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

discussed. Subsequently, a group representative presented an oral report and defence of their work, which was commented on by the instructor. Finally, the students revised their proposals according to the feedback received. To test whether the scoring consistency among group members improved after peer interaction, Kendall's coefficients of concordance were calculated on the group members' ratings before and after PA and discussion. The coefficients increased for all groups, which led the authors to conclude that the group members were better able to reach consensus on the quality of their proposals after discussing these proposals as well as the assessment results.

Six other reports on psychometric qualities of PA were found, but these did not ascribe psychometrics to specific methods or conditions. Haaga (1993) identified modest single-reviewer reliability. In this research, 45 psychology students' term papers were both peer and staff assessed. Each student handed in three copies of their own term paper and then anonymously reviewed two of their peers' papers in line with the instructor's guidelines. The instructor subsequently wrote a cover letter to each student author, combining the comments of the two peer reviews and the instructor's feedback. This cover letter and the peer reviews were returned to the student authors, while the instructor also provided each student with feedback on their reviewing. The students were then required to revise their papers on the basis of the feedback and hand them in along with a cover letter explaining if, how, and why they had incorporated the feedback into the revised paper. A single-reviewer reliability of $r = .55$ was found for the peer ratings.

Levine, Kelly, Karakoc, and Haidet (2007) reported modest correlations between PA and traditional forms of student assessment. They gathered data from 152 students enrolled in a psychiatry clerkship. Working in groups, each student was asked to assess their peers' performances, given a "supply" of 10 points per group member. They were required to give at least one member a higher than average score (≥ 11) and at least one a lower than average score (≤ 9). They were also asked to provide their rationales for awarding the points, which could be anonymously delivered to the students as feedback. Pearson correlations between PA scores and traditional forms of student assessments were computed. These traditional forms included the National Board of Medical Examiners' (NBME) psychiatry subject test scores, clinical grades, Individual Readiness Assurance Test (IRAT; Levine et al., 2007) scores and Group Readiness Assurance Test (GRAT; Levine et al., 2007) scores. Correlations between mean PA scores and these traditional test scores were modest to low ($r = .37$, $r = .28$, $r = .41$, and $r = .03$ for the NBME psychiatry subject test scores, clinical scores, IRAT, and GRAT, respectively).

Stefani (1994) also reported a modest correlation between PA scores and traditional examination scores. As part of their training in biochemical techniques, 67 undergraduates wrote practical laboratory reports. These reports were given marks out of 100 by their peers and a tutor according to a marking scheme established by the students. The PA scores of the scientific reports and student rankings after traditionally assessed examinations were then compared, and their correlation found to be modest ($r = .47$). Likewise, the correlation between the tutors' assessment scores of the reports and the traditionally assessed exam scores was modest as well ($r = .58$). The peer and tutor marks were also compared with each other, and although the students used a more restricted range of marks than the tutors, the correlation between the PA and the tutor marks was high ($r = .89$).

Hughes and Large (1993) also found a high correlation between PA and staff assessment. Forty-four pharmacology students gave oral presentations for their classmates and staff – a procedure with which they were familiar. Their peers as well as seven staff members gave marks out of 100 for their communication and presentation skills according to criteria defined in collaboration between the students and staff. The mean values for the two parties were quite close (peer group: $M = 60.2$, $SD = 6.1$; staff group: $M = 63.2$, $SD = 7.8$). The correlation between peer assessments and staff assessments was .83.

Similar to the latter two studies, Smith (1990) found high agreement between peer and staff ratings. Forty-two psychology students participated in debates that were rated by their peers and a tutor on an evaluation sheet which was handed out beforehand. The evaluation sheet consisted of 10 items (e.g., "How effective was the debater's attack?"). The items were to be rated on a 5-point Likert-type scale from 1 (poor) to 5 (good). The correlation between the overall peer score and tutor score was high ($r = .80$), and considered an indicator of the high validity of PA because the staff rating was perceived as the criterion. It was also found that students did not rate leniently, that is, their overall ratings were lower than the tutor ratings.

Magin (2001) examined whether reciprocity effects (i.e., lack of fairness in assessing others due to personal relationships) biased PA, and found these effects to be negligible. Magin (2001) used the PA data of 169 medicine students participating in a course on behavioural studies. The students worked in 16 groups consisting of 9–11 members on a report on a chosen topic related to the course. Their individual group process skills were assessed by their peers and tutors according to two criteria ("contribution to discussion" and "contribution to group development"), which were established in agreement with the students. Reciprocity effects appeared to account for only 1% of the variance in PA scores.

