ELSEVIER

Learning and Instruction 20 (2010) 304–315

Learning and Instruction

www.elsevier.com/locate/learninstruc

Improving the effectiveness of peer feedback for learning

Sarah Gielen [a,b,*], Elien Peeters [a], Filip Dochy [a], Patrick Onghena [c], Katrien Struyven [a]

[a] *Katholieke Universiteit Leuven, Department of Educational Sciences, Centre for Research on Teaching and Training, Dekenstraat 2 Box 3772, BE-3000 Leuven, Belgium*

[b] *Katholieke Universiteit Leuven, Department of Educational Sciences, Centre for Educational Effectiveness and Evaluation, Dekenstraat 2 Box 3773, BE-3000 Leuven, Belgium*

[c] *Katholieke Universiteit Leuven, Department of Educational Sciences, Centre for Methodology of Educational Research, Vesaliusstraat 2 Box 3700, BE-3000 Leuven, Belgium*

Abstract

The present study examined the effectiveness of (a) peer feedback for learning, more specifically of certain characteristics of the content and style of the provided feedback, and (b) a particular instructional intervention to support the use of the feedback. A quasi-experimental repeated measures design was adopted. Writing assignments of 43 students of Grade 7 in secondary education showed that receiving 'justified' comments in feedback improves performance, but this effect diminishes for students with better pretest performance. Justification was superior to the accuracy of comments. The instructional intervention of asking assessees to reflect upon feedback after peer assessment did not increase learning gains significantly.

Keywords: Peer assessment; Peer feedback; Writing; Revision; Feedback accuracy

1. Introduction

In the last two decades there has been strong interest in ''formative assessment'', that is, assessment designed to provide rich feedback and support for learning (Black & Wiliam, 1998), and renewed interest in peer assessment as a tool for learning (Falchikov, 1995). However, to increase the potential impact of peer assessment on learning, it is crucial to understand which mechanisms affect learning, and how these mechanisms can be supported.

During formative peer assessment, judgements often include qualitative comments in addition to (or instead of) marks. These comments are labelled ''peer feedback''. Peer feedback is expected to support the learning process by providing an intermediate check of the performance against the criteria, accompanied by feedback on strengths, weaknesses and/or tips for improvement (Falchikov, 1996). There can also be learning benefits for the peer assessor, arising from seeing other examples or approaches, and from internalisation of criteria and standards (Topping, 1998).

Not all feedback leads to performance improvement (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Kluger & DeNisi, 1996). Gibbs and Simpson (2004) describe several conditions under which feedback has a positive influence on learning. Feedback should be (a) sufficient in frequency and detail; (b) focused on students' performance, on their learning, and on the actions under students' control, rather than on the students themselves and/or on personal characteristics; (c) timely in that it is received by students while it still matters and in time for application or for asking further assistance; (d) appropriate to the aim of the assignment and its criteria; (e) appropriate in relation to students' conception of learning, of knowledge, and of the discourse of the discipline; (f) attended to, and (g) acted upon.

To date, literature that empirically links quality criteria for feedback to performance improvement in the case of peer

* Corresponding author. Tel.: +32 16 326181; fax: +32 16 325790.
E-mail address: sarah.gielen@ped.kuleuven.be (S. Gielen).

0959-4752/$ - see front matter © 2009 Published by Elsevier Ltd.
doi:10.1016/j.learninstruc.2009.08.007

assessment is scarce (Kim, 2005) and few studies (Sluijsmans, Brand-Gruwel, & Van Merriënboer, 2002) adopt a quasi-experimental approach to study the impact of instructional interventions on peer feedback effectiveness and learning (Van Zundert, Sluijsmans, & Van Merriënboer, 2010). The present study addressed the question whether general quality criteria for feedback are applicable to peer feedback and whether the effectiveness of peer feedback for learning can be raised through instructional interventions that aim to meet the conditions described by Gibbs and Simpson (2004).

Before exploring the specific literature on peer feedback, we first present an overview of feedback types and their effectiveness. Subsequently, we discuss the particularities of peer feedback compared to teacher feedback, as well as the effectiveness of peer feedback. Next, we provide an overview of existing operationalisations of peer feedback quality. Finally, we elaborate on instructional interventions aimed at stimulating effective use of peer feedback.

1.1. Feedback and performance

Narciss (2008) defines feedback as "all post-response information that is provided to a learner to inform the learner on his or her actual state of learning or performance" (p. 127) and differentiates between external (e.g., peer or teacher) and internal (the learner) sources of feedback. Feedback can have a strong positive effect on learning under certain conditions (see Bangert-Drowns et al., 1991; Kluger & DeNisi, 1996), however, effects can be absent or even negative depending on the instructional conditions.

Feedback research not only addresses *whether* feedback improves learning, but also *how* feedback improves learning. Mory (2003) discusses four perspectives on how feedback supports learning. First, feedback can be considered as an incentive for increasing response rate and/or accuracy. Second, feedback can be regarded as a reinforcer that automatically connects responses to prior stimuli (focused on correct responses). Third, feedback can be considered as information that learners can use to validate or change a previous response (focused on erroneous responses). Finally, feedback can be regarded as the provision of scaffolds to help students construct internal schemata and analyse their learning processes.

Apart from these perspectives on how feedback supports learning, the type of feedback varies considerably as well. Sometimes feedback is mere "knowledge of performance" (e.g., percentage of correctly solved tasks), "knowledge of result" (e.g., correct/incorrect) or "knowledge of correct response" (e.g., the correct answer to the given task), whereas in other cases it includes elaborated information strategically useful for task completion (e.g., "Do this/Add that/Avoid this", without giving the answer) or explanations for error correction (e.g., "Your response is incorrect, because…") (Narciss, 2008). Feedback messages differ in the volume of the elaborated informational component, and this appears to be related to their effectiveness in altering performance (Narciss & Huth, 2006). An example is the learning effect of the elaboration of feedback and presence of explanations in help during collaborative learning (Webb, 1991). Regardless of differences in the extent, content and style of feedback (i.e., amount and type of information), and the learning processes that are expected to take place (i.e., perspectives on how feedback supports learning), the information in traditional feedback research, however, can always be considered accurate.

