Studies in Higher Education
Vol. 31, No. 3, June 2006, pp. 341–356

Design principles and outcomes of peer assessment in higher education

Ineke van den Berg*, Wilfried Admiraal and Albert Pilot
Utrecht University, Netherlands

This study was aimed at finding effective ways of organising peer assessment of written assignments in the context of teaching history at university level. To discover features yielding optimal results, several peer assessment designs were developed, implemented in courses and their learning outcomes evaluated. Outcomes were defined in terms of the revisions students made, the grades of the written products, and the perceived progress of products and writing skills. Most students processed peer feedback and perceived improvement in their writing as a result of peer assessment. Significant differences between grades of groups using or not using peer assessment were not found. Most teachers saw better-structured interaction on the subject of writing problems in their classes. Important design features seemed to be the timing of peer assessment, so that it will not coincide with staff assessment, the assessment being reciprocal, and the assessment being performed in feedback groups of three or four students.

There is an increasing amount of attention being given in higher education to the concept of peer assessment, which can be understood as an educational arrangement in which students assess the quality of their fellow students' work and provide each other with feedback (Dochy *et al.*, 1999, Van den Berg, 2003). This development is in line with other recent developments in university teaching, such as collaborative learning and writing, and real-life task performance (see, for example, Van Weert & Pilot, 2003). Studies on peer assessment in higher education have shown positive effects on students' writing skills. Students learn from assessing and commenting on the writings of peers (Topping, 1998).

The purpose of our study was to find effective ways of organising peer assessment of written assignments ('products') in the context of teaching history at university level. 'Effective' is here understood as easily implemented and producing good learning outcomes. In order to examine features of peer assessment designs contributing

*Corresponding author: IVLOS, Utrecht University, P.O. Box 80127, 3508 TC Utrecht, The Netherlands. Email: b.a.m.vandenberg@ivlos.uu.nl

ISSN 0307-5079 (print)/ISSN 1470-174X (online)/06/030341–16

DOI: 10.1080/03075070600680836

to a course's optimal results, we developed, implemented and evaluated several such designs.

Design of peer assessment in higher education

Peer assessment can be understood as a type of collaborative learning (see Falchikov, 2001), but is more limited. It simply means that students assess each other's work using relevant criteria, and give feedback, not only for the benefits of the receiver but also for the purpose of their own development.

Flower *et al.* (1986), studying reviewing processes of experts and novices, observed that beginning writers are not yet able to systematically identify, evaluate and resolve the inaccuracies and problems in their text. According to Flower *et al.*, they have no clear idea about the standards it has to meet. Therefore, we suggest that it would appear profitable for students to review others' work and give them supportive feedback, as a means of internalising the standards for academic writing. Moreover, peer assessment of students' writing presents them with an authentic task, as it closely resembles students' future professional practice at the level of a higher education graduate, in which their texts will be assessed and commented upon by colleagues, or, for example, the editors of a journal. This 'real-life' character will make it easier to motivate and instruct students as to the proper performance of the task (Ten Berge *et al.*, 2004).

Since there are many ways of organising peer assessment, it is important to know which combination of design characteristics, in a certain context, would be likely to yield the best results. To research this question, we developed different designs, employing the 17 variables Topping (1998) found in reported systems (Table 1). We clustered these variables into four groups and looked for opportunities to vary within the given situation. Seven variables were not to be varied, for reasons of practicality (curriculum area/subject, year, time, requirement), or pedagogics (objectives, focus), or because the teachers preferred not to vary (official weight).

Cluster 1 (variables 1–6) relates to the function of peer assessment as an assessment instrument, 4 and 5 offering opportunities for variation. With regard to 4 (Product), peer assessment is applicable to different types of products and performances. For variable 5 (Relation to staff assessment), peer assessment can be intended to either supplement or substitute teacher assessment. Cluster 2 (variables 7–9) concerns the mode of interaction. With regard to 7 (Directionality), two-way assessment has the assessors and assessees consecutively switch roles. One-way assessment means that the assessor is to be assessed by students other than the one(s) they have to assess. For variable 8 (Privacy), the outcomes of the assessment may be presented in a plenary session, or within the 'feedback groups': i.e. small groups of students assessing and discussing one another's work. Finally, assessment may take place, partly or entirely, with or without face-to-face contact (9, Contact). Cluster 3 (variables 10–13) relates to the composition of the feedback group. With regard to 11 (Ability), feedback groups may be composed at random, or according to a plan exploiting the differences or similarities in prior knowledge and/or skills between students. For variable 12