3.1.1. Summary of findings

The psychometric qualities of PA were expressed in several ways and the findings were diverse. However, a high correlation between multiple PA practices is fostered by training and by experience. Training was defined as the provision of information about and preparation for PA such as to actively engage students. Experience was operationalised as having conducted a PA task at least once. Six out of eight studies reported mixed results for different psychometric qualities, without ascribing these qualities to specific methods (e.g., providing training) or conditions (e.g., students' experience

with PA). When expressed in terms of agreement between PA and staff assessment, the psychometrics are generally satisfactory. Correlations between PA practices and traditional examinations, however, appear to be modest.

3.2. Domain-specific skill

Domain-specific skill was reported as the main outcome in five of the selected studies. Four studies showing relations between domain-specific skill and methods and/or conditions are discussed first, followed by one study describing domain-specific skill without relating it to particular methods or conditions.

Olson (1990) found a positive relation between providing students with the opportunity to revise their work on the basis of peer feedback and writing performance. In Olson's elementary school study, 93 sixth-graders participated in six autobiographical writing lessons. They were divided into four groups: (a) the group of "revision instruction/peer partners" (prior to the writing lessons the students were instructed in revision strategies approximately twice a week for one month, then during the writing lessons they met with their peer partners to respond to and revise drafts); (b) the group of "peer partners only" (no revision instruction, but meetings with peer partners); (c) the group of "revision instruction only" (no peer meetings); and (d) the control group (students participated in the same writing lessons, but received neither instructions nor peer partners). Peer feedback appeared to have positive effects on the quality of the students' writing. Those in the group of "revision instruction/peer partners" had significantly higher quality autobiographical writing ($M = 117.26$, $SD = 21.74$ for final drafts) than all other groups. Specifically, the work of students in the group "peer partner only" ranked second in quality ($M = 106.83$, $SD = 19.09$), while students in the group "revision instruction only" and the control group achieved means of 97.13 ($SD = 21.56$) and 97.94 ($SD = 23.25$) respectively. Olson's (1990) study was one of the two studies in the present review that used a true experimental design, the other being the study of Sluijsmans, Brand-Gruwel, Van Merriënboer, and Martens (2004), described in subchapter 3.3.

In Sung et al. (2003) study (see subchapter 3.1), psychology students had to write research proposals and were later given the opportunity to revise their work on the basis of PA ratings. The original and the revised proposals were then compared using an evaluation scale consisting of eight items (see subchapter 3.1.). Similar to Olson (1990), the students' work improved: although the instructors were not informed which were the original and which were the revised proposals, the ratings of the revised proposals were significantly higher than those of the original proposals. Like all studies in this section, this study only measured short-term learning effects (improvement of revised products) and long-term or transfer learning was beyond the scope of these studies.

Van den Berg, Admiraal, and Pilot (2006) found that adequate timing and small group work were beneficial for learning from revisions on the basis of peer feedback. Seven PA designs were developed to identify which characteristics fostered effective PA. The participants were history students ($N = 168$). Of them 131 were allocated to groups using PA, and 37 to groups not using PA. The students exchanged draft reports and those in the PA groups were provided with peer feedback. They revised their drafts into final versions, which were graded by a teacher. No significant differences were found between the grades of groups with PA and without PA. However, for learning outcomes such as processing feedback, PA was facilitated by working in small groups of three to four students. These students were better able to compare feedback from different peers to determine its relevance. Van den Berg et al. (2006) also stated that providing students with sufficient time to revise their work (i.e., sufficient time between the PA and teacher assessment) was preferable.

Lin, Liu, and Yuan (2001) found that in the case of PA and revision, students' thinking style influences the quality of their work when they get the opportunity to revise it on the basis of feedback. A total of 58 computer science students participated in this study. Their thinking style was measured using the Thinking Style Inventory (Lin & Chao, 1999; Sternberg, 1994), which categorised them as either "high executive thinkers" ($N = 30$, i.e., willing to follow instructional rules) or "low executive thinkers" ($N = 28$, i.e., emphasis on independence and creativity). The students then wrote an exploratory reading summary, which was subsequently uploaded onto a web-based PA environment to be assessed by their peers. They were randomly assigned to one of two conditions regarding the type of feedback they would receive: specific feedback or holistic feedback. The specific feedback consisted of six criteria: (a) relevance of the project to the course contents, (b) thoroughness of the assignment, (c) sufficiency of the references, (d) perspective or theoretical clarity, (e) clarity of discussion, and (f) significance of the conclusion. Students in the holistic feedback group were required to give an overall score and general feedback. The results showed that students with a high executive thinking style benefited more from PA in terms of the quality of their written assignment than students with a low executive thinking style. Moreover, feedback format and thinking style were found to have an interaction effect on students' domain-specific skill. Those with a low executive thinking style performed better when receiving specific instead of holistic feedback ($M = 7.44$, $SD = 0.70$ and $M = 5.73$, $SD = 0.92$, respectively). For students with a high executive thinking style there was no noticeable difference between specific and holistic feedback ($M = 7.36$, $SD = 0.80$ and $M = 7.24$, $SD = 0.63$, respectively). This study was the only study in the present review that adopted a quasi-experimental design.

