THE TRAINING OF PEER ASSESSMENT SKILLS TO PROMOTE THE
DEVELOPMENT OF REFLECTION SKILLS IN TEACHER EDUCATION

Dominique M.A. Sluijsmans, Saskia Brand-Gruwel,
Jeroen J.G. van Merriënboer, and Theo J. Bastiaens

Open University of the Netherlands, Heerlen, the Netherlands

Effective assessment approaches, based on constructivist views, receive special attention in current innovations in higher education. These assessment approaches promote integration of assessment and instruction, seeing the student as an active person who shares responsibility, reflects, collaborates and conducts a continuous dialogue with the teacher. Assessment has no longer purely the function of crediting students with recognized certificates but is above all valuable for the monitoring of students' progress and to support them in improving their learning activities. The emphasis shifts to a representation of assessment as a tool for learning (Arter, 1996; Boud, 1990, 1995; Dochy & McDowell, 1997).

Assessment is an important issue in the current developments towards more student-centered learning. Students are more responsible for their own learning process and are increasingly regarded as active participants in instructional activities. An assessment approach has to be chosen that is in alignment with the learning goals of students. To realize the implementation of assessment as a learning tool, a number of changes are desirable on different levels in the organization of institutions in higher education - such as the level of the student, the level of the teachers, and the management level. In this article the level of student teachers is addressed. Specifically, the role of student teachers as assessors of their own work and that of peers is investigated.

There is an increasing demand for self and peer assessments in teacher training colleges, because these forms of assessment fit in well with the latest view on the education
of student teachers. In this view, it is strongly supported that student teachers debate with peers about required teaching skills and their implications in real-life class situations. These discussions are then based on personal teaching and learning experiences. Using students as assessors may provide teachers in teacher training colleges with a valuable strategy for effectively and efficiently implementing student involvement in assessment, and hence it is even more important to investigate the instructional benefits that students gain as assessors (Bangert, 1995). But how can students be involved in assessment and what are the effects on learning outcomes and students' views on instruction and assessment?

Despite many efforts towards student-centered education it is still not a natural practice to give students the role of assessor or designer of their own education. Although all assessment should be in partnership between students and teachers (Stefani, 1998), the question is how this partnership should be developed in educational practice.

To get more insight in how self and peer assessment should be implemented in higher education, a number of studies were analyzed (Sluijsmans, Dochy, & Moerkerke, 1999). Peer assessment practices appear to be mainly focused on peer ranking (each group member ranks all of the others on one of more factors), peer nomination (each group member nominates the highest in the group on a particular dimension of performance), or peer rating (each group member rating each other group member on a set of performance or personal characteristics, using one of several kinds of rating scales) (Kane & Lawler, 1978). The majority of the studies describe a rather quantitative approach in peer assessment, in which tutor scores are compared with student scores (e.g., Topping, 1998). Although peer-tutor correlations show promising results, students also found it difficult to criticize their friends. Hanrahan and Isaacs (2001) mention the students' lack of enthusiasm for peer assessment in their university courses. Some quotations of students illustrate this reaction: "You don't want to mark a fellow student too harshly", or "I feel uncomfortable evaluating another student's paper". Investigators actually found increased opposition to peer assessment after student exposure to it (Rushton, Ramsey, & Rada, 1993). Grading especially concerns students and is seen as "risky and unfair" (Kwan & Leung, 1996).

Results of the literature review and a pilot study (Sluijsmans, Moerkerke, Dochy, & Van Merriënboer, 2001) led to the formulation of three basic assumptions that could be supportive for a more structural implementation of self and peer assessment.

The first assumption is that conducting a self or peer assessment is a complex skill which involves more than giving scores to peers. Before they are put into the role of assessor, students must understand which skills are involved in judging of themselves or a peer. Students need explicit training in assessment techniques to make reliable and acceptable assessment reports (Boud, 1990; Hanrahan & Isaacs, 2001).

