Design principles and outcomes of peer assessment in higher education

There is increasing attention in higher education for 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 report positive effects of peer assessment, especially on students' writing skills. Students learn from communicating about their work with peers, even more so when assessing both their products and the writing processes (see, for example, a review study by Topping, 1998).

With a view to successfully implementing them in education it is important to know which designs may produce the best results. Our study aims to find effective ways of organizing peer assessment of written assignments ('products') in the context of university teaching.

'Effective' is understood as easily implemented and producing good learning outcomes. In order to examine features of peer assessment designs contributing to a course's optimal results, we developed, implemented and evaluated several of such designs.

Design of peer assessment in higher education

Peer assessment can be understood as a type of collaborative learning (see Falchikov, 2001). However, compared to collaborative, co-operative, or peer learning – in all of which students produce a collective product, share knowledge, and learn from the collaboration - peer assessment is more limited. It simply means students assess each other's work using relevant criteria in crediting the work and the effort, also for the purpose of their own development. Peer assessment in a formative sense indeed would appear helpful to students in developing their ability of reviewing their own texts. Flower *et al*., (1986), studying reviewing processes of experts and novices, observed that especially beginning writers are not yet able systematically to identify, evaluate and solve 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 and give them supportive feedback, as a means of interiorizing the standards for academic writing. Moreover, peer assessment of students' writing presents them with an authentic task, as it closely resembles students' future professional academic practice, 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 peer assessment task (Ten Berge *et al*., 2004).

As there are many ways of organizing 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 of peer assessment (Figure 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) reasons, or because the teachers preferred not to vary (official weight).

= Figure 1 =

Cluster 1 (variables 1-6) relates to the function of peer assessment (PA) as an assessment instrument, the aspects 4 and 5 offering opportunities for variation. As to aspect 4-Product, PA is applicable to different types of products and performances. As to 5-Relation to staff assessment, PA can be intended to supplement or substitute teacher assessment. Cluster 2 (variables 7-9) concerns the mode of interaction, the aspects 7, 8 and 9 offering opportunities for variation. As to 7-Directionality, two-way assessment has the assessors and assessees consecutively switch roles. In one-way assessment, the assessor is assessed by student(s) other than the one(s) he has assessed (directionality). Next, as to 8-Privacy, the outcomes of the assessment may be presented in a plenary session, or in feedback group. 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, the aspects 11, 12 and 13 offering opportunities for variation. As 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. As to 12-Constellation assessors, and 13- Constellation assessees, students may have to assess the products of group members individually, or may be required for example, first to reach consensus about their judgments before communicating their feedback to the assessees. The size of the peer groups can vary from two to more participants. As to 14-Place, PA may take place inside or outside the classroom.

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

Designing peer assessment

To discover factors crucial for effective peer assessment at course level, we developed several peer assessment designs, which were implemented in seven courses of the 1999 History curriculum of Utrecht University. The then standard four-year combined university curriculum in The Netherlands may best be compared to a combined bachelor and master programme. The courses were distributed over

the entire programme, covering different levels of ability, and different types of writing assignments. Thus, we developed designs for:

- A first-year course in which students take their first steps in learning how to report on historical research (course 1);
- Two second-year courses. In one of them (course 2), students plan, perform and report on a limited piece of historical research. Subsequently, course 3 has the students perform a more extensive historical study;
- A third-year course, in which the students have to write a biography of an important historian (course 4);
- A third/fourth-year specialization course, in which students learn to write a newspaper article under a strict deadline (course 5) and
- Two third/fourth-year specialization courses, one an introduction to cultural education, which has the students write an exhibition analysis (course 6), the other one requiring them to summarize 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. As a basic method for all seven designs, we adopted Bean's elaboration of the concept of 'advice-centred feedback' (Bean, 1996). 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 not to grade them, but merely to record their findings in a standardized assessment form, at the end of which they were also asked to reflect on their judgments 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 are based on the operationalization and combination of the ten variables afore mentioned from the typology of Topping (1998). The designs are summarized in Figure 2, which includes the features we used as variables.