One other study also reported improved domain-specific skill as a result of PA (Tsai, Liu, Lin, & Yuan, 2001), but without ascribing the improvement to certain conditions or methods. Pre-service teachers were required to design a science activity for secondary school students. A networked PA model was used, consisting of the following steps. First, the students discussed their homework assignments with the teacher; then, they uploaded their science activities to the system. Subsequently, they reviewed and commented on the homework of

a peer assigned to them by the system. The teacher then graded each homework assignment and read the peer comments. The system informed the students of their grades and the comments, after which they revised their homework. The process from uploading to revision was repeated once or twice, after which the teacher performed a final assessment. It appeared that after two rounds of PA, both the teacher and peer grades were higher, which suggested that the quality of students' homework assignments had improved.

3.2.1. Summary of findings

Enabling students to revise their work on the basis of peer feedback, small group size, sufficient time for revision, and high executive thinking style, as well as specific feedback format positively influenced domain-specific skill. The study by Lin et al. (2001) also revealed an interaction effect, namely low executive thinkers did better when receiving specific feedback instead of holistic feedback, whereas the feedback specificity did not matter for high executive thinkers. One out of five studies also reported improved domain-specific skill in students, but without ascribing this to certain conditions or PA methods.

3.3. PA skill

Five of the selected studies reported findings on PA skill, four of which showed specific relations between outcomes on the one hand and methods and/or conditions on the other hand. These four studies are described here first, followed by the study that did not investigate specific relations between variables.

Sluijsmans et al. (2004) found that training positively influenced PA skill. They studied 93 students from a teacher training college, 43 of whom received PA training (experimental group). This training consisted of four PA tasks focused on defining assessment criteria. The tasks were integrated into the course "Designing Lesson Plans for Discovery Learning". The students had to define criteria related to skills for this course, such as introducing a problem in a classroom. The remaining students received no training (control group). Pairs of students had to design an elementary school lesson plan, which was assessed by four other pairs. The quality of the PA ratings was determined by the use of criteria, word use, the presence of a consistent structure, and the provision of criticism, of a conclusion, of questions, of marks, and of suggestions for improvement. Overall, the students who had received training conducted better PA ratings than those who had not, that is, the total mean score for students in the experimental group was 16.77 ($SD = 9.65$), whereas for students in the control group was 12.89 ($SD = 6.33$). Specifically, the experimental group scored significantly higher than the control group on "use of criteria" ($M = 13.95$, $SD = 8.31$ and $M = 10.45$, $SD = 5.05$, respectively), on "the presence of a consistent structure" ($M = 0.79$, $SD = 1.42$ and $M = 0.27$, $SD = 0.80$, respectively), and on "provision of marks" ($M = 0.66$, $SD = 1.48$ and $M = 0.19$, $SD = 0.81$, respectively). Only on the aspect "provision of suggestions for improvement" the students in the control group performed significantly better than those in the experimental group ($M = 0.46$, $SD = 0.73$ and $M = 0.11$, $SD = 0.31$ respectively). This study was one of the two studies in the present review that adopted a true experimental design.

Three other studies also revealed relations between student characteristics and PA skill. Lin et al. (2001; see subchapter 3.2) found that students with a high executive thinking style provided peer feedback of better quality than students with a low executive thinking style. In their study, in which students wrote and assessed reading summaries, feedback quality was defined as high when it offered suggestions for the next step of modifying and explaining the peers' reading summary. Feedback scores were assigned by the teacher. These scores indicated that the feedback quality of students with a high executive thinking style was higher ($M = 6.07$, $SD = 2.22$) than that of students with a low executive thinking style ($M = 3.70$, $SD = 1.49$).

Yu, Liu, and Chan (2005) found relations between levels of academic achievement and PA skill. Two classes of primary education students ($N = 52$) participated in this study. They enrolled in the web-based learning system Question Posing and Peer Assessment (QPPA; Yu et al., 2005). On the topics of mathematics, natural sciences and social sciences, they were invited to construct questions, comment upon the questions posed by their peers, and view the peer comments in the web-based learning system. Computation of the product-moment correlation coefficient revealed a significant positive correlation between students' academic achievement and the number of questions generated ($r = .40$). This study distinguishes itself from many other studies in the current review by the fact that both assessing peers and being assessed by peers were taken into account. However, the outcomes were not specifically ascribed to either assessing peers or being assessed by peers.