Table 1. Typology of peer assessment in higher education (Topping, 1998)

Variable	Range of variation
1 Curriculum area/subject	All
2 Objectives	Of staff and/or students? Time saving or cognitive/affective gains?
3 Focus	Quantitative/summative or qualitative/formative or both?
4 Product/output	*Tests/marks/grades or writing or oral presentations or other skilled behaviours?*
5 Relation to staff assessment	*Substitutional or supplementary?*
6 Official weight	Contributing to assessee final official grade or not?
7 Directionality	*One-way, reciprocal, mutual?*
8 Privacy	*Anonymous/confidential/public?*
9 *Contact*	*Distance or face to face?*
10 Year	Same or cross year of study?
11 Ability	*Same or cross ability?*
12 Constellation Assessors	*Individuals or pairs or groups?*
13 Constellation Assessed	*Individuals or pairs or groups?*
14 Place	*In/out of class?*
15 Time	Class time/free time/informally?
16 Requirement	Compulsory or voluntary for assessors/assessees?
17 Reward	*Course credit or other incentives or reinforcement for participation?*

Note. Variables in which the designs differ are italicised.

(Constellation assessors) and 13 (Constellation assessees), students may have to assess the products of group members individually, or may be required, for example, to first reach consensus on their comments before communicating their feedback to the assessees. The size of the peer groups can vary from two to more participants. Regarding variable 14 (Place), peer assessment may take place inside or outside the classroom.

Next, cluster 4 (variables 14–17) includes such external factors as requirement and reward, with only variable 17 offering an opportunity for variation. For variable 17 (Reward), the teacher can decide to encourage participation by giving course credits.

Designing peer assessment

To discover factors crucial for effective peer assessment at course level, we developed seven types, each of which was implemented in one course of the history curriculum of Utrecht University. The courses were distributed over the entire programme, covering different levels of ability and different types of writing assignments. The reason for this selection was that these teachers showed interest in peer assessment as a helpful tool for students learning to write. Next, a peer assessment design was developed for every course in consultation with the teacher. The design had to be workable and motivating for the teacher. Thus, we developed designs for:

- a first-year course in which students took their first steps in learning how to report on historical research (course 1);
- two second-year courses. In one of them (course 2), students planned, performed and reported on a limited piece of historical research. Subsequently, course 3 had the students perform a more extensive historical study;
- a third-year course, in which the students had to write a biography of an important historian (course 4);
- a third/fourth-year specialisation course, in which students learned to write a newspaper article to a strict deadline (course 5); and
- two third/fourth-year specialisation courses, one an introduction to cultural education, which had the students write an exhibition analysis (course 6), the other one requiring them to summarise and discuss literature in the form of an article (course 7).

At every first meeting of each course, the students were informed about the objectives of peer assessment and the applicable procedure. Besides this, the assessment criteria were explained and illustrated. As a basic method for all seven designs, we adopted the concept of 'advice-centred feedback' (see, for example, Bean, 1996). It was not our intention to have the students grading each others' work. Students were asked to exchange their drafts, which were then assessed according to the same criteria the teacher would use for the final versions. They were asked to record their findings in a standardised assessment form, at the end of which they were also asked to reflect on their comments and formulate at least three recommendations for the writer. Upon receiving peer assessment, one could rewrite one's draft. The teacher monitored the whole process, only providing feedback after the students had received peer feedback.

The differences between designs were based on the operationalisation and combination of the 10 variables identified in Topping's typology (1998). The designs are summarised in Table 2, which includes the features we used as variables.

The first cluster of variables (peer assessment as an assessment instrument) was operationalised by varying the required length and phase of completion of the writing assignments (the products). In all designs, peer assessment was intended to be formative and to yield qualitative feedback. Peer assessment was in none of our designs meant to be substitutional, but the relation to teacher assessment differed. Four designs (courses 1, 2, 3 and 5) featured supplementary peer assessment, in the sense of an extra source of feedback on the draft which was also assessed by the teacher. In course 1, the teacher provided written feedback on the draft without grading it. In course 2, the teacher provided oral feedback on the draft only. In courses 3 and 5, the teachers gave written feedback and marked the draft. When providing written feedback on the draft, the teacher was asked not to give detailed feedback, but only to complete the assessment form, which was similar to the students' form and based on the same criteria. The teacher was asked to have the students always pass their comments first. In courses 4, 6 and 7, the teacher did not assess the draft. In these courses, peer assessment was the only formative assessment, coming before the teacher's end of course assessment.