1.2. Peer feedback and performance

Peer feedback is provided by equal status learners and can be regarded both as a form of formative assessment – the counterpart of teacher feedback – (Topping, 1998), and as a form of collaborative learning (Van Gennip, Segers, & Tillema, 2010; Webb, 1991). Taking the perspective of formative assessment, the main difference between teacher and peer feedback is that peers are not domain experts, as opposed to teachers. As a consequence the accuracy of peer feedback varies. Peer judgements or advice may be partially correct, fully incorrect or misleading. Moreover, the peer assessor is usually not regarded as a "knowledge authority" by an assessee, leading to more reticence in accepting a peer's judgement or advice (Hanrahan & Isaacs, 2001; Strijbos, Narciss, & Dünnebier, 2010).

Nevertheless, peer feedback can be beneficial for learning, which might even be due to the difference from teacher feedback (Topping, 1998), since the absence of a clear "knowledge authority" (e.g., the teacher) alters the meaning and impact of feedback. Bangert-Drowns et al. (1991) argue that "mindful reception" is crucial for the instructional benefits of feedback, and this might be stimulated through the uncertainty induced by a peer's relative status. In the study by Yang, Badger, and Yu (2006) revision initiated by teacher feedback was less successful than revision initiated by peer feedback, probably because peer feedback induced uncertainty. Teacher feedback was accepted as such, but proved to be associated with misinterpretation and miscommunication, whereas reservations regarding the accuracy of peer feedback induced discussion about the interpretation. Students' reservations prompted them to search for confirmation by checking instruction manuals, asking the teacher, and/or performing more self-corrections. As a result, students acquired a deeper understanding of the subject. In contrast, teacher feedback lowered students' self-corrections, perhaps students assumed that the teacher had addressed all errors and that no further corrections were required (Yang et al., 2006).

In addition to stimulating the "mindful reception", peer feedback may also increase the frequency, extent and speed of feedback for students while keeping workload for teachers under control. Involving students in the assessment process increases the number of assessors and feedback opportunities. Although the accuracy might be lower compared to teacher feedback, this can be considered an acceptable trade-off for increased follow-up of students' progress (Gibbs & Simpson, 2004).

1.2.1. Peer feedback in the domain of writing

Peer feedback as part of first language composition classes (L1 writing) has yielded beneficial effects. In this domain peer feedback is often referred to as "peer review" or "peer assistance when writing". In a recent meta-analysis Graham and Perin (2007) report a large positive effect size for peer feedback during writing instruction (Grade 4 through high school) when compared to students writing individually. Some studies have found peer comments to be as effective as teacher comments (Cho & Schunn, 2007, in the case of single peer feedback; Gielen, Tops, Dochy, Onghena, & Smeets, 2010), or even enhance performance beyond teacher feedback (Cho & Schunn, 2007, in the case of multiple peer feedback; Karegianes, Pascarella, & Pflaum, 1980).

Nevertheless, peer feedback is not always as effective as teacher feedback. Tsui and Ng (2000) found that teacher feedback was more often incorporated in revisions than peer comments when students received both peer and teacher comments. Students also perceived teacher comments as more useful, but the impact of comments on the quality of final assignments was not examined. The study outlines some of the problems associated with peer comments, such as depth of the feedback, accuracy and credibility. However, these appear to be more present in second language classes than first language classes (Nelson & Murphy, 1993).

A series of studies by Cho et al. (Cho, Chung, King, & Schunn, 2008; Cho & MacArthur, 2010; Cho, Schunn, & Charney, 2006) revealed qualitative differences in peer and expert feedback. Experts provide more ideas and longer explanations and typically include less praise, whereas peers' comments request more clarification and elaboration. Yet, students (as opposed to other experts) perceive peer and expert feedback as equally helpful, given a blinded source. In addition, there were qualitative differences in the type of initiated revisions. Feedback by multiple peers induced more complex repairs (compared to teacher or single peer feedback) and extension of content (compared to single peer feedback). Expert feedback induced more simple repairs than single peers, but these had no effect on writing quality after revision. Complex repairs improved writing quality, whereas adding new content had a negative influence.

1.3. Peer feedback quality

There are various perspectives on peer feedback quality. A first perspective defines peer feedback quality in terms of accuracy, consistency across assessors and/or concordance with teacher feedback (see Van Steendam, Rijlaarsdam, Sercu, & Van den Bergh, 2010). Examples of quality criteria for peer feedback from this perspective are (a) the number of errors detected from the total number of errors; (b) the number of errors accurately and completely corrected and justified out of the total number of errors, and (c) a holistic score for the correctness, exhaustiveness and explicitness of peer comments. This definition originates from a summative view on peer assessment, where scoring validity and reliability are leading concepts. However, from an interventional point of view it is problematic, because – even if more accurate feedback is assumed to be better than inaccurate feedback – peers are not experts. Peer assessors are inevitably novices, unless peer assessment is transformed into cross-level peer tutoring.

A second perspective defines peer feedback quality in terms of content and/or style characteristics. The advantage of this approach is that such characteristics are not domain- and/or task-specific, thus teaching students to focus on content and style characteristics results in a generic skill transferable to other settings. Examples of this perspective are the studies of Kim (2005), Sluijsmans et al. (2002), and Prins, Sluijsmans, and Kirschner (2006) (see Table 1).

Kim (2005) studied the relationship between feedback composition (in terms of four characteristics listed in the first column of Table 1) and performance. She considered feedback as constructive when marks and comments for each content criterion were present and supported by a rationale and revision suggestion. However, no performance increase was observed for assessees who received this type of high quality peer feedback. Kim (2005) argues that this might have been due to the limited variance in peer feedback quality ($SD = 0.53$) and/or assessees' scepticism toward their peers' ability as assessors, preventing students from internalising peer feedback. This lack of internalisation might have decreased the impact of constructive feedback on performance. Hence, peer feedback quality might still be important for performance, provided that assessees are stimulated to apply the feedback.

It should be noted that the scepticism toward peers' ability is precisely the argument used by Yang et al. (2006) to explain a *higher* impact of peer feedback, as compared to teacher feedback. The arguments by Kim (2005) and Yang et al. (2006) appear to be contradictory; however, assessees' reservations in Yang et al.'s (2006) study prompted them to initiate discussion and self-correction, which in turn led to successful revisions. Performance improvement was not directly related to feedback composition (and thus the quality of peer feedback) – as is the focus in the study by Kim (2005) – but rather to the critical attitude of the assessee toward peer feedback.

