ELSEVIER

Learning and Instruction 20 (2010) 270–279

Learning and Instruction

www.elsevier.com/locate/learninstruc

Effective peer assessment processes: Research findings and future directions

Marjo van Zundert*,[1], Dominique Sluijsmans, Jeroen van Merriënboer

Open University of the Netherlands, Centre for Learning Sciences and Technologies, P.O. Box 2960, 6401 DL, Heerlen, The Netherlands

Abstract

Despite the popularity of peer assessment (PA), gaps in the literature make it difficult to describe exactly what constitutes effective PA. In a literature review, we divided PA into variables and then investigated their interrelatedness. We found that (a) PA's psychometric qualities are improved by the training and experience of peer assessors; (b) the development of domain-specific skills benefits from PA-based revision; (c) the development of PA skills benefits from training and is related to students' thinking style and academic achievement, and (d) student attitudes towards PA are positively influenced by training and experience. We conclude with recommendations for future research.

Keywords: Peer assessment; Development of peer assessment skills; Attitudes towards peer assessment; Training of peer assessment skills

1. Introduction

Due to the growing complexity of the workplace and professional tasks, modern education increasingly aims at self-directed and collaborative learning (Boud, Cohen, & Sampson, 1999). Because self-directed learning implies that learners be actively involved in shaping their own learning processes, and collaborative learning implies joint effort in carrying out tasks, peer assessment (PA) fits these new goals. PA can be described generally as a process whereby students evaluate, or are evaluated by, their peers. In educational practice, this occurs in many different forms. Several types of PA exist, such as grading a peer's research report, providing qualitative feedback on a classmate's presentation, or evaluating a fellow trainee's professional task performance.

In all its forms, PA has become increasingly popular in education. As a learning tool, assessing their peers can provide students with skills to form judgements about what constitutes high-quality work (Topping, 1998). As an assessment tool, PA can provide teachers with a more accurate picture of individual performance in group work (Cheng & Warren, 2000).

Despite PA's popularity and advantages, one major problem remains unresolved. At present it is impossible to make claims about what exactly constitutes effective PA; in other words, which PA measures benefit student learning and yield satisfactory psychometric qualities such as reliability and validity. The deadlock is due to an enormous variety both in PA practices and in research on their effects (Van Gennip, Segers, & Tillema, 2009). The conditions under which PA occurs differ, a diversity of methods can be applied, and many different outcomes can emerge. For example, one might imagine that students who already have some experience in assessing their peers (condition) might gain fewer learning benefits (outcome) from extensive assessment training (method) than students who have never assessed their peers before. This multiplicity in itself is positive, that is, PA can be customised to individual needs. However, it does complicate the drawing of inferences about causes and effects. This is because the literature usually describes PA in a holistic fashion, that is, without specifying all the variables present in terms of conditions, methods and outcomes.

Several research reviews have already recognised the large variety in PA practices, but explicit relations between variables that underlie the PA practices, such as conditions, methods,

* Corresponding author. Tel.: +31 43 3885726 (secretariat); fax: +31 43 3885779.

E-mail address: m.vanzundert@educ.unimaas.nl (M. van Zundert).

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

0959-4752/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.learninstruc.2009.08.004

and outcomes, have rarely been investigated (i.e., the variables are not held to account for causes and effects). Topping (1998), for example, provides a comprehensive overview of PA variables in higher education, but no indication of the *relations* between these variables. Hence, under which conditions certain methods result in preferred outcomes remains unknown. The main question of this study was, therefore, "under which specific circumstances are particular types of PA beneficial for particular types of student learning?" and following on from this question another question was posed, namely "what precisely leads to satisfactory psychometric qualities in PA, such as acceptable reliability and validity?" (e.g., correlations between the peers' and the staff's marks).

The added value of this study in comparison to previous reviews is to investigate how PA conditions, methods and outcomes are related, not merely to provide an overview of these variables per se.

2. Methodology

The selected literature had to meet the following criteria: (a) be published between 1990 and 2007; (b) be published in a journal; (c) the journal be listed in the Education and Educational Research domain of the Social Sciences Citation Index; (d) be an empirical study, and (e) the main topic be PA between students in an educational setting (related search terms for the abstracts included peer assessment, peer evaluation, peer ranking, peer rating and peer feedback). The search was conducted in PsycINFO and Academic Search Elite. A subsequent search in ERIC did not lead to additional sources. The procedure resulted in 26 articles for inclusion in the study (see also Table 1 in the Discussion).

The selected literature was analysed to identify which conditions, methods and outcomes were studied, and which relations – if any – between these variables were investigated. The distinction between the three variable groups (conditions, methods and outcomes) is common in instructional design theory (Reigeluth, 1983). The experimental studies were further categorised as either pre-experimental (either a pre- and posttest of one group or posttest only), quasi-experimental (participants were not randomly assigned to the conditions), or true experimental (participants were randomly assigned to experimental and control groups, making it possible to draw inferences in terms of cause and effect with confidence) (Campbell & Stanley, 1963). The overwhelming majority were pre-experimental (mostly case studies).

In the literature analysis, the reported outcome variables were first identified and listed. Based on these listed variables we extracted four variable categories by which the studies could be compared. Some studies reported on the range of marks students used to assess their peers or on differences between student and tutor marks. These and similar outcome variables related to validity and reliability formed the first category, *psychometric qualities of PA*. Besides psychometrics, many studies made claims about learning from PA. Some focused mainly on the quality of students' work, for example their writing performance or science homework assignments. Such outcome variables were included in the second category, *domain-specific skill*. Other studies focused on PA skills, including the quality of students' feedback and feedback styles. These and analogous outcome variables comprised the third category, *PA skill*. Finally, the majority of studies reported on students' views of PA, such as their confidence in assessing their peers and the perceived learning benefits of PA. These formed the fourth category, *student attitudes towards PA*. The results of the current review are addressed according to the four outcome categories (it was possible for a study to report on variables of more than one category). For all studies, conditions and methods will be traced that influence the outcome(s) for each category.

3. Results

3.1. Psychometric qualities of PA

Eight studies reported findings on the psychometric qualities of PA. Two of these showed distinct relations between psychometric qualities and method and/or condition. These two studies are described first, followed by six studies that reported findings on psychometric qualities without ascribing them to specific variables.

Smith, Cooper, and Lancaster (2002) reported positive effects of PA training on psychometric qualities. Prior to their intervention, data were gathered from a cohort of 103 psychology students participating in a particular course. The next year, an intervention was designed for this course in which a second cohort of 90 students received PA training before they conducted the task of marking posters. Students in both cohorts were already familiar with PA. Before the intervention, the students had to assess posters made by their peers on the basis of PA information acquired via a lecture and a handbook (i.e., no active student engagement). In addition to the lecture and handbook, the intervention included a workshop on devising assessment criteria, and a second workshop on applying the criteria. The poster marks pre- and post-intervention were subsequently compared. Analyses revealed that the trained students used an increased range of marks across all posters post-intervention, and less varied marks for each individual poster.

Similarly, Sung, Lin, Lee, and Chang (2003) found experience in PA to positively influence psychometric qualities. In their study, 34 psychology undergraduates were arranged in groups of six to eight to write a research proposal. They had six weeks to prepare their proposals after which these were uploaded onto Web-SPA, a web-based self-assessment and PA system. Students subsequently performed individual self-assessment and PA based on a list of eight items. Responses were given on a 5-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree). Example item was "The research design used proper statistical tests for the hypotheses". The results were discussed within groups and the students were able to re-observe and re-score their proposals. Comparisons with other groups' work were also made and