Table 2. Seven peer assessment designs

Variable	Course						
	1 *n* = 17	2 *n* = 20	3 *n* = 11	4 *n* = 40	5 *n* = 13	6 *n* = 22	7 *n* = 8
(4) Product	Draft paper (10 pp.)	Outline (1–2 pp.) + draft chapter (3–5 pp.)	Draft paper (15 pp.)	Draft biography (10 pp.)	Draft article (5 pp.)	Draft analysis of an exhibition (3–5 pp.)	Draft article (5 pp.)
(5) Relation to staff assessment	Supplementary; peer + teacher feedback	Supplementary; peer + teacher feedback	Supplementary; peer + teacher feedback + teacher grades	Supplementary; peer feedback only	Supplementary; peer+ teacher feedback +teacher grades	Supplementary; peer feedback only	Supplementary; peer feedback only
(7) Directionality	One-way (2 assessments)	Mutual (2 or 3 assessments of both products)	Mutual (1 assessment)	Mutual (2 collective assessments)	Mutual (2 assessments)	Mutual (3 assessments)	One-way (2 assessments)
(8) Privacy	In public (teacher + all students)	Confidential (within feedback group); teacher receives copy	Confidential (within feedback group); teacher receives copy	Confidential (within feedback group); teacher receives copy	Confidential (within feedback group); teacher receives and assesses copy	Confidential (within feedback group); teacher receives copy	Confidential; teacher receives copy
(9) Contact	Written + oral peer feedback; plenary discussion with supplementary feedback from teacher	Written + oral peer feedback	Written + oral peer feedback; plenary discussion of themes brought up by feedback groups	Written + oral peer feedback; plenary discussion of themes brought up by feedback groups	Written + oral peer feedback; plenary discussion of themes brought up by feedback groups	Written + oral peer feedback	Written peer feedback only

Table 2. *Continued*

Variable	Course						
	1 *n* = 17	2 *n* = 20	3 *n* = 11	4 *n* = 40	5 *n* = 13	6 *n* = 22	7 *n* = 8
(11) Ability	Feedback groups selected by teacher (related topics)	Feedback groups selected by teacher (related topics)	Feedback groups selected by students (related topics)	Feedback groups selected at random by teacher (different topics)	Feedback groups selected at random by teacher (same topics)	Feedback groups selected by students (same topics)	Feedback groups selected at random by teacher (different topics)
(12) Constellation assessors	Two students and teacher	Small groups (3–4 students), teacher participates	Two students and teacher	Small groups (4 students, 2 pairs)	Small groups (3 students) and teacher	Small groups (4 students)	Two students
(13) Constellation assessees	Two other students	Same small groups	Same two students	Same small groups	Same small groups	Same small groups	Two other students
(14) Place	Written feedback outside, oral feedback in classroom (plenary discussion and tutorial with teacher)	Written feedback outside, oral feedback in classroom (small groups with teacher)	Written feedback outside, oral feedback in classroom (small groups and plenary discussion)	Written feedback outside, oral feedback in classroom (small groups and plenary discussion)	Written feedback outside, oral feedback in classroom (small groups and plenary discussion)	Written feedback outside, oral feedback in classroom (small groups)	Written feedback outside classroom
(17) Reward	Participation in PA not rewarded	Participation in PA not rewarded	Participation in PA not rewarded	Participation in PA not rewarded	Max.1/4 points bonus for written feedback	Participation in PA not rewarded	Participation in PA not rewarded

Note. The numbers 4–17 in the first column refer to the corresponding topics in Toppings's typology of peer assessment (Topping, 1998); *n* = number of students in the PA condition.

The second cluster (mode of interaction) was operationalised as follows. Only courses 1 and 7 were designed with one-way directionality. Course 7 required written feedback only. Only course 1 had the students presenting oral feedback publicly, at a plenary session. In all courses, students performed the written parts of peer assessment outside the classroom. This included reading the draft, writing down remarks and completing the assessment form. All oral feedback was provided face to face.