Another example of the content and style perspective is the study by Sluijsmans et al. (2002). They extracted characteristics of constructive peer feedback from expert assessment reports and identified seven key characteristics and associated criteria (see second column of Table 1). Contrary to Kim (2005), Sluijsmans et al. (2002) adopted an interventional perspective and examined how students can be instructed to apply the key characteristics more frequently in their feedback. However, the relationship between the feedback characteristics and effectiveness of the peer feedback was not investigated.

A third example is the study by Prins et al. (2006). They compared the style and quality of peer feedback by general practitioners in training to that by expert trainers, but they did not relate style or quality to performance. Prins et al. (2006) developed the Feedback Quality Index, an elaboration of the form used by Sluijsmans et al. (2002). However, instead of counting the number of criteria used and the number of

Table 1
Summary and comparison of criteria used for "good" peer feedback.

Kim (2005)	Sluijsmans et al. (2002)	Prins et al. (2006)	The present study
Criterion-orientation	Use of criteria (1 point per criterion)	Presence of content-related remarks (weight 3)	- Comments related to the assessment criteria (Appropriateness) - Explanation of judgement 1: Reference to specific behaviour (Specificity)
Justification	–	Presence of explanations of remarks (weight 2)	Explanation of judgement 2: Justification
Suggestion	Constructive suggestions (1 point per comment)	Presence of good and clear suggestions for improvement/advice (weight 1)	Presence of suggestions for improvement
–	- Positive comments (1 point per comment) - Negative comments (1 point per comment)	Balance of positive and negative remarks (weight 1)	* Presence of both positive and negative comments (unless no negative possible)
–	Posing questions (1 point per question)	Presence of questions fostering reflection (weight 1)	* Presence of thought-provoking questions
–	–	Clear formulation (descriptions instead of keywords) (weight 0.5)	Clear formulation
–	Structure (max. 4 points for presence of clear judgement, summary of suggestions for improvement, positive comment at beginning or end, and length of conclusion)	Clear structure in report (weight 0.5)	(Not applicable in pre-structured feedback form)
Completeness	–	–	–
	No use of 'naïve words' (minus 1 point per word, such as nice, good, excellent, fine)	–	–
–	–	Presence of external examples (weight 0.5)	–
–	–	Style (first person instead of judging) (weight 0.5)	–

Criteria with an asterisk were omitted for analyses.

comments or certain words present, the index evaluates the presence of a set of necessary ingredients – with a certain weight applied (third column in Table 1). Although the Feedback Quality Index is derived from expert feedback reports and grounded in learning theories, the contribution of the identified feedback characteristics on performance was not empirically tested.

1.4. Instructional interventions to foster peer feedback effectiveness

Studies on instructional interventions to enhance the effectiveness of peer feedback have focused thus far on the impact of (a) indicators to clarify evaluation criteria (Orsmond, Merry, & Reiling, 2002), (b) the number of peer assessors by comparing a single peer assessor versus multiple peer assessors (Cho & Schunn, 2007), (c) the training of peer assessors in assessment skills (Sluijsmans et al., 2002), (d) methods of teaching students how to provide peer feedback (Van Steendam et al., 2010), (e) the matching principles for peers (Van den Berg, Admiraal, & Pilot, 2006), and (f) teacher support for peer feedback (Van den Berg et al., 2006). All studies have in common that the focus mainly lies on the assessor.

Instructional interventions to raise peer feedback quality include, for example, the use of directed questions (such as "Did the assessee cover all relevant topics?") to stimulate comments on assessment criteria (Miller, 2003) and sentence openers (such as "I propose to…/I think that…") to promote task-focused and reflective interaction between learners (Baker & Lund, 1997). Yet, these types of interventions might have negative motivational effects in the long run, because they can interrupt the natural interaction process by enforcing the use of the same communication structures on all occasions. Another type of intervention used to raise the quality of peer feedback is training students to adopt specific quality criteria. A third type of intervention is the use of a quality control system that rewards or sanctions assessors for the quality of their feedback (Bloxham & West, 2004; Kali & Ronen, 2008; Searby & Ewers, 1997) or that filters unreliable feedback (Cho & MacArthur, 2010; Rada & Hu, 2002). In line with Gibbs and Simpson (2004), however, it is not sufficient to focus on the type or quality of peer feedback to foster its effectiveness, but the assessee's response should be addressed as well. A fourth type of intervention aiming both at raising feedback quality and the response to it, is the adoption of an 'a priori question form' (assessees formulate their feedback needs; see, e.g., Gielen et al., 2010) combined with a feedback form prompting

the assessor to address these needs in the feedback. Such an intervention may enhance both ''individual accountability'' and ''positive interdependence'' (Slavin, 1989), and motivate and guide assessors to provide ''responsive'' feedback (Webb, 1991). It may also result in more appropriate feedback (Webb, 1991) and promote ''mindful reception'' (Bangert-Drowns et al., 1991), that is, make assessees feel more personally addressed and subsequently more inclined to apply the feedback. Finally, a fifth type of intervention to foster the use of feedback is an ''a posteriori reply form'' (assessee reflects on and replies to the assessor's comments; see, e.g., Gielen et al., 2010; Kim, 2005). The ''a posteriori reply form'' stimulates students to reflect on the peer feedback they received and demonstrate how they used the peer feedback in their revisions, closing the feedback loop (Boud, 2000; Webb & Mastergeorge, 2003).

1.5. Aims of the present study – Hypotheses

The present study aimed, first, to assess the role of seven characteristics (fourth column in Table 1) of the content and style of the provided feedback in terms of their impact on learning. These characteristics are a synthesis of key characteristics identified by Kim (2005), Sluijsmans et al. (2002) and Prins et al. (2006), and are termed ''constructive feedback''. It also aimed to assess the effectiveness of an ''a posteriori reply form'' to support learning after peer feedback.

Specifically, the present study focused on peer feedback in secondary education. These students are less experienced compared to higher education students (Kim, 2005; Sluijsmans et al., 2002) and professionals (Prins et al., 2006), who were targeted in the previous studies. Hence, the feedback form was not as open as the one used in the above mentioned studies; rather a scripting approach was applied, with directed questions (Miller, 2003) and sentence openers (Baker & Lund, 1997) being used as guiding prompts.

The main research questions were the following: (a) Are the constructive feedback characteristics able to raise performance? (b) Do constructive feedback characteristics add to the effects of feedback accuracy? (c) Is the ''a posteriori reply form'' able to enhance performance?

