2009; Prins, Sluijsmans, & Kirschner, 2006).

3.3. Need for conceptual development

In parallel to a strong focus on the summative aspect (see Section 3.2), peer assessment has been approached as an assessment issue rather than as an interactive and communicative process in the service of learning. In other words, most peer assessment practices limit peer assessment to a one-off event, rather than approaching the peer assessment as a cyclical and interactive process (Strijbos et al., 2009).

Furthermore, social aspects of peer assessment – labelled as "reciprocity effects" (Cheng & Warren, 1997; Pond, Ul-Haq, & Wade, 1995; Williams, 1992) – have been approached from a control perspective, that is, controlling (a) high ratings to friends, (b) high ratings to fellow group members, (c) high ratings to dominant group members, and (d) high profit from effort invested by fellow group members. Moreover, reciprocity effects are commonly considered assessment errors since they decrease the reliability of peer assessment (Magin, 2001), rather than investigating how these social aspects affect the peer assessment process and subsequent student performance and learning.

Peer assessment is increasingly applied to evaluate the collaborative process during group work (Sluijsmans & Strijbos, 2010). Interestingly, neither the interactive opportunities offered by the collaborative settings are applied to peer assessment practice (Strijbos et al., 2009), nor are important aspects from collaborative learning research applied to the study of peer assessment.

4. Meeting the three needs: overview of the contributions to this special issue

Regarding the hypothesised learning benefits of peer assessment it is crucial to determine systematically the mechanisms that influence students' performance and learning. This requires a richer and broader perspective, and necessitates methodological, functional and conceptual development – as reflected in the studies of this special issue.

The contribution by Van Zundert, Sluijsmans, and Van Merriënboer (2010) acts as a rationale for the empirical studies applying (quasi-)experimental designs (methodological development) and signals the need for functional and conceptual development as well. A thorough literature review of 26 studies, focused on the relations between methods, conditions and outcomes, identified four variables that contribute to effective peer assessment: (a) psychometric qualities; (b) domain-specific skills; (c) peer assessment skills, and (d) student attitudes. The literature review also underlines the variety in peer assessment practices and the lack of transparency in methods, conditions and outcomes. The empirical studies in this special issue provide a first coherent set of (quasi-)experimental studies, where methods and conditions are clearly described and related to outcome variables.

The study by Van Gennip, Segers, and Tillema (2010) contributes especially to the functional and conceptual development of peer assessment research, by examining the role of psychological safety, value diversity, interdependence, trust, and peer assessment conceptions (conceptual), during peer assessment of team work (functional) in vocational education. The experimental group received training on peer assessment and variables of interest were measured before and after a six week project; the control group received no training. The results indicated change in psychological safety, value diversity, and trust in the peer as an assessor. Peer assessment contributed to psychological safety and lower value diversity – students' perceived learning was predicted by value diversity and conceptions. The study shows that the social aspects of peer assessment are not detrimental by definition – as is presumed for 'reciprocity effects' – and thus that they need not be automatically 'controlled' for that reason.

The contribution by Strijbos, Narciss, and Dünnebier (2010) adds to the functional development of peer assessment research by investigating the impact of various contents of feedback, and to conceptual development by investigating the impact of sender's competence level in the context of peer assessment of academic writing. Students were assigned to four experimental and a control group; the experimental groups received a scenario with either Concise General (CGF) or Elaborated Specific (ESF) feedback by a high or low competent peer. The study revealed that ESF by a high competent peer was perceived as more adequate, but led to more negative affect. CGF groups outperformed ESF groups during treatment, whereas during the posttest, groups with a low competent peer outperformed the groups with a high competent peer. This study clearly shows that sender's competence level affects student perceptions, and negates the implicit assumption that correlated peers' and teacher' ratings lead to similar reactions.

The study by Gielen, Peeters, Dochy, Onghena, and Struyven (2010) also examined the effectiveness of peer feedback for learning by focusing on the core characteristics of constructive peer feedback (functional development). In addition, an instructional intervention, which aimed to support the use of the feedback by asking assesses to reflect upon the feedback after peer assessment, was studied (conceptual development). This study clearly shows that the characteristics of the peer feedback content and style of the provided feedback, in particular justification, can play a significant role in a peer assessment exercise. Moreover, the instructional intervention used was akin to 'scripts' that are widely used in collaborative learning research, reflecting that findings from specific scripts and script design features are highly relevant for peer assessment research.