The third cluster of variables (composition of feedback group) was operationalised as follows. We varied the size of our feedback groups from two (course 3) to three (course 5) or four (course 2, 4 and 6). In courses 1 and 7, each with one-way assessment, every student assessed the work of two others. Except for courses 2 and 6, the teacher generally grouped students at random. In course 2, the teacher formed feedback groups of students working on related subjects. In course 6, the students had already formed groups of their own, and we saw no reason to change this. With the exception of course 4, students assessed the products of their fellow students individually. In course 4, the assessors had to reach consensus on their feedback before communicating it to the assessees. In courses 5 and 6, students studied the same subjects and material. The subjects of the other courses were the same, similar, or non-related.

In the fourth cluster (external factors), all courses featured mandatory participation in the peer assessment procedures. The quality of the peer assessment was only rewarded in course 5, where students could earn a (small) bonus on top of their final course grade.

In order to determine which design principles are most effective in fostering learning outcomes we focused on two types of results. The first type of outcome to be studied was the quality of the final products. For this purpose, the writings were examined with respect to the kind of revisions students had made in response to the feedback from their fellow students. The second type of outcome was students' progress in writing as perceived both by themselves and the teacher. We related the results to differences in designs of peer assessment.

Method

This study describes peer assessment in seven courses in relation to 10 design features from Topping (1998), in order to determine which combination of design features yields the best results.

Subjects and data collection

Our study involved nine teachers from the History Department of Utrecht University and 168 students from the history programme of the Faculty of Arts, 131 of whom were in peer assessment groups and 37 of whom were not. The latter group did not differ from the former with respect to prior knowledge, efforts and achievements. The peer assessment groups and the 'parallel groups' had the same teacher. These parallel groups were used in courses 4, 6 and 7. Courses 2 and 4 were randomly divided into

two peer assessment groups (a and b) because there were too many participants for one group. Neither teachers nor students had had any previous experience with peer assessment.

Implementation of peer assessment

In order to monitor the implementation of the peer assessment procedures, classroom activities were observed. All writing products and written and oral peer feedback were gathered, and two student questionnaires were administered, the first directly after the students had provided their oral feedback (so before having received the credits for their final version of the product). The questions related to students' time investment in assessing their peers, the workability of the procedures, the usefulness of the received feedback and the usefulness of the assessment form. The second questionnaire was administered at the end of each course (but before the students received the credits for the final examination). In this questionnaire, students from the peer assessment and parallel groups were asked how much time they had spent on writing their products. Classroom observations covered the entire process. This process involved the introduction of peer assessment by the teacher and students' responses, the formation of feedback groups, the exchange of written products and written feedback, participation in oral feedback, plenary discussion after oral feedback, interaction between students, and interaction between students and teacher.

Revisions and grades of the writing product

Data on the revisions students had carried out were obtained by gathering the products in both draft and final version. In order to decide whether revisions had been occasioned by peer assessment, the researchers also collected the written and oral feedback of both students and teachers. The written feedback was gathered by means of the completed assessment forms and the remarks, if any, which the teacher had scribbled in the margins of the product. The oral feedback was collected by tape-recording classroom sessions.

Students' revisions were compared, firstly, with the peer feedback received, and then, to establish whether any revision had been induced by peers or teacher, with the teacher feedback. To gain insight into the intensity of the reviewing process, the revisions and the feedback were coded in terms of content, structure and style (Steehouder *et al.*, 1992). 'Content' refers to presented subject matter, the definition of the problem, argumentation, and usage of conceptual language. 'Structure' refers to inner consistency, especially as to how the main problem is related to the sub-questions, and how the conclusion provides an answer to the main problem. 'Style' refers to the outer form of the text, such as layout and language (including grammar and spelling). The inter-rater reliability of the coding instrument was satisfactory (Cohen's $\kappa = .85$ for feedback function, $n = 88$ oral comments; Cohen's $\kappa = .93$ for feedback aspect, $n = 79$ oral comments). This coding instrument was also used to categorise the type of revisions ($\kappa = .89$, $n = 48$ revisions in 12 writings).