The following hypotheses were tested: (a) Constructive feedback characteristics, namely appropriateness, specificity and formulation of peer feedback, presence of positive and negative comments, of justifications, of suggestions, and of thought-provoking questions will have a positive effect on students' performance (Hypothesis 1). (b) The accuracy of the critique in peer feedback will be positively associated with performance improvement, but the appropriateness, specificity and formulation of peer feedback, presence of positive and negative comments, of justifications, of suggestions and of thought-provoking questions, will have an additional positive effect (Hypothesis 2). (c) Students with an ''a posteriori reply form'' will show higher performance improvement compared to students without this form (Hypothesis 3).

2. Method

2.1. Participants

Participants in the study, which took place during the second and third trimester, were 43 seventh-grade (first year secondary school) students ($M = 13$ years, $SD = 0.22$; 28 males). They were recruited from two classes of the same school, taught by the same teacher (class sizes were 22 and 21 students). All students were enrolled in the theory-oriented general secondary education track.

2.2. Research design

The study adopted a quasi-experimental repeated measures pretest-treatment-posttest design (Fig. 1). It was embedded in the Dutch language writing curriculum. Students studied characteristics of several types of essays – a story,

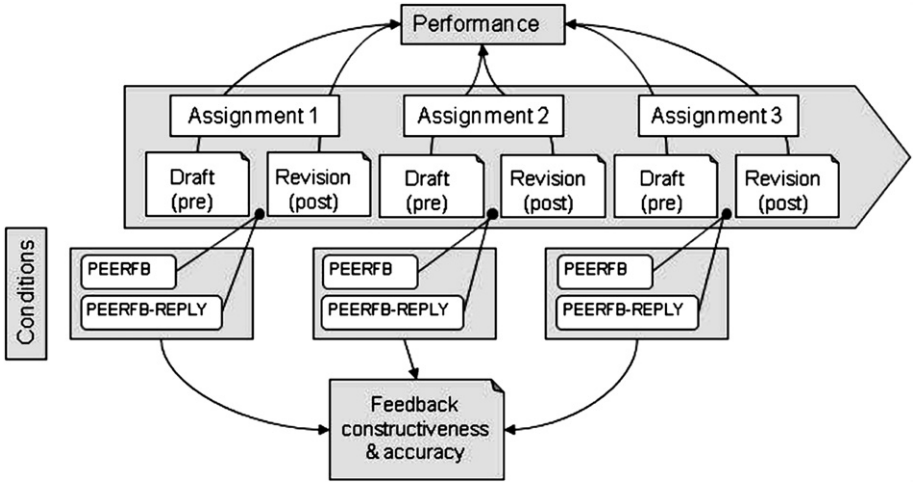

Fig. 1. Research design.

a newspaper article, a letter to the editor – and wrote an essay in each genre for summative assessment. For the present study, the assignments were transformed into two-stage tasks, that is, draft (pretest) and final text (posttest), with intermediate feedback by peers. Students conducted three successive writing assignments (at the beginning of March, the end of April and the end of May).

To provide peer feedback, each student was paired with a classmate of approximately equal ability in line with the "equal status students" principle. First trimester writing scores were used to pair students to a peer who was just above or below in the rank order. Students were not paired bidirectionally – they provided feedback to a peer other than the peer from whom they received feedback – to prevent the possibility that the quality or tone of comments would become conditional on comments received in the previous round. As a result, there were 43 pairs of an assessor and an assessee. For each assignment students wrote a draft essay (pretest); then peers provided written feedback after which the essay was rewritten (posttest) and submitted for summative marking by the teacher. The assignments and peer feedback were not anonymous. Providing peer feedback was mandatory, but peer feedback quality was neither rewarded nor sanctioned.

2.2.1. Conditions

There were two conditions in terms of feedback, namely the peer feedback with reply (PEERFB-REPLY) condition and peer feedback without reply (PEERFB) condition. The two classes were randomly assigned to the conditions (Fig. 1). In the PEERFB-REPLY condition students are asked to report – in a written reply to the teacher – which comments they took into account, how they addressed them, what they learned from providing peer feedback, and to reflect on their own accomplishments. The PEERFB condition was the "plain peer-feedback" condition.

2.3. Instruments

2.3.1. Feedback forms

A description of the feedback forms' structure and guiding prompts is provided in Appendix A. There were two types of feedback forms, an "assessor form" used in both conditions (Feedback form) and an "assessee form" only used in the feedback with reply condition (A Posteriori Reply form). In the Feedback form, students provided open-ended feedback for each criterion listed in the scoring rubric of Appendix B. Within each of six pre-structured paragraphs, guiding questions prompted for a positive and a negative comment with an explanation and suggestions for improvement (which is why the "structure" criterion was not included as a quality criterion in this study, see fourth column in Table 1). In the A Posteriori Reply form, several prompts for reflection were provided. Most prompts applied to the assignment in general and were given once; the prompt on revisions was repeated three times to allow assessees to describe their three most important revisions.

2.3.2. Peer feedback constructiveness and accuracy

Constructiveness of feedback was conceptualized as "quality criteria" based on Prins et al. (2006). The quality criteria were appropriateness to the assessment criteria, specificity, presence of justifications, presence of suggestions for improvement and clear formulation (see Table 1). "Presence of thought-provoking questions" was originally considered, but since it appeared only once in our data it was dropped during the analyses. The scoring of "presence of positive as well as negative comments" posed difficulties. The definitions by Sluijsmans et al. (2002) and Prins et al. (2006) were not considered completely valid, since a zero score could mean either that important comments were missing or that there were no substantial comments to be made (e.g., the content criterion was met). We explored an adapted definition, in which "unless no negative comment possible" was added, but this appeared to be dependent on the judgement of performance, resulting in an unsatisfactory interrater agreement on the second judgement. As a result the category "presence of positive as well as negative comments" was omitted.