To design such training, the skill of making assessments had to be decomposed. Based on results of literature review and interviews with assessment experts, a model was designed in which the necessary skills for conducting a reliable assessment were elaborated (Sluijsmans, et al., 1999; Sluijsmans & Van Merriënboer, 2000). In this model, three main skills are taken into account, namely: (1) defining assessment criteria; (2) giving feedback for future learning; and (3) writing a qualitative assessment report. In this model self and peer assessment are interpreted as learning activities, instead of scoring or ranking tools, which is the case in most self and peer assessment studies (see Falchikov & Goldfinch, 2001; Sluijsmans, et al., 1999).
A second assumption is that the training of self and peer assessment might positively affect the development of content-related skills if the training is embedded in the existing course material (Mehrens, Popham, & Ryan, 1998). In this view, the assessment skill is not trained in isolation but is directly linked to course content. If a teacher for example integrates a training of the assessment skill "defining criteria" in his course on presentation skills, students will learn to negotiate about criteria for a good presentation. Understanding these criteria helps the students to improve their own presentations, and thus the assessment training will support students' development of their presentation skills (Boud, 1995; Orsmond, Merry, & Reiling, 1996, 1997, 2000).

The last assumption states that developed reflection skills are a prerequisite for conducting reliable self and peer assessments. To assess their own work or that of a peer, students have to employ reflection skills to recognise strengths and weaknesses. In this view self and peer assessment are strongly related: analyzing the work of peers can lead to an improved awareness of the quality of one's own work (Falchikov, 1995; Freeman, 1995).

Reflection is already a familiar concept in teacher training (Korthagen & Wubbels, 2000; Krcmer-Ilayon & Tillema, 1999; Newman, 1996; Richert, 1999). The literature on reflection in teacher training is exhaustive but it is a widely accepted truth that it is important to actively and carefully examine one's thoughts in order to improve performance (Reilly Freese, 1999). Reflection in teacher training is mostly linked to actual teaching performance (e.g., Anderson & Freiberg, 1995). Schön (1987) in this context distinguishes between "reflection on action", which refers to thinking about a lesson before and after, and "reflection in action", which refers to the thinking that occurs during the lesson. Loughran (1996) developed a three-part conceptual framework concerning reflection on teaching (reflection during the planning of a lesson, during the actual teaching, after the teaching). Another interpretation is the spiral model developed by Korthagen (1985), in which five phases are distinguished: action, looking at or looking back, awareness of essential aspects, creation of alternative solutions or methods of action, and trial. The fifth phase forms the first phase of a new cycle.

A specific form of a reflective activity that is generally a recurring activity of student teachers, is the writing of reflection papers after a course in which the actual teaching is just one aspect. Teacher educators have to both structure the way student teachers may write their reports on practical experiences as well as create an atmosphere of safety in which student teachers are willing to "open up" and write about their strengths and weaknesses. Korthagen (2001) states that student teachers need to understand the principles of writing reflection papers on a meta-cognitive level, because this helps them to monitor their own progress in reflecting through writing. Students however are not explicitly guided in the process of writing such papers and the implication of reflecting on one's own learning process.

For this study, a teacher training college in the Netherlands was selected. A training in assessment was embedded in first-year courses. The basis for selecting these specific courses was that the teachers of these courses had been using reflection papers for several years. This is in line with the third assumption, that good reflection skills are necessary to become a reliable assessor. Teachers at teacher training colleges, however, often express a desire to systematically guide students in the process of writing good reflection papers. Students gradually exhibit routine behaviors in their writing, so actual learning is difficult.
to investigate. In other words, teachers doubt the level of students' reflection skills. To combine the assessment training with the development of reflection skills, it was decided to regard the writing of reflection papers as the content-related skill. Thus, according to the model, students were trained in defining and applying appropriate criteria for a good reflection paper, how to give feedback on a peer's reflection paper, and how to write an assessment report of such a reflection paper.

Three existing first-year courses on mathematics were redefined to make them suitable for the assessment training. Two research questions were chosen for in-depth exploration. The first research question focuses on the effects of the integrated training in assessment on two types of outcomes: (1) the quality of the assessment skill, which is operationalised in three subskills: using appropriate criteria, giving feedback, and writing an assessment report; and (2) the quality of written reflection papers. The hypothesis is that the training improves both outcomes.