= Figure 2 =

The first cluster of variables (PA as an assessment instrument), was operationalized by varying the required length and completeness 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 with teacher assessment differed. Four designs (courses 1, 2, 3 and 5) featured supplementary peer assessment in the sense of 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 credited the draft. When providing written feedback on the draft, the teacher was asked not to give detailed feedback but to restrict herself to completing her assessment form, which was similar to the students' form and based on the same criteria. The teacher was asked always to have the students pass their comments first. In the courses 4, 6 and 7, the teacher did not assess the draft. In the latter peer assessment served as the only formative assessment, coming before the teacher's end of course assessment.

The second cluster (mode of interaction) was operationalized 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 includes reading the draft, making notes and completing the assessment form. All oral feedback was provided face-to-face.

The third cluster of variables (composition of feedback group) was operationalized 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 from 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 the courses 5 and 6, students

studied the same subjects and material, the subjects of the other courses were the same, similar, or non-related.

As to 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 receive up to 2.5% of their final course grade upon having assessed their written feedback

In order to conclude which design principles of peer assessment are most effective in fostering learning outcomes we will focus on two types of results. The first type is outcomes in terms of the quality of the written products in their final versions, which are examined with respect to the kind of revisions students had made in response to the feedback from their fellow students and the grades they received from the teacher. The second type is students' progress in writing as perceived both by themselves and the teacher. We will then relate the results to differences in designs of peer assessment

Method

This study is a multiple design study, describing seven courses in relation to ten design features from Topping (1998). In order to decide which combination of design features yields the best results, we made a set up a course-ordered descriptive meta-matrix (cf. Miles & Huberman, 1994, p.187).

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 in peer assessment groups, 37 in groups without peer assessment. The latter did not differ from those in the peer assessment groups with respect to prior knowledge, efforts and achievements. The PA-groups and these 'parallel groups' had the same teacher. Such parallel groups were employed in the courses 4, 6 and 7. The courses 2 and 4 were divided in two PA groups (a and b), because there were too many participants for one group, the groups were divided at random. Neither teachers nor students had had any previous experience with peer assessment.

Implementation of peer assessment

In order to check the implementation of the peer assessment procedures we observed classroom activities, gathered all writing products and peer feedback, and administered two student questionnaires, the first of which directly after the students had provided their oral feedback (so before having received the credits for their final version of the product). The items related to students' time investment in assessing their peers, the practicability of the peer assessment procedures, the usefulness of the received feedback and the workability 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 exam). In this questionnaire, students from the peer assessment groups and parallel groups were asked about the total amount of time they had spent on writing their products. In terms of reliability and validity, the quality of both questionnaires was satisfactory. Classroom observations covered the entire process, from the introduction of peer assessment by the teacher and students' responses, to how feedback groups were formed, exchange of written products and written feedback, participation in oral feedback, plenary

discussion after oral feedback, interaction between students, and between students and teacher.

Revisions and grades of the writing product

Data on the revisions students 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, we also collected the feedback, written and oral, of both students and teachers. The written feedback was gathered by means of the completed assessment forms and the written notes, if any, the teacher had left scribbled in the margins of the product. The oral feedback was collected by tape-recording classroom sessions.

We proceeded by comparing the revisions made, first with the peer feedback received, then, to establish whether any revision had been induced by peer or teacher feedback, with the teacher feedback. To gain insight into the intensity of the reviewing process, we coded the revisions and the feedback 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 constitutes an answer to the main problem. 'Style' refers to the outer form of the text, for example layout and language (including grammar and spelling). The inter-rater reliability of the coding instrument was satisfactory (Cohen's $\kappa = .93$. This coding instrument was also used to categorize the type of revisions ($\kappa = .89$).

Student evaluation of peer assessment

Students' perceived progress of 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 (both PA and non-PA) at the end of the course. The evaluative questions of the first questionnaire were open-ended; the answers 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.

The second questionnaire, at the end of the course, was answered by scoring on a five-point scale (1= no progress at all, 5= very much 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 not a member of the research team. The interview included topics similar to the student questionnaires, including teachers' time spent on assessment, the workability of the assessment form, the practicality of the PA procedures, and the perceived effects of peer assessment. The research team checked the questions for relevance. The interviewer verified the answers by means of 'member check', i.e. by at several times in the interview summarizing the answers to the respondent, enabling him/her to provide corrections and additions, which occasionally occurred.