Whereas Prins et al. (2006) used the entire report as their unit of analysis, in the present study the unit was the feedback paragraph, that is, each of the six paragraphs with a focus on one particular content criterion. Paragraph scores were averaged to represent the scoring of the essay (cf. Kim, 2005). The averaged score for each feedback characteristic (across the criteria mentioned above) was calculated for each measurement occasion. Interrater reliability indices (independent scoring of a subset of the Feedback forms by the research assistants and the first author) for the different criteria was as follows: appropriateness, $r = .73$, $p < .01$; specificity, $r = .63$, $p < .01$; justification, $r = .49$, $p < .05$; suggestion, $r = .82$, $p < .01$; and clear formulation, $r = .81$, $p < .01$. When scoring feedback constructiveness, we also registered how many content criteria, that is, how many feedback paragraphs, contained accurate negative comments. A first condition was that the paragraph contained at least one critique, identified weakness or suggestion for improvement with an implicit critique. A second condition was that this remark was accurate. This judgement was based on a content analysis of the draft version of the assignment, and was based on the subject-related expertise of the judges. The interrater reliability was acceptable, $r = .65$, $p < .01$.

2.3.3. Performance measure

Two research assistants (the second author and a colleague) both rated the quality of the draft (pretest) and final (posttest) essays. A scoring rubric (Appendix B) was used to assist reliable assessment of performance, based on criteria for essay assignments collaboratively defined by the students and their teacher. The maximum score for each individual essay was 12. A subset of the essays was rated independently by the research assistants and the teacher, with an acceptable interrater reliability, $r = .74$, $p < .01$.

2.4. Procedure

Peer assessment was novel to the students and teacher. At the start of the study the rationale for peer feedback and the Feedback and A Posteriori Reply forms were explained. A worked-out example was used to model the peer assessment process and teachers formulated the assessment criteria together with their students. During the three peer-assessment sessions help was provided if students did not understand how to give feedback. Although training students in giving feedback is important in peer assessment (Sluijsmans et al., 2002), it would have made the feedback quality more homogeneous, and potentially counteract our goal of studying the natural variety of peer feedback skills and their relation with learning progress.

Some data were missing. For ten assignments (three times assignment one, six times assignment two and once assignment three) there was no pretest available. In five of these cases, also the Feedback form and the posttest was missing, in two cases the Feedback form was missing but the posttest not, and finally in the other three cases only the pretest was missing. Since the pretest score was always the first variable being controlled for in the analyses, these 10 observations were deleted listwise. There was one extra assignment for which only the posttest –being the dependent variable– was missing, and one assignment for which only the Feedback form, and thus information on peer feedback constructiveness and accuracy, was missing. To be able to compare the models with variables on feedback quality as nested models to the simpler models, this last observation has also been deleted for the simpler models.

These 12 cases with missing data appeared to be significantly associated with the mean pretest performance on the other assignments, $F(1, 41) = 5.73$, $p = .021$. There were no students who had missing data for all three assignments. The repeated measures structure of the data collection allowed to a certain extent to compensate for the missing information on some students at some measurements.

2.5. Analyses

Observations (performance scores) were nested within students; hence a multilevel approach was used instead of ordinary least squares regression or analysis of variance. Multilevel modeling provides more correct standard errors, confidence intervals and significance tests, it enables a more flexible exploration of covariates, and it can deal with a non-balanced data structure (Goldstein, 1995; Snijders & Bosker, 1999). Students' final essay scores are located at the first level, and the second level is the student. A cross-classification of student by assignment was also explored, but yielded no better fit. In testing the hypotheses, the effect of the feedback characteristics, of accuracy and of condition on performance was examined, with initial performance (pretest score) as a covariate. The SAS system (SAS Institute, 2004) was used to conduct all analyses.

3. Results

3.1. Performance

The descriptives of the performance measure at the pre- and posttest per condition and per assignment are presented in Table 2. The entry level (i.e., pretest scores for the first assignment) of the two classes did not differ, $F(1, 37) = 2.87$, $p = .099$.

3.2. Constructiveness of feedback

The constructiveness criteria (feedback quality indicators) were not all present to the same extent in students' Feedback forms. Table 3 provides the mean and standard deviations for the constructiveness criteria per assignment and condition. The mean score for "clear formulation" and "appropriateness" was highest and for "specificity" and "justification" lowest.

Table 4 reveals that there was a large variation in the number of feedback paragraphs with accurate critique (negative comments). Some Feedback forms contained no accurate critique at all, whereas others contained accurate critique in all feedback paragraphs.

3.3. Effects on performance

3.3.1. Unconditional means model and model with pretest as covariate

The multilevel null model (Model A, Table 5) shows that most variance was attributable to the assignment level. The intraclass correlation showed that 35% of the total variance in students' performance was situated at the student level. Performance on the pretest (draft essay; Model B, Table 5) was as expected, a highly significant predictor of performance on the posttest as indicated by the fact that the original variance of the null model (Model A) was reduced with 69% (in Model B). Of the remaining variance in students' performance (in Model B) 11% was situated at the student level.

3.3.2. The constructiveness of feedback

The multilevel analyses revealed that no constructiveness criteria had a significant main effect on performance, if the

Table 2
Mean performance scores (and *SD*) per assignment, measurement occasion (pre-and posttest), and condition.

			Pretest		Posttest	
Assignment	Condition	*N*	*M*	*SD*	*M*	*SD*
1	PEERFB	19	7.58	1.55	8.37	1.40
	PEERFB-REPLY	20	6.75	1.51	7.83	1.67
2	PEERFB	20	7.55	1.21	8.23	1.31
	PEERFB-REPLY	16	7.66	2.13	8.22	2.02
3	PEERFB	22	8.02	1.37	8.75	1.03
	PEERFB-REPLY	20	7.15	1.55	8.33	1.79

Table 3
Mean scores and *SD* across students for the constructiveness criteria per assignment and condition.

	Appropriateness		Specificity		Justification		Suggestion		Formulation	
Assignment	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
	PEERFB condition									
1	1.89	0.15	0.61	0.47	0.62	0.40	0.65	0.39	1.89	0.25
2	1.67	0.31	0.41	0.37	0.62	0.29	0.74	0.46	1.75	0.36
3	1.67	0.36	0.39	0.39	0.57	0.29	0.78	0.53	1.62	0.52
Overall	1.74	0.31	0.47	0.41	0.60	0.32	0.73	0.46	1.75	0.41
	PEERFB-REPLY condition									
1	1.70	0.19	0.52	0.36	0.59	0.28	0.85	0.47	1.87	0.27
2	1.57	0.35	0.49	0.35	0.36	0.28	0.70	0.37	1.63	0.47
3	1.76	0.31	0.41	0.40	0.48	0.37	0.91	0.47	1.68	0.52
Overall	1.68	0.29	0.47	0.37	0.49	0.32	0.83	0.45	1.73	0.44

An assessor's score, (e.g., for appropriateness) is the mean score for appropriateness of the feedback across all feedback paragraphs in the feedback form of a certain assignment (max. score = 2).
An assessor's overall score (e.g., for appropriateness) is his mean score across all three assignments.

only covariate is performance on the pretest (these models are not mentioned in Table 5). However, adding the interaction term of justification with the pretest performance scores showed that justification had a significant effect ($p = 0.007$, tested one-sided; Model C, Table 5) on posttest performance scores depending on the pretest performance of the student. Justifications in the feedback raised the performance of the assessee, but this effect was weaker when the pretest performance was already high.