The second research question explores students' perceptions regarding several aspects of assessment and mathematics, such as mathematics self-esteem and test anxiety, before and after training. Trujillo and Hadfield (1999) concluded that there are nearly no studies that examine interventions that influence self-esteem and test anxiety. It is hypothesized that receiving feedback and writing reflection papers positively changes students' views towards mathematics and that their anxiety decreases.

Method

Participants

The sample in this study consisted of 110 first-year students of a primary teacher training college in the Netherlands (15 male, 95 female) with an average age of 19.4 years. The group of students was heterogeneous in educational background with 79 students entering from secondary education, 21 after vocational training for 16-18 year-olds, 6 from another area of higher education and 4 after having worked in a different setting.

Materials

Training Program

In the present study three successive courses on mathematics were selected. These courses lasted six weeks each and were instructed during a period of seven months (November 2000-May 2001). In the three courses students were confronted with some basic skills for mathematics' teaching. Students wrote a personal reflection paper after each course. An assessment training was embedded in each of the three courses. In the first course four assessment tasks about defining criteria for a reflection paper were integrated while in the second course two assessment tasks focused on giving feedback to a peer. In the third course two tasks for writing a structured assessment report about a reflection paper were embedded in the regular course material. These assessment tasks were instructed by the mathematics teacher and the teaching methods used were negotiation, discussion and elaboration. Students reported their feedback to the peers in organized feedback sessions.
**Peer Assessment Form**

The students had to assess a reflection paper by one peer on a peer assessment form on five occasions: in the intake session, at the end of each course, and in the outtake session. This was a blank form after the first and second courses. Developing a structured form in group discussions with the teacher was part of the training in the third course. Students could apply the structured form after this part of the training. A copy of the peer assessment form was given to the assessed peer after each course. A second copy was used by the researcher to investigate the effects of the training on the assessment skill. Students received no grades for their peer assessment forms.

**Rating Form**

A rating form with a number of dichotomous items was developed to analyze the quality of the peer assessments that were written by the students and to find an answer on the first research question if training leads to an improvement in the assessment skill. A prior study was used for the development of this rating form (Sluijsmans, Brand-Gruwel, Van Merrienboer, 2002). In this rating form three variables are included that correspond with the three subskills in the peer assessment model. These variables are: (1) the use of criteria, (2) giving feedback, and (3) writing an assessment report. The variable of giving feedback is subsequently operationalised in four variables: positive comments, negative comments, constructive comments, and posing questions. For the first variable, use of criteria, we checked if the student applied the appropriate criteria. Each student could score a maximum of 19, because the students had defined 19 criteria in the first part of the peer assessment training, which focused on the definition of criteria for a reflection paper. The variable giving feedback was measured by counting the number of positive, negative and constructive comments and the number of questions the student used in the peer assessment. Six items in the rating form were included to measure the variable writing an assessment report. One item for example checked if the student started his or her assessment report with a positive comment.

Finally, three dichotomous items were included to obtain some additional information about the effects of the training. One item was only measured at intake and outtake. This item was: Does the student mark the reflection paper as satisfactory or unsatisfactory? The two other items were included to gather information about the testing after each course. These items were: Does the student elaborate on the criteria? and, Does the student mention own learning experiences?

Two research assistants independently scored the peer assessment forms by means of the rating form. These research assistants received a one-day training, in which they had to conduct a number of exercises with the rating form on samples of peer assessment reports. By analyzing the research data, sum scores were compared for each variable by calculating the interrater-reliabilities. These reliabilities were acceptable for all three variables (Cohen's Kappa > .95).
Examinations

To measure the effect of the assessment training on the results of the reflection papers, the marks given by the mathematics teacher on the first and the final reflection paper were analyzed. The scores ranged from 0 to 100. It was decided that a score of 75 was to be given when the student used the criteria. Additional points were assigned if the criteria were elaborated. If aspects of structure were missing, five points per aspect were subtracted. Points were also subtracted when the reflection paper was too superficial.