In a study by Davies (2006), university students participating in a computing course used the computerised PA system Computerised Assessment by Peers. Each student wrote an essay, which was subsequently marked and commented on anonymously by their peers. Davies analysed the peer feedback and produced a "feedback index", that is, a measure of the quality of a piece of assessed work. The variation in the feedback indices revealed that the peers in the lower performance quartiles tended to be less critical, whereas the upper quartiles were more critical.

As opposed to the previous four studies, Mathews (1994) reported findings on PA skill without ascribing them to specific methods or conditions. Moreover, rather than making quantitative claims, Mathews merely described several response styles adopted by groups of students when undertaking PA in higher education. In this study, management students assessed one another's group work contributions. Five response patterns were distinguished: (a) equal equality (all group members were claimed to have contributed equally); (b) normal distribution (despite ups and downs, there was an overall feeling that things evened out); (c) reluctant finger (group members consistently indicated under-contributions

from one person who had free-wheeled[2]); (d) stitch them up (collusion between some group members, which resulted in evaluations that clearly indicated poor performance by one or more members); and (e) out of kilter (perceptions about contributions greatly varied between group members).

3.3.1. Summary of findings

It was found that PA skill improved by using a method that incorporated training. Several conditions also played a role: students with high executive thinking styles were generally better at assessing peers than low executive thinkers and students with a high level of academic achievement were generally more skilful in PA and critical than low achievers.

3.4. Student attitudes towards PA

The most often reported outcome category in the selected studies was student attitudes towards PA. A total of 15 studies reported findings on attitudes. This subchapter first describes the six sources that revealed relations between specific methods and/or conditions and student attitudes. It then discusses the remaining nine sources according to their outcomes: studies reporting positive effects, studies reporting mixed results, and finally studies reporting negative effects on attitudes.

In Smith et al. (2002) study (see subchapter 3.1) student attitudes towards PA were measured as well. For both psychology cohorts, course feedback questionnaires were used consisting of 10 items to be rated on a 5-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree); example item was "The assessment was appropriate". Students could also write qualitative comments. Students of the second cohort, which had received the intervention, also received open-ended qualitative questionnaires at three points in time: (a) at the end of the course introduction, (b) at the end of the second workshop, and (c) after the course had been completed. Before the intervention, they wrote extremely negative comments in the course feedback questionnaires. The largest category of criticism involved PA, that is, they were not confident that students could mark other students' work. After the intervention, the qualitative questionnaires revealed more positive attitudes. At time point 1, the students still had some concerns about PA but also indicated that it could help to bridge the gap between learners and assessors. At time point 2, there was an increase of confidence in the PA process. At time point 3, the students perceived ownership of marking and expressed more awareness of the goals of the posters. Moreover, comparing questionnaires of both cohorts revealed not only more positive student attitudes towards PA overall, but also more positive attitudes towards the entire course.

Cheng and Warren (1997) found PA training and experience to have similar positive effects on student attitudes. As part of a group project, 52 first-year electrical engineering undergraduates had to give presentations and write reports for an English language course. They participated in a training programme in which they discussed the advantages and disadvantages of PA, examined the assessment criteria developed by staff, and practiced using PA. Subsequently, they assessed the group presentations and reports by other groups, as well as the contributions by their own group members to the preparation and execution of their project. Student attitudes were measured by means of questionnaires consisting of four open-ended questions (e.g., "Do/did you feel comfortable doing peer assessments?") before and after the training and PA task. Almost two thirds of the students indicated both before and after the training and experience with the PA task that students should participate in PA, but fewer than half thought that first-year students were able to assess their peers fairly and responsibly. Attitude shifts occurred in both positive and negative directions in relation to training and experience with PA, but overall there was a positive attitude shift. After training and experience with PA, the majority of students felt more comfortable with PA and had more confidence in PA.

The positive effects of experience on student attitudes were underlined by Sluijsmans et al. (2004; see subchapter 3.3). Students completed a perception of instruction and assessment questionnaire before and after the PA intervention. This questionnaire consisted of 92 items, categorised in 15 clusters on instruction and assessment as well as on the role of students in assessment. The answers were rated on a 5-point Likert-type scale ranging from 1 (totally disagree) to 5 (totally agree). Although training did not affect student perceptions for 12 of the 15 clusters, these perceptions were significantly more positive after the PA than before. The significant positive shift in perceptions on the cluster 'assessment skill' was particularly important, because this cluster was the focus of the training.