3.3.3. Constructiveness versus accuracy of feedback

A similar relationship was found between the number of accurate negative comments and posttest performance, after controlling for initial performance (Model D). Specifically, a significant positive effect was found for accuracy as well as an interaction of accuracy with pretest performance ($p = .014$, tested one-sided; Model D, Table 5). The main effect of accuracy, however, was significant at the .05 alpha level, whereas the main effect of justification (Model C) was significant at the .01 alpha level. Moreover, the model fit of Model C (see AIC and BIC in Table 5) was better than that of Model D.

Adding accuracy and its interaction with pretest performance to Model C along with the significant criterion "justification" and its interaction with pretest performance, accuracy and justification both lost their predictive power (Model E, Table 5). Nevertheless, the Pearson correlation between the two predictors, $r = .52$, $p < .01$, suggests no severe multicollinearity.

Table 4
Mean and *SD* for the number of feedback paragraphs with accurate negative comments per assignment and condition.

Assignment	*M*	*SD*	Min.	Max.[a]	*N*
	PEERFB condition				
1	2.42	0.90	1	4	19
2	2.50	1.24	0	5	20
3	2.36	1.40	0	5	22
Overall	2.43	1.19	0	5	61
	PEERFB-REPLY condition				
1	2.55	0.83	1	4	20
2	1.75	1.13	0	4	16
3	2.65	1.39	0	6	20
Overall	2.36	1.18	0	6	56

[a] Observed maximum out of the six feedback paragraphs.

3.4. Effects of condition

The multilevel analysis with condition as predictor (Model F) revealed that the PEERFB-REPLY condition had no significant effect as compared to the PEERFB condition, when controlling for pretest performance (Model F, Table 5). The interaction between condition and pretest performance did not change this outcome.

4. Discussion

The present study investigated which peer feedback characteristics affect performance in a secondary education setting, and examined whether these characteristics added to the effect of feedback accuracy. The impact of an 'a posteriori reply form' (an instructional intervention to raise mindful reception of feedback and close the feedback loop) on performance improvement was also studied.

In partial agreement with Hypothesis 1 that feedback constructiveness would affect performance improvement, the presence of justification (if accuracy of feedback is not taken into account) significantly improved performance, but only for those with low pretest performance. The effect of "justification" is in line with findings on the informative value of feedback (Bangert-Drowns et al., 1991; Narciss & Huth, 2006) and on the importance of explanations and justifications in help (Webb, 1991) and in collaborative problem-solving (Chiu, 2008; Coleman, 1998). On the contrary, Kim (2005) found no effects for marks and feedback with a justification and specific revision suggestion, but in that study the "constructiveness criteria" were included in a global measure and this might have masked the effect of the separate criteria. Sluijsmans et al. (2002) and Prins et al. (2006) also applied a global measure. However, our findings suggest that it is better to treat the quality criteria separately.

Table 5
Model estimates for the repeated measures analyses of performance.

Parameter	Model A	Model B	Model C	Model D	Model E	Model F
N	117	117	117	117	117	117
Fixed						
Intercept	8.25***	2.32***	0.64	0.30	−0.07	2.23***
	(0.18)	(0.40)	(0.76)	(0.92)	(0.94)	(0.43)
Pretest performance		0.80***	1.01***	1.03***	1.07***	0.81***
		(0.05)	(0.10)	(0.11)	(0.12)	(0.05)
Justification			2.69**		2.15	
			(1.08)		(1.38)	
Justification × pretest performance			−0.33*		−0.28	
			(0.14)		(0.18)	
Number of accurate negative comments				0.70*	0.34	
				(0.31)	(0.40)	
Number of accurate negative comments				−0.08*	−0.03	
× pretest performance				(0.04)	(0.05)	
Condition PEERFB-REPLY						0.11
						(0.18)
Random						
Level 2 − student: σ^2_{u0}	0.83	0.08	0.04	0.04	0.03	0.08
	(0.33)	(0.08)	(0.08)	(0.08)	(0.08)	(0.09)
Level 1 − assignment: σ^2_r	1.56	0.67	0.69	0.68	0.69	0.68
	(0.26)	(0.11)	(0.12)	(0.11)	(0.12)	(0.11)
Model fit						
-2 res log likelihood	422.5	303.4	300.3	305.7	305.6	304.6
AIC	426.5	307.4	304.3	309.7	309.6	308.6
BIC	430.0	311.0	307.9	313.3	313.1	312.2

Values in parenthesis are standard errors.
AIC: Akaike's information criterion. BIC: Bayesian information criterion.
For *fixed* effects: * $p < .05$, ** $p < .01$, *** $p < .001$ (one-sided *t*-tests).

Other quality criteria, such as appropriateness, specificity, clear formulation, and presence of suggestions, did not have a significant impact on performance improvement. Unfortunately, the only characteristic with a significant impact on performance, namely justification, is also among the most difficult to teach. Our prompts may not have been sufficient to generate justification, as can be concluded from the low mean scores for justification indicating a low occurrence. This signifies a need for peer assessment training, guidance and quality control, in order to teach students to provide justifications more frequently (Webb & Mastergeorge, 2003). Results of studies that aim to support collaborative problem-solving by encouraging students to justify their own ideas and explain their answers to each others' questions can be highly informative for developing peer assessment training (Chiu, 2008; Coleman, 1998).