Student evaluation of peer assessment

Students' perceived progress in their writing was measured by means of questions in which students were asked to evaluate their own writing abilities, especially in terms of progress made. The first questionnaire was administered upon assessment and in peer assessment groups only, the second in all groups at the end of the course. The evaluative questions of the first questionnaire were open-ended; the answers were scaled by the researcher on a three-point scale (1 = mainly negative, 3 = mainly positive), inter-rater reliability averaging $\kappa \geq .70$, ranging from .64 to .90, n = 22 questionnaires. The second questionnaire, at the end of the course, was answered by scoring on a five-point scale (1 = no progress at all, 5 = a lot of progress).

Teacher evaluation of peer assessment

A semi-structured interview was conducted with each teacher at the end of each course, immediately after they had graded the final products. The interviewer was a research assistant and not a member of the research team (composed of the three authors of this article). The interview included topics similar to the student questionnaires, including teachers' time spent on assessment, the usefulness of the assessment form, the workability of the procedures, and the perceived effects of peer assessment.

Results

Implementation of peer assessment

The implementation is summarised in Table 3. Generally, procedures were followed as scheduled. About 80% of the students handed in their drafts on time, received feedback from at least one other student, and assessed the work of at least one peer. However, not all carried out the planned number of assessments. Only some two-thirds received the number of comments prescribed in the peer assessment design.

Some participants of course 2a did not receive any feedback at all, despite the teacher's monitoring efforts. Their peers did not regularly attend classes, and did not hand in a draft on time or comply with any peer assessment activity, for reasons unknown to us. In course 7 none of the draft versions were exchanged on time. As it

Table 3. Processing scheduled peer assessment activities (percentage of students participating in each activity)

Courses	1	2	3	4	5	6	7
Draft versions handed in on time	88	78	67	65	100	91	0
Assessment by all members of feedback group	80	55	100	31	75	75	50
Not assessed by peers	0	26	0	0	0	0	0

Note. For course 2, we used an average for two products: outline for a paper, and a draft of a chapter.

turned out, the assignment to prepare a presentation in the field of art took more time than was estimated in the course design, so it superseded peer feedback activities.

The students from course 4, the only course requiring collective assessment, mostly produced individual assessments, explaining that a collective assessment made the process unnecessarily complicated, as it was not easy to find time for meeting outside classes. A factor that negatively influenced the willingness to comply with the procedure was that many students, more than in the other courses, did not know each other. In group 4a the implementation of peer assessment was complicated by a small group of students who raised objections. They saw it as a negative consequence of peer assessment that they had to start their writing earlier, when they themselves had planned to start writing at the end of the course. Next to this, they were afraid that peer assessment would take too much time, because they would have to read and assess the work of other students.

The feedback groups from course 3 consisted of four pairs of students mutually assessing each other and one trio. Some pairs performed poorly, in one case due to one's overbearing behaviour, in the other due to illness, which resulted in written feedback only.

In practice, the amount of time taken for reading and assessing averaged 75% of the time we considered minimal for a serious execution of the assessment task. In conformity with Wijnen *et al.* (1992), we considered the assessment task 'heavy' and the degree of difficulty of the products that had to be assessed of average complexity. According to Wijnen *et al.* (1992), this should result in students reading and assessing about seven pages per hour. At one end of the scale, course 5 resulted in students investing twice the time we thought necessary. This may have been caused by, in this design, the quality of the peer assessment influencing the final grade. At the other extreme, the students invested much less time in course 3 than the average 75% of the time expected to be required. In this case, the students prioritising the input from teacher assessment may have caused them not to put in the required effort. As in this course the first draft would largely determine the final grade, the moment for handing in the assignment was postponed—the teacher succumbed to the pressure of the students—to allow the students more time for producing a good first draft. The teacher also gave written feedback on the first draft. As a result, the peer and teacher feedback coincided. Contrary to the initial design, the teacher feedback was very detailed, and the students were left less time for revision than was planned. Thus, the role of peer assessment became unclear, making it no surprise that it was deemed only a second priority.

Our observations showed that oral feedback in most feedback groups was lively and to the point, albeit that one or two feedback groups in most courses were not as task-oriented as they should have been. This holds specifically for courses 2 (in peer assessment group 2a), 3, 4 (group 4a) and 6.