Venables and Summit (2003) also provided empirical support for the positive influence of PA experience on student attitudes. In their study, 125 computer science undergraduates without prior PA experience anonymously assessed their peers' literature review essays. They received detailed instructions and assessment criteria for the task, and were informed that the PAs of their essays were recommendations. If necessary, tutors altered skewed PA ratings (i.e., far too positive/negative assessments), but this was restricted to a few cases. Surveys with responses on a 5-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree) were administered, in which students were asked to give their opinions on the essay being assessed in terms of task description, intellectual level, length, and its contribution to understanding of the subject. The results showed that the students had reservations prior to the PA task, and little confidence in assessing their peers. After the task, however, the opinions of most students were more positive. They indicated that they had learned a lot and considered their peers' perspectives beneficial.

Wen and Tsai (2006) found further support for the notion that PA experience contributes to more positive student

[2] A "free-wheeler", in this context, is someone who contributes very little to the group work. Often, a free-wheeler 'parasites' on other group members' work.

attitudes. In their sample, 280 university students participated of whom 59% had PA experience. Student attitudes towards online PA were measured by means of a 34-item questionnaire with answers rated on a 5-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree). The questionnaire consisted of four subscales, including a Positive Attitude subscale (e.g., ''PA is helpful to my learning''), an Online Attitude subscale (e.g., ''Online PA activities can be time saving''), an Understanding-And-Action subscale (e.g., ''PA activities help me understand what my classmates think'') and a Negative Attitude subscale (e.g., 'I think students should not be responsible for assessments'). Students with previous PA experience had less negative attitudes towards PA than those without such experience. Additionally, male students had more positive attitudes towards PA than females, a finding also reported in Wen, Tsai, and Chang (2006).

Although clear relations between methods and/or conditions and outcomes were described in these six studies, the other nine studies were less transparent regarding these relations. Freeman and McKenzie (2002), for example, reported four case studies involving the web-based template 'Self and Peer Assessment Resource Kit' (SPARK; Freeman & McKenzie, 2002). The PA training was an integral part of this template and generally positive student attitudes (measured by student focus groups, ratings and open-ended survey questions) were reported.

Haaga's (1993) study (see subchapter 3.1) reported a peer review exercise of draft papers in psychology courses. After the peer review process, students were asked to provide anonymous ratings of the educational value of PA on a Likert-type scale ranging from 0 (worthless) to 10 (highly educational). They were found to have positive attitudes towards PA.

Liu and Tsai (2005) reported on web-based PA systems using repertory grid analysis. Computer science students submitted portfolios and were assigned to groups of six to eight students. Within these groups they assessed the portfolios of their group members using dynamic, student-designed assessment criteria (e.g., data structure, searching methodology, and convenience of usage) on 5-point Likert-type scales; for example, responses for the criterion convenience of usage ranged from 1 (good user interface) to 5 (poor user interface). Students' perceptions of this particular PA process were then measured by a questionnaire also containing eight items on a 5-point Likert-type scale, such as ''Feedback from others reflects the weak points of my portfolio''. Similar to Haaga's (1993) findings, students were generally positive about the PA process: they felt that it had helped their learning and provided substantial feedback.

Stanier (1997) also found positive student attitudes towards PA. In this study, 36 environmental sciences students were involved in group work. One of their tasks was to produce a brochure for assessment by their peers and tutors. Most of the students (98%) were unfamiliar with PA. They were involved in the design of the assessment criteria and received clear instructions on how to conduct PA followed by discussions. Their evaluations of this PA experience revealed positive attitudes, that is, 94% claimed that they had learned from the experience and 74% felt that PA should be part of other courses as well.

Smith (1990; see subchapter 3.1) provided further evidence of positive student attitudes towards PA. Psychology students ($N = 42$) were asked to rate their views on the evaluative aspects of a debate on a scale ranging from 1 (poor) to 10 (great). The aspect ''Evaluating other students' debates'' was rated positively ($M = 6.98$, $SD = 1.78$), as were ''Being evaluated by fellow students'' ($M = 6.62$, $SD = 2.08$) and ''How much did evaluating other students' debates help you prepare for your own debate?'' ($M = 6.67$, $SD = 2.27$).

Strachan and Wilcox (1996) also found positive student attitudes towards PA. Geography students ($N = 30$) worked in ten groups of three members to develop a research topic, present their research to their peers and write a paper. They were trained in the assessment procedures and involved in devising the assessment criteria. They ranked their fellow group members using the zero-sum technique (i.e., a student's mark in the group may go up and down, as long as the sum of the movements for all group members is zero). On a simple evaluation form they could write comments about the assessment procedure. Analysis of these evaluation forms revealed that most students perceived the assessment as positive, although some expressed reservations (e.g., they disliked rating their friends).