Agreeing with the first part of Hypothesis 2, the results suggest that the accuracy of the critique in peer feedback is positively related to performance improvement. However, the present study could not find evidence for an additional effect of constructive feedback, being the second part of Hypothesis 2. However, it was not expected either that accuracy would lose its predictive power when the constructiveness of the feedback was taken into account. The present study showed that if the accuracy of feedback is taken into account together with the presence of justifications in the feedback, neither of them has a significant effect on performance. Nevertheless, comparison of the simpler models with only one of the two predictors suggest that although feedback accuracy is also important (when constructiveness criteria are not taken into account), the effect tends to be smaller than that of justification (when accuracy is not taken into account), in the case of performance improvement. The analyses showed that the model with justification as predictor had a better model fit than the model with accuracy as predictor. This finding suggests that it is reasonable to use peer feedback for learning, although peer feedback cannot be expected to be as accurate as feedback by an expert. It also indicates that it is more important for a peer assessor to provide justification rather than accurate critique in the form of negative comments.

The use of an a posteriori reply form had no significant effect on performance, as was predicted in Hypothesis 3. This finding is in contrast to the suggested importance of applying the received help, as described by Webb (1991), Webb and Mastergeorge (2003) and Gibbs and Simpson (2004). Perhaps the a posteriori reply form did not add anything to the peer feedback procedure, because students in both conditions were expected to revise their essay. Nevertheless, these findings are in contrast with a finding by Gielen et al. (2010) who used the same peer assessment procedure and feedback forms. They found that assessees using the a posteriori reply form showed more performance improvement compared to the plain peer feedback condition. An important difference, however, is that

the present study focused on short-term effects in the revision phase, whereas Gielen et al. (2010) examined writing progress between the start and the end of a full semester, after experiencing peer feedback on three intermediate assignments. Instructional interventions might not pay directly in the short term, but appear to do so in the long term. Perhaps students become more reflective self-regulated learners by using the a posteriori reply form, and this helps them to perform better on a later assignment without peer feedback (Gielen et al., 2010). Further research is needed to investigate short-term and long-term learning effects of peer feedback and the effects of the a posteriori reply form.

4.1. Methodological limitations

The results of the present study are conditional upon certain choices in the research design and procedure. Performance measures were based on scores for essays written and revised at home. Although this is common practice in writing instruction, it results in a less controlled research design: peers, friends or parents might have influenced the writing process. Moreover, students did not only receive peer feedback in the revision phase, but also provided peer feedback themselves. Providing peer feedback is a learning experience, which may also lead to improvement of students' own writing. It was not, however, the aim of this study to examine whether peer feedback had an impact on performance compared to a control condition without peer feedback (e.g., Gielen et al., 2010), nor to examine whether it is the only factor that affects performance improvement; rather the aim was to explore possible features of peer feedback that are able to enhance its effect on learning.

A second limitation of the study was that only one class was assigned to each condition. Although the teacher, the timing and the assignments were the same for both classes, and no differences were found regarding entry level, this limitation should be taken into account when considering the extent to which our results can be generalised.

Finally, although acceptable interrater reliability was achieved for the essay scores and judgements on feedback quality, this type of 'performance assessment' of essays and written feedback remains a challenging task and requires extensive rater training. Future research could focus on the development of more elaborate rating schemes, such as including more aspects of peer feedback as well as increasing the accuracy of rating.

4.2. Practical implications

In line with Sluijsmans et al. (2002) and Prins et al. (2006), the findings of the present study suggest that training students in providing constructive feedback could be at least as efficient in raising performance of assessees as trying to avoid that peer assessors make inaccurate comments. An important message for practice is that apart from validity and reliability – which have been the main focus in many prior studies (Falchikov & Goldfinch, 2000; Magin & Helmore, 2001) – the quality of peer feedback can affect its impact. This quality can likely be enhanced by guiding prompts and specific training of assessors. Furthermore, instructional interventions to raise a mindful reception of the feedback should be explored further, because feedback left unattended or not acted upon cannot be effective.

Acknowledgements

This research was conducted with the financial support of the Research Foundation, Flanders (FWO - Vlaanderen). The authors want to thank the reviewers, the guest editors, and the journal editor for their valuable feedback.

Appendix A. Forms and prompts

Form	Condition	Structure	Prompt type	Prompt formulation
Feedback form	PEERFB and PEERFB-REPLY	Per criterion (max. 6, see Appendix B)	1. Strengths + justification	What did he/she do well and why?
			2. Weaknesses + justification	What didn't he/she do well and why?
			3. Suggestions	If I were you I would …, Maybe you could …, It would even be better if you …
A posteriori reply form	PEERFB-REPLY	Once	1. Reflection on comments	From the comments of my critical friend, I particularly remember that …
			2. Reflection on assessor role	Assessing the essay of somebody else, I learned that …
		Max. 3	3. Reflection on revisions: identification of criterion with revision + justification + clarification of revision	After the 'critical friend-assignment' I revised my essay with regard to … (criterion) because … and I tried to solve this by …
		Once	4. Reflection on strength + justification	My best piece is, … because …
		Once	5. Reflection on point of attention + justification	I paid this time special attention to … since …

Appendix B. Scoring rubric

Assignment 1: Urban Legend (UL) or Story*	Assignment 2: Newspaper article	Assignment 3: Letter to the editor
Criterion 1: genre-specific elements (max. 2 points)		
UL: Remarkable story (1 point) UL: Can be a true story (1 point) Story: IPADE-structure** (Introduction - Problem - Action - Denouement - End) (2 points)	Presence of a lead (1 point) Stipulation of place and/or time (1 point)	Reference to original article (1 point) Address to audience or name of the author (1 point)
Criterion 2: genre-specific elements (max 2 points)		
Suspense/Originality/Humour (2 points) (one of these should be present)	Objectivity (2 points) (no personal opinion)	Good arguments (2 points): Clear arguments (1 point) and personal opinion (1 point)
Criterion 3: clarity (max. 2 points)		
No inconsistency (0.5 point) Good sentence sequence (0.5 point) Story is clear after first reading (1 point) If story is too short (+/− 0.5 page), then maximum score is 1 of 2 points	Objectivity (2 points) (no personal opinion)	Good arguments (2 points): Clear arguments (1 point) and personal opinion (1 point)
Criterion 4: variety (max. 2 points)		
Sentence structure, use of words (1 point) Limited or no "vague" verbs ("passe-partout verbs")*** (1 point)	Objectivity (2 points) (no personal opinion)	Good arguments (2 points): Clear arguments (1 point) and personal opinion (1 point)
Criterion 5: readability (max. 2 points)		
Not too long sentences (0.5 point) Correct vocabulary, not too difficult words (0.5 point) Consistency of tense (1 point)	Objectivity (2 points) (no personal opinion)	Good arguments (2 points): Clear arguments (1 point) and personal opinion (1 point)
Criterion 6: spelling (max. 2 points)	Objectivity (2 points) (no personal opinion)	Good arguments (2 points): Clear arguments (1 point) and personal opinion (1 point)
For 0 or 1 error the score is 2; For 2 or 3 errors the score is 1; For more than 3 errors the score is 0.		