Revision of the writing products

Most students used some elements of the peer feedback for revision. On average, students processed about one-third of all suggestions, which mainly dealt with

Table 4. Number and type of comments for revision (suggested and processed) in written and oral peer feedback

	Content		Structure		Style		Total	
Course	S	P	S	P	S	P	S	P
4a	20	10	4	1	25	9	49	20
4b	29	7	3	2	26	8	58	17
5	19	8	5	2	46	14	70	24
6	39	11	0	0	57	11	96	22
7	23	7	5	1	19	8	47	16
Total	130	43	17	6	173	50	320	99

Note. Courses 1, 2 and 3 are left out, as teachers' feedback interfered with peer feedback; S = number of suggestions for revision, P = number of suggestions for revision processed.

content and style. Most revisions concerned style, and to a lesser degree content. Comments on structure, as far as they occurred, were hardly used. Parts of the text which had not received any feedback were barely revised, if at all. The extent to which students processed suggestions for revision from their peers is shown in Table 4.

There were differences between courses as to the amount and type of revisions actually made. The absence of structural revisions in course 6 may have been caused by the prescribed themes, which had been presented in considerable detail, more so than in the other courses. Feedback was generally processed less in this course than in the other courses. This may be explained by the absence of a plenary discussion on the oral feedback phase, which would have enabled the students to discuss the questions and differences of opinion presented by the feedback group. On the other hand, about 40% of the suggestions for revision were processed in course 4a, which may be explained by this course producing relatively more drafts not resembling a final version, with some essential parts not written yet. This left writers with more room to complete their work on the basis of their fellow students' remarks.

Student grades

Courses 4, 6 and 7 provided the opportunity to compare the (teacher) grades for the final products of the peer assessment group members with those of the non-peer assessment groups. We found no significant differences ($\alpha = 0.05$).

Courses 3 and 5 allowed us to compare the grades before and after peer assessment, as the teacher graded both the draft and the final product and did so using the same criteria. In both cases the grades for the final products were significantly higher than those for the drafts (course 3: 6.8 for the draft vs. 7.3 for the final version with $t = 3.3$, $df = 9$, $p = .01$; course 5: 5.7 vs. 6.5 with $t = 4{,}3$, $df = 11$, $p = .001$). Correlation is high ($r = .92$ in course 3, $r = .88$ in course 5), and all participants made progress. In course 5 students revised their draft mainly based on peer feedback. In course 3 the revisions are mainly based on teacher feedback.

Teachers' evaluation of peer assessment

The teachers of most courses approved of peer assessment, especially appreciating its influence on students' interaction and involvement in the course. Compared with groups from the same courses the year before, the teachers of courses 1 and 2 experienced improved content-related interaction and increased involvement. As peer assessment had the students studying each other's work, their participation in discussions increased. Moreover, the teachers witnessed better-structured plenary discussion on writing problems and their solutions. According to these teachers this was the result of the use of an assessment form, in which the criteria were clearly formulated and known to all.

When asked whether peer assessment resulted in better writing, teachers' answers varied. Those from courses 4, 6 and 7 had not observed any differences between the (final) products of their groups and those of the control groups. Those from courses 1, 2 and 5, having assessed and commented upon both drafts and final products, agreed that the latter were of higher quality, but were not sure whether this resulted from peer assessment or from their own feedback. As it happened, the teacher of course 2a could compare the assessed final versions with the final versions of four students of the same group who had not participated in peer assessment. According to the teacher, the writings of the latter were lacking in structure: for example, the main problem had not been clearly stated, the research questions did not seem logical, or the conclusion did not provide an answer to the problem. The teacher of course 2b observed that students were more attentive to the process of writing, and thought this to be the result of students having commented upon the work of their fellow students at different stages of completion.

The teachers of courses 1, 2 (a and b) and 5 experienced difficulty in determining their role in the peer assessment system. Complying with the procedures, they still wanted to provide more assistance but found opportunities restricted, as the students had yet to give their feedback first. For this reason, the teacher of course 3 had, on second thoughts, decided to provide detailed instead of general feedback.