The contribution by Van Steendam, Rijlaarsdam, Sercu, and Van den Bergh (2010) further aids our understanding of peer feedback by examining the effects of instruction type (observation versus practising) on higher-order peer feedback

(functional development), and whether the subsequent emulation in dyads or individually would be more efficient (conceptual development) for the quality of revision in the context of English as a Foreign Language (EFL). Results showed a significant interaction of instruction and emulation. If emulation takes place individually, then observation and practice are equally effective for strategy acquisition. For dyadic emulation to be productive, it needs to be preceded by observation. This study clearly shows that not only the interactive aspects of receiving peer feedback affect students' performance, but that the effectiveness of an instructional format, aimed at teaching revision criteria in order to stimulate more higher-order peer feedback when evaluating a peer's text, is affected by whether the students work individually or in dyads.

The study by Cho and MacArthur (2010) contributes to our understanding of both the role of peer feedback (functional development), and the pivotal issue of peer versus expert feedback as well as whether peer feedback by multiple peers is more efficient for subsequent revision of a research proposal (conceptual development). Students received either feedback from a single expert, a single peer, or multiple peers. The findings revealed that the students receiving feedback from multiple peers received more feedback of all types (directive, non-directive, and praise). Non-directive feedback predicted complex repair revision that students in the multiple peers group made more than both other groups. This study clearly showed that students perform better at using feedback from their peers rather than feedback by a subject-matter expert. These findings signify that the specific contribution of peers to peer assessment is yet to be determined.

The commentary by Topping (2010) summarises the strengths and weaknesses of each contribution from a methodological perspective, whereas the commentary by Kollar and Fischer (2010) focuses on functional and conceptual development of peer assessment. They propose a process-related model of peer assessment (functional development), by using evidence from collaborative learning research (conceptual development). Both commentaries stress the need for embedding peer assessment research in a broader scientific framework.

5. Conclusion

Obviously, qualitative, and non-(quasi-)experimental research designs and their associated analytic techniques are equally important for peer assessment research to advance. The rich description in case studies of specific peer assessment settings provides a wealth of evidence for hypothesis generation, which can subsequently be tested in (quasi-)experimental settings. In sum, the six contributions provide an instructive overview of current (quasi-)experimental research on peer assessment, which may stimulate a wider adoption of (quasi-)experimental studies enabling the investigation of specific components and conditions derived from case studies. With this special issue, we attempted to break new ground in peer assessment research regarding the methodology of peer assessment research, the function of peer assessment and the affiliation with other research disciplines. The commentaries also signify that the variety of variables, contexts and domains as presented in the contributions provide a fruitful footing to advance the science of peer assessment.

Acknowledgements

The guest editors would like to thank the reviewers and Prof. Dr. Anastasia Efklides, the editor of Learning and Instruction, for their contribution and support to this special issue.

References

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*, 347–364.

Birenbaum, M. (1996). Assessment 2000: towards a pluralistic approach to assessment. In M. Birenbaum, & F. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 3–29). Boston, MA: Kluwer.

Birenbaum, M. (2003). New insights into learning and teaching and their implications for assessment. In M. Segers, F. Dochy, & E. Cascallar (Eds.), *Optimising new modes of assessment: In search of qualities and standards* (pp. 13–37). Dordrecht, The Netherlands: Kluwer.

Boud, D. (1995). *Enhancing learning through self-assessment*. London: Kogan Page.

Boud, D. J. (1990). Assessment and promotion of academic values. *Studies in Higher Education, 15*, 101–113.

Cheng, W., & Warren, M. (1997). Having second thoughts: student perceptions before and after a peer assessment exercise. *Studies in Higher Education, 22*, 233–239.

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction, 20*(4), 328–338.

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98*, 891–901.

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*, 331–350.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: a meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*, 287–322.

Gielen, S. (2007). *Peer assessment as a tool for learning*. Unpublished doctoral dissertation, Leuven University, Leuven, Belgium.

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*(4), 304–315.

Hattie, J. (2003). *Formative and summative interpretations of assessment information*. Retrieved July 6, 2009, from. http://www.education.auckland.ac.nz/webdav/site/education/shared/hattie/docs/formative-and-summative-assessment-(2003).pdf.

Kember, D. (2003). To control or not to control: the question of whether experimental designs are appropriate for evaluating teaching innovations in higher education. *Assessment and Evaluation in Higher Education, 28*, 89–101.

Kollar, I., & Fischer, F. (2010). Commentary – peer assessment as collaborative learning: a cognitive perspective. *Learning and Instruction, 20*(4), 344–348.

Magin, D. J. (2001). A novel technique for comparing the reliability of multiple peer assessments with that of single teacher assessments of group work. *Assessment and Evaluation in Higher Education, 26*, 139–152.

Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: how different types of peer feedback affect writing performance. *Instructional Science, 37*, 375–401.

Orsmond, P., Merry, S., & Callaghan, A. (2004). Implementation of a formative assessment model incorporating peer and self-assessment. *Innovations in Education and Teaching International, 41*, 273–290.