Student Questionnaire

Before the first course – at the intake– and after the third course – at the outtake – the students filled out a questionnaire about their perceptions on mathematics, instruction and assessment.

Table 1: Variables, Number of Items, Reliability Coefficients and Sample Items in the Twelve Variables of the Student Questionnaire

<table>
<thead>
<tr>
<th>Variable</th>
<th>#</th>
<th>α</th>
<th>Description</th>
<th>Sample items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Role of the mathematics teacher</td>
<td>7</td>
<td>.65</td>
<td>Perception of the activities a mathematics teacher should perform</td>
<td>The mathematics teacher has to stimulate creative thinking</td>
</tr>
<tr>
<td>Self esteem of the student</td>
<td>16</td>
<td>.74</td>
<td>Self-perception on learning mathematics</td>
<td>I'm capable of working independently on assignments</td>
</tr>
<tr>
<td>Perceptions about mathematical didactics</td>
<td>16</td>
<td>.68</td>
<td>Perception about mathematical didactics</td>
<td>Pupils learn mathematics by working in groups of peers</td>
</tr>
<tr>
<td>Mathematics skills</td>
<td>7</td>
<td>.85</td>
<td>Self-perception on mathematical skills</td>
<td>I'm capable of helping peers with their mathematical problems</td>
</tr>
<tr>
<td>Fear of assessment</td>
<td>6</td>
<td>.86</td>
<td>Students' test anxiety</td>
<td>I'm usually very nervous before taking an exam</td>
</tr>
<tr>
<td>Obtrusiveness assessment</td>
<td>5</td>
<td>.60</td>
<td>Perception of the extent to which math exams have been previously made public to them</td>
<td>The questions on an exam have to be made public to students before the exam is taken</td>
</tr>
<tr>
<td>Predictability of assessment</td>
<td>4</td>
<td>.77</td>
<td>Perception of the extent to which mathematics exams are predictable</td>
<td>I know before I take the test whether I will pass or fail</td>
</tr>
<tr>
<td>Involvement in assessment</td>
<td>7</td>
<td>.82</td>
<td>Perception of the extent to which they want to be involved in assessment</td>
<td>I think that students should participate more in the development of assessment criteria</td>
</tr>
<tr>
<td>Group behavior</td>
<td>4</td>
<td>.68</td>
<td>Perception of peer group behavior</td>
<td>I don't like it when students don't make an individual contribution in groups</td>
</tr>
<tr>
<td>Collaborative learning</td>
<td>8</td>
<td>.84</td>
<td>Perception of collaborative learning</td>
<td>I prefer to elaborate on problems with my peers</td>
</tr>
<tr>
<td>Peer assessment</td>
<td>7</td>
<td>.74</td>
<td>Perception of peer assessment</td>
<td>Peer assessment is useful</td>
</tr>
<tr>
<td>Assessment skill</td>
<td>18</td>
<td>.84</td>
<td>Self-perception on assessment skills</td>
<td>I'm able to analyze a peer's work</td>
</tr>
</tbody>
</table>
On the basis of literature review and prior studies (Sluijsmans, et al., 1999; Sluijsmans, et al., 2001) twelve components or variables were defined for which 116 items in total were constructed. These components were related to expectations regarding mathematics, instruction, assessment and the assessment skill. The students had to answer the items on a five point Likert-scale, ranging from I totally disagree to I totally agree. The intake was organized to investigate the students' perceptions on prior mathematics and assessment experiences. The outtake concerned students' perceptions after the three mathematics courses and the peer assessment training. Reliability analyses of each variable showed that 11 items were responsible for a considerable decrease of the reliability coefficient. Further analyses made clear that these items were less suitable or transparent in their formulation and they were therefore removed from the questionnaire. The variables, number of items, reliability coefficients and a sample of items are presented in Table 1.