Students' evaluation of peer assessment

Most students perceived some progress in their writing product as a result of their having processed peer feedback. Some explained they had improved their style of writing, in particular the grammar, others said they had restructured their product, or had changed some of the content to achieve more relevance. A Kruskal-Wallis analysis of variance shows that students in different designs differ in their opinions (KW = 15.0; df = 7; p = .04). The students from courses 1, 4b, and 5 noted considerable progress in their writing, whereas those from courses 2a and 7 saw hardly any progress. Some students from course 2a felt they had not received any useful peer feedback, assuming their fellow students had not looked at their work seriously, but had made do with some comments on spelling. Table 5 presents students' perception of the progress of their writing product as a result of peer assessment.

Table 5. Students' perception of progress in their writing product as a result of peer feedback

Courses	Mean	Sd	*n*
1	2.1	0.67	12
2a	1.6	0.55	5
3	2.0	0.00	6
4a	2.0	0.53	8
4b	2.4	0.51	12
5	2.0	0.64	13
6	1.8	0.43	14
7	1.5	0.55	6

Note. 1 = no progress; 2 = some progress; 3 = much progress; course 2b omitted, because of the low number of frequencies; *n* = number of respondents.

Students felt their writing skills had progressed in some courses, especially courses 1, 2, 4b and 5. There were differences between the designs, which were tested by means of a parametric variant of analysis of variance ($F = 4.9$; $df = 7.82$; $p \leq .001$) with a considerable effect (with an effect-size f = .65; see Cohen, 1988). As courses 1, 2 and 5 focused more strongly on writing than the other courses, such outcomes are not very remarkable. Besides the writing assignments, courses 6 and 7 presented the students with assignments of quite another type, course 7 including a presentation which proved such a time-consuming task that some found themselves pressed for time in their writing assignment. Table 6 presents the results of students' perception of the progress in their writing skills.

As shown in Table 7, only course 6 shows significant differences (with $\alpha = .05$) between the estimations of peer assessment group students and non-peer assessment group students. Although both groups perceived the writing to have progressed rather little, the difference may have been caused by the implementation of peer assessment.

Table 6. Students' perception of own improved writing skills

Courses	Mean	Sd	*n*
1	4.1	0.67	12
2a	4.2	0.83	9
2b	4.3	0.50	4
4a	2.9	0.88	15
4b	3.5	1.09	12
5	3.4	0.70	11
6	2.9	0.88	19
7	2.6	1.30	8

Note. Scale 1–5: 1 = no progress at all; 5 = a lot of progress; course 3 omitted as erroneously students were not questioned on this aspect; *n* = number of respondents.

Table 7. Descriptive statistics and *t*-test of students' perception of own improved writing: PA vs. non-PA group

Courses	N1	N2	M1	M2	*t*	df	*p*
4a	15	14	2.9	3.0	–.23	27	.82
4b	12	14	3.5	3.0	1.38	18	.19
6	19	12	2.9	2.3	2.25	28	.03
7	8	7	2.6	2.6	.09	13	.93

Note. Scale 1–5: 1 = no progress at all; 5 = a lot of progress; M1 = mean PA group; M2 = mean non-PA group; N1 = number of respondents PA group; N2 = number of respondents non-PA group.

In sum, in terms of students' perceptions, we conclude that peer assessment did indeed produce positive learning outcomes, particularly in courses 1, 4b, 5 and 6. In courses 1, 4b and 5 students thought their final version better for having used peer feedback revising, while in course 6 they felt their writing skills had improved as a result of peer assessment.

Conclusion and discussion

The purpose of our study was to find effective ways of organising peer assessment of writing assignments in the context of university teaching. 'Effective' was understood as being easily implemented and providing optimal learning outcomes. To discover factors crucial to the organisation of peer assessment in a course, we developed and implemented seven designs which we evaluated with respect to their short-term learning outcomes. Outcomes were defined in terms of the revisions students made, the grades of the written products, and the progress students and teachers perceived in products and writing skills.

Generally, peer assessment procedures were carried out as scheduled. After having received peer feedback, most students revised their work, using about one-third of the suggestions. Most revisions related to the style, and to a lesser degree to the content of the draft. When students did comment on structure, the writer rarely used the suggestion. Revision did not lead to higher grades in the peer assessment groups, although most students found their revised products better than the drafts, and felt this was the result of them processing peer feedback.

Where, according to some teachers, the products were of better quality as a result of peer assessment, others observed no difference at all. Most teachers appreciated peer assessment for breaking with the usual one-to-one communication between student and teacher. Also, they experienced a better structured plenary discussion on problems in writing and adequate solutions.