Pain and Mowl (1996) found the effects on student attitudes to be less unidirectional. In their study, 53 geography students received assessment training and then wrote essays for peer and self-assessment. Of these students, 67% indicated that PA was difficult, 45% said it helped their understanding of assessment, 18% found peer and self-assessment to be less fair than staff assessment, and 64% felt that the assessment procedure would help them write better essays in the future.

Brindley and Scoffield (1998) also found mixed results on student attitudes. A total of 80 business students, equally distributed over two cohorts, received marking schemes to assess other groups' performances in either role plays/presentations or the organisation and running of a trade fair. The number of students with and without PA experience was about equal in both cohorts. Questionnaires revealed that the majority of the students understood the process, but also had difficulty assigning marks to peers. Half of the students preferred to be involved in devising assessment criteria, whereas half did not.

In an anonymous PA setting of a clinical clerkship (Levine et al., 2007; see subchapter 3.1), students evaluated their teammates' performances. When asked to provide rationales for why they had given certain marks, students often chose to comment on the PA process instead. They expressed negative attitudes towards PA and indicated being hesitant about giving negative feedback to other students. A total of 72 unsolicited statements (statements not providing a rationale for the peer performance evaluation) were given. Using qualitative analysis, 56 were coded as 'negative comments about the PA process' and the remaining 16 comments expressed the general opinion that 'everyone in our team did well'.

3.4.1. Summary of findings

Twelve studies revealed positive student attitudes overall, and six of these studies reported positive relations between methods and/or conditions (training, experience) and student attitudes. The remaining six studies did not ascribe attitudes to particular methods or conditions. Two studies reported mixed findings on attitudes and one study reported negative student attitudes towards PA. These three studies also did not ascribe attitudes to particular methods or conditions. It is notable that, whereas the procedures varied tremendously, there was also an enormous variety in the instruments used to measure student attitudes.

4. Discussion

The present review aimed to identify relations between variables that foster effective PA. Generally, the psychometric qualities of PA seem to be sufficient and PA was found to have positive effect on domain-specific skill, PA skill, and student attitudes towards PA. Table 1 provides a summary of all findings. For each study, it shows the experimental nature, the educational context, the reported outcome variable(s), whether particular methods and/or conditions were related to the outcome variable(s), and whether the effects on the outcome variable were positive, moderate or negative.

Table 1
Summary of reported findings per study.

Source	Nature of study	Context	Outcome variables				Related condition/Method variables
			Psychometric qualities	Domain-specific skill	PA skill	Student attitudes towards PA	
Brindley and Scoffield (1998)	PEX	HE				+_	
Cheng and Warren (1997)	PEX	HE				[+]	Method: training & condition: experience
Davies (2006)	PEX	HE			[+]		Condition: high academic achievement
Freeman and McKenzie (2002)	PEX	HE				+	
Haaga (1993)	PEX	HE	+_			+	
Hughes and Large (1993)	PEX	HE	+				
Levine et al. (2007)	PEX	HE	+/_			_	
Lin et al. (2001)	QEX	HE		[+]	[+]		Method: specific feedback format (> domain-specific skill) Condition: high executive thinking (> domain-specific skill & PA skill)
Liu and Tsai (2005)	PEX	HE				+	
Magin (2001)	PEX	HE	+				
Mathews (1994)	PEX	HE			+_		
Olson (1990)	TEX	PE		[+]			Method: PA-based revision
Pain and Mowl (1996)	PEX	HE				+_	
Sluijsmans et al. (2004)	TEX	HE			[+]	[+]	Method: training (> PA skill) Condition: experience (> attitudes)
Smith (1990)	PEX	HE	+			+	
Smith et al. (2002)	PEX	HE	[+]			[+]	Method: training (> psychometrics & attitudes)
Stanier (1997)	PEX	HE				+	
Stefani (1994)	PEX	HE	+/+_				
Strachan and Wilcox (1996)	PEX	HE				+	
Sung et al. (2003)	PEX	HE	[+]	[+]			Condition: experience (> psychometrics) Method: PA-based revision (> domain-specific skill)
Tsai et al. (2001)	PEX	HE		+			
Van den Berg et al. (2006)	PEX	HE		[+]			Methods: small groups & revision time
Venables and Summit (2003)	PEX	HE				[+]	Condition: experience
Wen and Tsai (2006)	PEX	HE				[+]	Conditions: experience & gender (male)
Wen et al. (2006)	PEX	HE				[+]	Condition: gender (male)
Yu et al. (2005)	PEX	PE			[+]		Condition: high academic achievement

HE = higher education; PE = primary education; PEX = pre-experimental; QEX = quasi-experimental; TEX = true experimental. +/+_/_ = positive/moderate/negative outcome variable not related to method and/or condition variable. [+] = positive outcome variable related to method and/or condition variable.