Teacher Interview, Field Notes and Students' Reflection Papers

Additional qualitative data were collected from three sources to support the findings of the quantitative data in the discussion section. A standardized open-ended interview approach was chosen to analyze the mathematics teacher's reflections regarding the courses, peer assessment, and feedback sessions (Patton, 1990, p. 284). Field notes were recorded by the researcher during the feedback sessions after each course in which the students reported their assessment reports to their peers. Some quotations from the final reflection papers were used to illustrate students' experiences.

Design and Procedure

The study was constructed according to a within-subject repeated-measures design. Students participated in the experiment for a period of seven months; in the experiment they followed three successive mathematics courses which also involved them in peer assessment tasks. In a two-hour intake session that took place a day before the start of the first mathematics course, the students carried out three activities: filling out the student questionnaire, writing a reflection paper about prior experiences in mathematics, and assessing an anonymous reflection paper. This anonymous reflection paper was previously marked as "unsatisfactory" by the mathematics teacher. We decided to select an unsatisfactory reflection paper for this activity because students tend to mark a product that they are unfamiliar with higher than they would mark the same, but familiar, paper. After the intake, all students attended three successive courses on mathematics. These were the first mathematics courses they attended in their teacher training. During each course, the students worked on products that were content-related. At the end of each course, students took a mathematics test. Besides, the students had to write a reflection paper after the first course, which they could improve after the second and third course, to submit a final version two weeks after concluding the last of the three courses. All students received a three-part training in assessment skill during the courses in relation to the reflection paper.
they were producing. The assessment training was directed at three topics: What are important criteria for a reflection paper (four tasks in the first course)? How to give feedback (two tasks in the second course)? and How to write an assessment report (two tasks in the third course)? In the third course, for example, students developed a peer assessment form based on an expert assessment report that was written by the mathematics teacher.

The output of the first part of the training was a list of 19 criteria for a reflection paper. Students agreed, in negotiation with the mathematics teacher, that a good reflection paper contains for example self-criticism, field work experiences, personal expectations and strengths/weaknesses.

In the second part of the training, which was integrated in the second course, students developed some guidelines for giving feedback. One guideline students agreed on was that it would be positive for a peer to mention his/her own learning experiences in the assessment report.

In the third and last part of the training, which was embedded in the third course, students worked on a peer assessment form and decided what is important in the writing of an assessment report. An expert assessment report served as an example.

Students were instructed that the criteria, feedback rules and structure guidelines derived from the peer assessment training could be helpful in writing the reflection papers and the peer assessment. At the conclusion of each course, the students had to send their reflection paper to the other students. This was done using the facilities of Blackboard, an electronic learning platform. Each student had to assess the reflection paper of another student according to a public "who assesses who?" scheme determined by the mathematics teacher. This scheme altered after each course so that every student had to assess, and was assessed by, different peers. The students had to pick up the reflection paper of their assigned peer from the Blackboard platform. This procedure made the students interdependent, because a student could not write an assessment report if the paper was not sent in. The students wrote their assessment report at home at their own pace. After each course, a feedback session was organized, chaired by the mathematics teacher. In these sessions, in which a group of ten to twelve students participated, each student had to present orally his or her assessment report. The written report was given to the assessed student after the feedback session. The students used the feedback of the peers to rewrite and improve their reflection paper. The student feedback can be regarded as the formative assessment of the papers. To decrease test anxiety and to lengthen the period of training the peer assessment skills, students received no grades from the mathematics teacher for their reflection paper after each course. The role of the teacher was limited to coaching and chairing in the feedback sessions. The reflection paper that was written based on the peer feedback given after each course and had to be sent in two weeks after the last feedback session, was used for the final grade given by the mathematics teacher.