Results

Implementation of peer assessment

The implementation of peer assessment is summarized in Table 1. Generally, procedures were followed as scheduled. Around 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.

= Table 1 =

The students of course 4, the only course requiring common assessment, mostly produced individual assessments, explaining that a common assessment turned the PA process unnecessarily complicated, as it was not easy to find the time for meeting outside classes. The feedback groups of course 3 consisted of pairs of students mutually assessing each other, some pairs performing 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 (see Van den Berg, 2003, p. 217 for

an explanation of our criteria). On one end of the scale, course 5 had the students invest twice the time we thought required. This may have been caused by the fact the quality of the peer assessment would influence the final grade. On the other extreme of the range, 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 prioritizing the input from teacher assessment may have caused non-performance of 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 to allow the students more time for producing a good first draft. The teacher wanted the students at least to be able to process her feedback. As a result, the deadlines for peer and teacher feedback came to coincide. Peer assessment was deemed only a second priority.

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

Revision of the writing products

Almost all students used at least some elements of the peer feedback for revision. On average, students processed about one third of all suggestions, which mainly dealt with content and style. Remarkably, those parts of the text which had not received any feedback were hardly revised, if at all. To what extent students did process suggestions for revision from their peers is presented in Table 2.

There are differences between courses as to that amount and type of revisions actually made. The absence of structural revisions in course 6 may have been caused by the prescribed themes to be covered, which had been presented in considerable detail, more so than in the other courses. In fact, feedback generally was processed less in this course, than in the other courses. This may be explained by the absence of a plenary postmortem on the oral feedback phase, which would have enabled the students to discuss the questions and differences of opinion presented by the feedback group. Thus, peer feedback was left suspended. On the other hand, more than half of the suggestions for revision were processed in course 4a, which may be explained by the fact that this course produced less completed drafts

than the other courses, which left writers with more room to complete their work on the basis of their fellow students' remarks.

=Table 2=

Student grades

Courses 4, 6 and 7 provided the opportunity to compare the grades for the final products of the PA group members with those of the non-PA groups. We found no significant differences (α= 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 by the same criteria. In both cases the grades for the final products were significantly higher than those for the drafts (course 3: 7.3 vs. 6.8 with t=3.3, df=9, p=0.01; course 5: 6.5 vs. 5.7 with t=4,3, df=11, p=0.001). Correlation is high (r= 0.92 in course 3, r= 0.88 in course 5), which means all participants made progress.

In course 5 students revised their draft mainly on base of peer feedback. In course 3 the teacher provided much detailed feedback, more detailed in fact than intended by the peer assessment procedure.

Moreover, as the deadlines for teacher and peer feedback came to coincide, students understandably discarded peer feedback.

All in all, the conclusions about peer assessment and students' grades are at best ambiguous. We found no significant difference between the PA groups and the non-PA groups. Also, other effects, such as teacher feedback, blur the results with respect to the progress from draft to final product

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 discussion in most classes, as according to the assessment form, the criteria were clearly formulated and known to all.

When asked whether peer assessment resulted in better writing, teachers' answers varied. Those of courses 4, 6 and 7 had not

observed any differences between the (final) products of their peer assessment groups and those of the control groups. Those of courses 1, 2 and 5, having assessed and commented upon both drafts and final products agreed the latter to be of improved quality, but were not sure whether this resulted from peer assessment or 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 with peer assessment, for several reasons. She found the writings of the latter to be lacking in structure, as for example the main problem had not been clearly stated, the research questions did not seem logical, the conclusion did not constitute an answer to the problem. The teacher of course 2b observed that the students were more attentive to the process of writing, she thought thanks to students having commented on the work of their fellow students in different stages of completion.

Students' evaluation of peer assessment

This section presents the students' perceived progress of their written product versions, and of their writing skills. Most students were of the opinion the revised version was better than the draft and that this was

the result of their having processed peer feedback. Some explained they had improved their style of writing, in particular the grammar, others said they had re-structured 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 of courses 1, 4b, and 5 noted considerable progress in their writing, whereas those of courses 2a and 7 saw hardly any progress. Some students of 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 3 presents students' perception of the progress of their writing product. There were some who related their progress not to the peer feedback, but to the plenary discussion upon oral peer feedback, or to the comments of the teacher, for which reason their answers are not included in the table.