With regard to variables that foster PA, our review revealed that the psychometric qualities of PA scores could be enhanced by training and experience. Enabling students to revise their work on the basis of peer feedback improved domain-specific skill. The PA skill appeared to be mainly fostered by training and dependent on student characteristics, such as thinking style and level of academic achievement. Finally, student attitudes towards PA were positively influenced by training and experience.

4.1. Gaps in content: topics that need further scrutiny

The eight studies that reported on the psychometric qualities of PA do not yet provide conclusive evidence as to what contributes to these qualities. As these qualities are especially relevant for teachers who want to implement PA, methods to improve psychometric qualities must be investigated more specifically in future studies. Also, in the studies selected for the current review, domain-specific skill was mostly measured by performance on the product to be peer assessed. Long-term learning and transfer of learning, measured by retention and transfer tasks, were not investigated. The slightly diverse findings on student attitudes reported in the selected studies might be clarified by future investigations on the role of interpersonal variables in PA, such as psychological safety and trust (Van Gennip, Segers, & Tillema, 2010). Also, it was striking that almost none of the studies clearly differentiated between the effects of assessing peers versus the effects of being assessed by peers. To truly account for the effects of PA, this distinction and the differential effects on different types of outcomes need further scrutiny.

Most of the selected studies reported research conducted in higher education. It would also be interesting to investigate learning effects of PA in secondary education (Gielen, Peeters, Dochy, Onghena, & Struyven, 2010) and vocational education (e.g., Van Gennip et al., 2010).

4.2. Gaps in methodology: research methods that need transparency

When interpreting the results, the methodologies of the selected studies should be taken into account. The holistic method of describing PA (see Section 1) was underlined by our review. Only 12 studies reported clear relations between methods, conditions and outcomes. Interventions were often described globally and outcomes were discussed without being ascribed to particular causes. Although these studies provided useful insights regarding PA practices, inferences about what caused the reported effects were difficult or impossible to draw.

The multiplicity of PA practices was underlined as well. One of the few common denominators was the fact that nearly all studies reported on PA in higher education. Apart from this commonality, a large diversity of PA processes was described. The instruments used to measure student attitudes were also very diverse, which makes it hard to compare studies.

The aforementioned remarks on holistic research reports and study comparisons gain extra weight when publication bias is taken into account. Studies reporting positive effects of a particular intervention are published more often than studies reporting no or negative effects, hence relations between variables may thus not occur systematically in the literature and the overall picture may be too optimistic.

The share of quasi-experimental and true experimental studies in the field of PA is very small (see Table 1). This means that the research outcomes of the reviewed studies must be interpreted with some caution. Experimental research in which methods and conditions are clearly described and related to outcome variables in factorial designs will contribute to more clarity on how to design optimal PA. Several other contributions in this special issue provide good examples (Cho & MacArthur, 2010; Gielen et al., 2010; Strijbos, Narciss, & Dünnebier, 2010; Van Gennip et al., 2010; Van Steendam, Rijlaarsdam, Sercu, & Van den Bergh, 2010).

4.3. Suggestions for further research

The present study revealed several gaps in existing research on PA in education that provide starting points for further experimental research. To provide full insight into effective PA processes, issues regarding content as well as research methodologies require more attention. Content-related topics that need further investigation include the psychometric qualities of PA, long-term effects and transfer of learning, and the effects of assessing a peer versus being assessed by a peer. Also, educational contexts other than higher education should be the focus of research. As for methodologies, more true experimental and quasi-experimental research is needed that describes relevant variables in a specific rather than a holistic fashion, with more consistency in measurement instruments.

Acknowledgements

This research project is funded by the Netherlands Organization for Scientific Research (NWO; The Hague, Project No. 411-05-110).

References

Boud, D., Cohen, R., & Sampson, J. (1999). Peer learning and assessment. *Assessment and Evaluation in Higher Education, 24*, 413–426.

Brindley, C., & Scoffield, S. (1998). Peer assessment in undergraduate programmes. *Teaching in Higher Education, 3*, 79–89.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally College.

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

Cheng, W., & Warren, M. (2000). Making a difference: using peers to assess individual students' contributions to a group project. *Teaching in Higher Education, 5*, 243–255.