Papinczak, T., Young, L., & Groves, M. (2007). Peer assessment in problem-based learning: a qualitative study. *Advances in Health Sciences Education, 12*, 169–186.

Pond, K., Ul-Haq, R., & Wade, W. (1995). Peer review: a precursor to peer assessment. *Innovations in Education and Training International, 32*, 314–323.

Prins, F. J., Sluijsmans, D. M. A., & Kirschner, P. A. (2006). Feedback for general practitioners in training: styles, quality and preferences. *Advances in Health Science Education, 11*, 289–303.

Sambell, K., & McDowell, L. (1998). The value of self and peer assessment to the developing lifelong learner. In C. Rust (Ed.), *Improving student learning – Improving students as learners* (pp. 56–66). Oxford, UK: Oxford Centre for Staff and Learning Development.

Segers, M., Dochy, F., & Cascallar, E. (Eds.). (2003). *Optimising new modes of assessment: In search of qualities and standards*. Dordrecht, The Netherlands: Kluwer.

Shute, V. J. (2007). Tensions, trends, tools, and technologies: time for an educational sea change. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 139–187). Mahwah, NJ: Erlbaum.

Sivan, A. (2000). The implementation of peer assessment: an action research approach. *Assessment in Education: Principles, Policy and Practice, 7*, 193–213.

Sluijsmans, D., Dochy, F., & Moerkerke, G. (1999). Creating a learning environment by using self-, peer- and co-assessment. *Learning Environments Research, 1*, 293–319.

Sluijsmans, D. M. A., & Strijbos, J. W. (2010). Flexible peer assessment formats to acknowledge individual contributions during (web-based) collaborative learning. In B. Ertl (Ed.), *E-collaborative knowledge construction: Learning from computer-supported and virtual environments* (pp. 139–161). Hershey, PA: IGI Global.

Smith, H., Cooper, A., & Lancaster, L. (2002). Improving the quality of undergraduate peer assessment: a case for student and staff development. *Innovations in Education and Teaching International, 39*, 71–81.

Stefani, L. A. J. (1994). Peer, self and tutor assessment: relative reliabilities. *Studies in Higher Education, 19*, 69–75.

Strijbos, J. W., & Fischer, F. (2007). Methodological challenges for collaborative learning research. *Learning and Instruction, 17*, 389–393.

Strijbos, J. W., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: are they critical for feedback perceptions and efficiency? *Learning and Instruction, 20*(4), 291–303.

Strijbos, J. W., Ochoa, T. A., Sluijsmans, D. M. A., Segers, M. S. R., & Tillema, H. H. (2009). Fostering interactivity through formative peer assessment in (web-based) collaborative learning environments. In C. Mourlas, N. Tsianos, & P. Germanakos (Eds.), *Cognitive and emotional processes in web-based education: Integrating human factors and personalization* (pp. 375–395). Hershey, PA: IGI Global.

Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*, 249–276.

Topping, K. J. (2003). Self and peer assessment in school and university: reliability, validity and utility. In M. Segers, F. Dochy, & E. Cascallar (Eds.), *Optimising new modes of assessment: In search of qualities and standards* (pp. 55–87). Dordrecht, The Netherlands: Kluwer.

Topping, K. J. (2010). Commentary – methodological quandaries in studying process and outcomes in peer assessment. *Learning and Instruction, 20*(4), 339–343.

Van den Berg, I., Admiraal, W., & Pilot, A. (2006). Design principles and outcomes of peer assessment in higher education. *Studies in Higher Education, 31*, 341–356.

Van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2009). Peer assessment for learning from a social perspective: the influence of interpersonal variables and structural features. *Educational Research Review, 4*, 41–54.

Van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2010). Peer assessment as a collaborative learning activity: the role of interpersonal variables and conceptions. *Learning and Instruction, 20*(4), 280–290.

Van Steendam, E., Rijlaarsdam, G., Sercu, L., & Van den Bergh, H. (2010). The effect of instruction type and dyadic or individual emulation on the quality of revision in ESL. *Learning and Instruction, 20*(4), 316–327.

Van Zundert, M., Sluijsmans, D. M. A., & Van Merriënboer, J. J. G. (2010). Effective peer assessment processes: research findings and future directions. *Learning and Instruction, 20*(4), 270–279.

Williams, E. (1992). Student attitudes towards approaches to learning and assessment. *Assessment and Evaluation in Higher Education, 17*, 45–58.

Zhang, B., Johnston, L., & Kilic, G. B. (2008). Assessing the reliability of self- and peer rating in student group work. *Assessment and Evaluation in Higher Education, 33*, 329–340.