In order to be able to draw conclusions with respect to an optimal design for peer assessment, we related the different kinds of outcomes to design features. In terms of results, courses 4 and 5 scored best on all types of outcomes, with students investing a considerable amount of time, processing, relatively, a lot of peer feedback, and

considering their final products to be an improvement on the drafts as a result of peer feedback. This perception was reflected in course 5 by the improved grades for each student as a result of peer feedback.

In terms of design, we may conclude that there are three design features which were particularly beneficial for the effects of peer assessment in the context of writing. First, with regard to the relation of peer and staff assessment (Topping variable number 5), there must be sufficient time between peer and teacher assessment for students to revise their drafts on the basis of peer feedback before they are required to hand in the product to the teacher. Next, for directionality (7): reciprocal two-way feedback is more easily organised, as it is clear that the assessor will in turn become assessee, making it easier to exchange products.

Finally, regarding the grouping of assessors (12) and assessees (13), the optimal size of feedback groups seemed to be three or four. In groups of this size, students had the opportunity to compare the comments of two or three peers, so that they were able to determine their relevance. Apart from lacking the opportunity to compare, 'groups' of two are vulnerable. As it turned out, there is a risk that two 'weaker' students end up with each other, without much to offer.

Our method, however, has its drawbacks. Our designs were developed in consultation with the teachers, which resulted in compromises. Especially in those courses in which students were assumed to receive a substantial part of their writing training, the teachers took more part in the peer feedback process than we would have liked. Some teachers experienced difficulty in determining their role in the system. The role of the teacher in a peer assessment system deserves more attention.

The outcomes we studied are limited, as we only studied the results within a course, and not the gains one finally aims for in introducing peer assessment. We want students to internalise the criteria of academic writing, becoming able to review their own work.

Finally, there is the issue of generalisation. We have no evidence of to what extent the results of our study are valid for other skills, other disciplines and other types of education. We studied peer assessment designs in a curriculum in which writing has an important place: history students are expected to write a lot in their future professions. This also means that the results of this study cannot be generalised for other curricula, such as physics or vocational practices. From the perspective of the potential relevance of peer assessment in higher education, we consider this to be a challenge for future research.

References

Bean, J. C. (1996) *Engaging ideas: the professors' guide to integrating writing, critical thinking and active learning in the classroom* (San Francisco, CA, Jossey-Bass).

Berg, B. A. M. van den (2003) *PA in universitair onderwijs. Een onderzoek naar bruikbare ont-werpen* [*PA in university teaching: an exploration of useful designs*]. Doctoral dissertation, Utrecht, University of Utrecht.

Berge, H. ten, Ramaekers, S. & Pilot, A. (2004) The design of authentic tasks that promote higher-order learning, *Motivation, Learning, and Knowledge Building in the 21st century*. Book of Abstracts, Conference EARLI-SIG HE (49) (Stockholm, EARLI).

Cohen, J. (1988) *Statistical power analysis for the behavioural sciences* (2nd edn) (Hillsdale, NJ, Lawrence Erlbaum).

Dochy, F., Segers, M. & Sluijsmans, D. (1999) The use of self-, peer and co-assessment in higher education: a review, *Studies in Higher Education*, 24, 331–350.

Falchikov, N. (2001) *Learning together; peer tutoring in higher education* (London, Routledge-Falmer).

Flower, L., Hayes, J. R., Carey, L., Schriver, K. & Stratman, J. (1986) Detection, diagnosis, and the strategies of revision, *College Composition and Communication*, 37, 16–55.

Steehouder, M., Jansen, C., Maat, K., Staak, J. van de, & Woudstra, E. (1992) *Leren communiceren* [*Learning to communicate*] (Groningen, Wolters-Noordhoff).

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

Weert, T. J. van & Pilot, A. (2003) Task-based team learning with ICT, design and development of new learning, *Education and Information Technologies*, 8, 195–214.

Wijnen, W. H. F.W., Wolfhagen, H. A. P., Bie, D. de, Brouwer, O. G., Ruijter, C. T. A. & Vos, P. (1992) *Te doen of niet te doen? Advies over de studeerbaarheid van onderwijsprogramma's in het hoger onderwijs* [*To do or not to do? Advice on the doability of higher education teaching curricula*] (Zoetermeer, Ministerie van Onderwijs).