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

Davies, P. (2006). Peer assessment: judging the quality of students' work by comments rather than marks. *Innovations in Education and Teaching International, 43*, 69–82.

Freeman, M., & McKenzie, J. (2002). SPARK, a confidential web-based template for self and peer assessment of student teamwork: benefits of

evaluating across different subjects. *British Journal of Educational Technology, 33*, 551–569.

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.

Haaga, D. A. F. (1993). Peer review of term papers in graduate psychology courses. *Teaching of Psychology, 20*, 28–32.

Hughes, I. E., & Large, B. J. (1993). Staff and peer-group assessment of oral communication skills. *Studies in Higher Education, 18*, 379–385.

Levine, R. E., Kelly, P. A., Karakoc, T., & Haidet, P. (2007). Peer evaluation in a clinical clerkship: students' attitudes, experiences, and correlations with traditional assessments. *Academic Psychiatry, 31*, 19–24.

Lin, S .S. J., & Chao, I. C. (1999). The manual for use with thinking style Inventory-Taiwan version Hsinchu, Taiwan: Institute of Education, National Chiao Tung University. Unpublished manual.

Lin, S. S. J., Liu, E. Z. F., & Yuan, S. M. (2001). Web-based peer assessment: feedback for students with various thinking-styles. *Journal of Computer Assisted Learning, 17*, 420–432.

Liu, C. C., & Tsai, C. M. (2005). Peer assessment through web-based knowledge acquisition: tools to support conceptual awareness. *Innovations in Education and Teaching International, 42*, 43–59.

Magin, D. (2001). Reciprocity as a source of bias in multiple peer assessment of group work. *Studies in Higher Education, 26*, 53–63.

Mathews, B. P. (1994). Assessing individual contributions: experience of peer evaluation in major group projects. *British Journal of Educational Technology, 25*, 19–28.

Olson, V. L. B. (1990). The revising process of sixth-grade writers with and without peer feedback. *Journal of Educational Research, 84*, 22–29.

Pain, R., & Mowl, G. (1996). Improving geography essay writing using innovative assessment. *Journal of Geography in Higher Education, 20*, 19–32.

Reigeluth, C. M. (Ed.). (1983). *Instructional design theories and models: An overview of their current status*. Hillsdale, NJ: Erlbaum.

Sluijsmans, D. M. A., Brand-Gruwel, S., Van Merriënboer, J. J. G., & Martens, R. L. (2004). Training teachers in peer-assessment skills: effects on performance and perceptions. *Innovations in Education and Teaching International, 41*, 60–78.

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.

Smith, R. A. (1990). Are peer ratings of student debates valid? *Teaching of Psychology, 17*, 188–189.

Stanier, L. (1997). Peer assessment and group work as vehicles for student empowerment: a module evaluation. *Journal of Geography in Higher Education, 21*, 95–98.

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

Sternberg, R. J. (1994). Thinking styles: theory and assessment at the interface between intelligence and personality. In R. J. Sternberg, & P. Ruzgis (Eds.), *Personality and intelligence* (pp. 169–187). New York: Cambridge University Press.

Strachan, I. B., & Wilcox, S. (1996). Peer and self assessment of group work: developing an effective response to increased enrolment in a third-year course in microclimatology. *Journal of Geography in Higher Education, 20*, 343–353.

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.

Sung, Y. T., Lin, C. S., Lee, C. L., & Chang, K. E. (2003). Evaluating proposals for experiments: an application of web-based self-assessment and peer-assessment. *Teaching of Psychology, 30*, 331–334.

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

Tsai, C. C., Liu, E. Z. F., Lin, S. S. J., & Yuan, S. M. (2001). A networked peer assessment system based on a vee heuristic. *Innovations in Education and Teaching International, 38*, 220–230.

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. M., & 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 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 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 higher-order peer feedback in EFL. *Learning and Instruction, 20*(4), 316–327.

Venables, A., & Summit, R. (2003). Enhancing scientific essay writing using peer assessment. *Innovations in Education and Teaching International, 40*, 281–290.

Wen, M. L., & Tsai, C. C. (2006). University students' perceptions of and attitudes toward (online) peer assessment. *Higher Education, 51*, 27–44.

Wen, M. L., Tsai, C. C., & Chang, C. Y. (2006). Attitudes toward peer assessment: a comparison of the perspectives of pre-service and in-service teachers. *Innovations in Education and Teaching International, 43*, 83–92.

Yu, F. Y., Liu, Y. H., & Chan, T. W. (2005). A web-based learning system for question-posing and peer assessment. *Innovations in Education and Teaching International, 42*, 337–348.