=Table 3=

Students felt their writing skills had progressed in some courses, especially courses 1, 2, 4b and 5. There are differences between the

designs (F= 4.9; df= 7.82; p≤ 0.001), with a considerable effect (f= 0.65, see Cohen, 1988). As courses 1, 2 and 5 were more strongly focused on writing than the other courses, such outcomes are not very remarkable. Next to the writing assignments, courses 6 and 7 presented the students with assignments of quite another type, course 7 including a Powerpoint presentation, which proved such a time-consuming task that some found themselves pressed in their writing assignment. Table 4 presents the results of students' perception of the progress of their writing skills.

=Table 4=

As shown in Table 5, only course 6 shows significant differences between the estimations of PA group students and non-PA group students. Although both type of groups perceived the writing to have progressed rather little, the difference may have been caused by the implementation of peer assessment.

=Table 5=

In sum, in terms of students' perception we conclude 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, in courses 1 and 6 they felt their writing skills had improved as a result of peer assessment. On the other hand, in course 7 students thought neither their writing product, nor their writing skills had improved. The progress in course 3 was largely to be the result of teacher feedback.

Conclusion and discussion

The aim of our study was to find effective ways of organizing 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 organization of peer assessment in a course, we developed and implemented seven designs of peer assessment, 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 students' and teachers' appraisal of the

perceived progress of the written products and students' writing skills. We will now first list the main results.

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, in a lesser degree to the content of the draft. If students ever commented on structure, the writer would rarely use 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. All teachers appreciated peer assessment for breaking with the usual one-to-one communication between student and teacher.

In order 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 type of outcomes, with students investing a considerable amount of time,

processing relatively much peer feedback, an perceiving their final products as better than 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 designing, we may conclude there are three design features which are particularly beneficial to the effects of peer assessment. First, as 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 pass the product to the teacher. Next, as to directionality (7): reciprocal two-way feedback is more easily organized, as it is clear the assessor will in turn become assessee, which makes it easier to exchange products.

Finally, as to the constellation of assessors (12) and assessees (13), the optimal size of feedback groups seems to be three or four, which allows students to compare their fellow members' remarks, and better to determine their relevance. Apart from lacking the opportunity of comparison, 'groups' of two are too vulnerable for several reasons. One might not perform properly, leaving the other demotivated for insufficient supply of feedback There is also a risk of for instance two

'weaker' students ending up with each other without much to offer. This multiple design study aimed to clarify which combination of design features of peer assessment yields the best results. According to Shavelson *et al*., (2003), the strength of design studies lies in their being implemented in real-life settings in order to find out what works in practice. However, the method 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 of peer assessment. Complying with procedures, they still wanted to give more assistance, but found opportunities restricted, as the students had yet to pass their feedback first. Some also found it difficult having to restrict themselves in their feedback. The role of the teacher in a peer assessment system deserves more attention.

The outcomes we studied are limited to short-term gains, as we did not study the metacognitive gains one finally aims for in introducing peer assessment. We want students to interiorize the criteria of

academic writing, becoming able to review their own work, and to provide their peers with constructive feedback. Such shortcomings present us with challenges 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: Jossey-Bass Publishers.

Berg, B.A.M. van den (2003). Peer assessment in universitair onderwijs. Een onderzoek naar bruikbare ontwerpen. *[Peer assessment in university teaching: an exploration of useful designs]*. Doctoral dissertation, University of Utrecht, The Netherlands.

Berge, H.ten, Ramaekers, S. & Pilot, A. (2004). *The design of authentic tasks that promote higher-order learning.* Paper presented at the EARLI-SIG Higher Education/IKIT-conference, June 18-21, 2004.

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 Palmer.

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.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: an expanded sourcebook (2nd ed.).* Thousand Oaks, California: SAGE.

Shavelson, R.J., Phillips, D.C., Towne, L., & Feuer, M.J. (2003). On the Science of Education Design Studies. *Educational Researcher, Vol.32, 1,* 25-28.

Steehouder, M., Jansen, C., Maat, K., Staak, J. van de, & Woudstra, E. (1992). Leren communiceren *[Learning to communicate]* (3e herziene druk ed.). 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.

[Word count including Abstract and References: 5431]