*Students could choose to write one of both.
**IPADE-structure = The presence of five consecutive phases (introduction, problem, action, denouement, end) in the story; Denouement = The end of a story in which everything is explained.
***"Passe-partout verb" = A verb that is so vague that it does not add much detail to the story (e.g., to go, to make, to do).

References

Baker, M., & Lund, K. (1997). Promoting reflective interactions in a CSCL environment. *Journal of Computer Assisted Learning, 13*, 174–193.

Bangert-Drowns, R., Kulik, C. L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*, 213–238.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*, 7–74.

Bloxham, S., & West, A. (2004). Understanding the rules of the game: marking peer assessment as a medium for developing students' conceptions of assessment. *Assessment and Evaluation in Higher Education, 29*, 721–733.

Boud, D. (2000). Sustainable assessment: rethinking assessment for the learning society. *Studies in Continuing Education, 22*, 151–167.

Chiu, M. M. (2008). Flowing toward correct contributions during groups' mathematics problem solving. *The Journal of the Learning Sciences, 17*, 415–463.

Cho, K., Chung, T. R., King, W. R., & Schunn, C. D. (2008). Peer-based computer-supported knowledge refinement: an empirical investigation. *Communications of the ACM, 51*(3), 83–88.

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

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline. *Computers and Education, 48*, 409–426.

Cho, K., Schunn, C. D., & Charney, D. (2006). Commenting on writing: typology and perceived helpfulness of comments from novice peer reviewers and subject matter experts. *Written Communication, 23*, 260–294.

Coleman, E. (1998). Using explanatory knowledge during collaborative problem solving in science. *The Journal of Learning Sciences, 7*, 387–427.

Falchikov, N. (1995). Improving feedback to and from students. In P. Knight (Ed.), *Assessment for learning in higher education* (pp. 157–166). London: Kogan Page.

Falchikov, N. (1996, July). *Improving learning through critical peer feedback and reflection*. Paper presented at the HERDSA Conference 1996: Different approaches: Theory and practice in Higher Education, Perth, Australia.

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.

Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education, 1*, 3–31.

Gielen, S., Tops, L., Dochy, F., Onghena, P., & Smeets, S. (2010). A comparative study of peer and teacher feedback and of various peer feedback forms in a secondary school writing curriculum. *British Educational Research Journal, 36*, 143–162.

Goldstein, H. (1995). *Multilevel statistical models*. London: Arnold.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99*, 445–476.

Hanrahan, S. J., & Isaacs, G. (2001). Assessing self- and peer-assessment: the students' views. *Higher Education Research and Development, 20*, 53–70.

Kali, Y., & Ronen, M. (2008). Assessing the assessors: added value in web-based multi-cycle peer assessment in higher education. *Research and Practice in Technology Enhanced Learning, 3*, 3–32.

Karegianes, M. L., Pascarella, E. T., & Pflaum, S. W. (1980). The effects of peer editing on the writing proficiency of low-achieving tenth grade students. *Journal of Educational Research, 73*, 203–207.

Kim, M. (2005). *The effects of the assessor and assessee's roles on preservice teachers' metacognitive awareness, performance, and attitude in a technology-related design task.* Unpublished doctoral dissertation, Florida State University, Tallahassee, USA.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254–284.

Magin, D., & Helmore, P. (2001). Peer and teacher assessments of oral presentation skills: how reliable are they? *Studies in Higher Education, 26*, 287–298.

Miller, P. (2003). The effect of scoring criteria specificity on peer and self-assessment. *Assessment and Evaluation in Higher Education, 28*, 383–394.

Mory, E. H. (2003). Feedback research revisited. In D. H. Jonassen (Ed.), *Handbook of research for educational communications and technology* (pp. 745–783). New York: Macmillan.

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merrill, J. Van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed.). (pp. 125–143) New York: Erlbaum.

Narciss, S., & Huth, K. (2006). Fostering achievement and motivation with bug-related tutoring feedback in a computer-based training for written subtraction. *Learning and Instruction, 16*, 310–322.

Nelson, G. L., & Murphy, J. M. (1993). Peer response groups: do L2 writers use peer comments in revising their drafts? *TESOL Quarterly, 27*, 135–142.

Orsmond, P., Merry, S., & Reiling, K. (2002). The use of exemplars and formative feedback when using student derived marking criteria in peer and self-assessment. *Assessment and Evaluation in Higher Education, 27*, 309–323.

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

Rada, R., & Hu, K. (2002). Patterns in student–student commenting. *IEEE Transactions on Education, 45*, 262–267.

SAS Institute. (2004). *SAS/STAT 9.1 user's guide*. Cary, NC: Author.

Searby, M., & Ewers, T. (1997). An evaluation of the use of peer assessment in higher education: a case study in the school of music, Kingston University. *Assessment and Evaluation in Higher Education, 22*, 371–383.

Slavin, R. E. (1989). Research on cooperative learning: an international perspective. *Scandinavian Journal of Educational Research, 33*, 231–243.

Sluijsmans, D. M. A., Brand-Gruwel, S., & Van Merriënboer, J. J. G. (2002). Peer assessment training in teacher education: effects on performance and perceptions. *Assessment and Evaluation in Higher Education, 27*, 443–454.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

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.

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

Tsui, A. B. M., & Ng, M. (2000). Do secondary L2 writers benefit from peer comments? *Journal of Second Language Writing, 9*, 147–170.

Van den Berg, I., Admiraal, W., & Pilot, A. (2006). Peer assessment in university teaching: evaluating seven course designs. *Assessment and Evaluation in Higher Education, 31*, 19–36.

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 Berg, 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.

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

Webb, N. M. (1991). Task-related verbal interaction and mathematics learning in small groups. *Journal for Research in Mathematics Education, 22*, 366–389.

Webb, N. M., & Mastergeorge, A. (2003). Promoting effective helping behavior in peer-directed groups. *International Journal of Educational Research, 39*, 73–97.

Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15*, 179–200.