After the third feedback session, an outtake session took place, similar to the intake. In this session, all students filled out the student questionnaire again. They also wrote an assessment report of the reflection paper that was also presented in the intake session.
Data-analyses

Two research assistants analyzed 550 peer assessment forms collected at five points in time during testing, 220 forms from the intake and outtake and 330 collected after completion of each course. These research assistants had prior experience with the use of rating forms. First, means and standard deviations were calculated for each variable of the rating form for the peer assessments written in the intake and the outtake. The data of the peer assessments written by the students in the intake and outtake were analyzed separately from the peer assessment forms written after the three courses. The reason for this was that the conditions in which the students wrote these peer assessments were not similar to the conditions of the other three testing occasions or "moments". Students had less time to write the report in the intake and outtake (half an hour), while they had the opportunity to write the report after each course at their own pace, at home.

A non-parametric test for two related samples (Wilcoxon Signed Ranks Test) was applied to detect significant effects. Frequencies were calculated for the three dichotomous items. Chi-square tests (McNemar) were conducted to retrieve significant effects.

The scores of each variable of the peer assessments gathered after each course (n = 330) were analyzed with an within-subjects analysis of variance with repeated measures on the factor time of testing. Univariate analyses of variance were applied to identify the significant effects after each training part. Paired sample t-tests were conducted to identify differences between the product learning outcomes of the reflection papers after the intake and outtake.

Means and standard deviations were calculated for the twelve variables of the student questionnaire of the intake and outtake. Paired sample t-tests were applied to identify significant differences between the intake and the outtake.

Results

Research Question I: The Effects of the Assessment Training

Effects on Assessment Skill

The peer assessment reports of the intake and outtake were analyzed to see if any overall effects occurred. These reports were analyzed separately from the three intermediate testing moments because the conditions under which they were written were different. In Table 2 the means and standard deviations of each variable of the intake and outtake are shown. Because of a high variance, a non-parametric test for two related samples (Wilcoxon Signed Ranks Test) was applied. This analysis revealed that the training had a positive effect on all variables. It for example appears that in the outtake 87 students used more criteria compared to the intake. The small significance values (p < .001) indicate that each pair of variables differed in distribution.
Table 2: Negative and Positive Ranks, and \( z \)-values of Assessment Forms after Intake and Outtake

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean Rank</th>
<th>( z )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use of criteria&lt;sup&gt;a&lt;/sup&gt;</td>
<td>10.50</td>
<td>-8.15*</td>
</tr>
<tr>
<td>Giving feedback</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Positive comments</td>
<td>23</td>
<td>-3.77*</td>
</tr>
<tr>
<td>- Negative comments</td>
<td>12</td>
<td>-6.37*</td>
</tr>
<tr>
<td>- Constructive comments</td>
<td>73</td>
<td>-7.25*</td>
</tr>
<tr>
<td>- Posed questions</td>
<td>9</td>
<td>-4.26*</td>
</tr>
<tr>
<td>Writing an assessment report&lt;sup&gt;b&lt;/sup&gt;</td>
<td>3</td>
<td>-8.77*</td>
</tr>
</tbody>
</table>

<sup>a</sup>maximum score = 19

<sup>b</sup>maximum score = 6

Table 3: Means and Standard Deviations of the Variables of the Rating Form After Each Course and Mean Square Errors, Degrees of Freedom and F-values for the Main Effect on Time of Testing

<table>
<thead>
<tr>
<th>Variable</th>
<th>Course I</th>
<th>Course II</th>
<th>Course III</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean SD</td>
<td>Mean SD</td>
<td>Mean SD</td>
</tr>
<tr>
<td>Use of criteria&lt;sup&gt;a&lt;/sup&gt;</td>
<td>9.25</td>
<td>6.35</td>
<td>10.45</td>
</tr>
<tr>
<td>Giving feedback</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Positive comments</td>
<td>2.75</td>
<td>3.31</td>
<td>6.08</td>
</tr>
<tr>
<td>- Negative comments</td>
<td>4.28</td>
<td>2.50</td>
<td>4.04</td>
</tr>
<tr>
<td>- Constructive comments</td>
<td>2.05</td>
<td>2.28</td>
<td>2.96</td>
</tr>
<tr>
<td>- Posed questions</td>
<td>1.71</td>
<td>1.92</td>
<td>3.56</td>
</tr>
<tr>
<td>Writing an assessment report&lt;sup&gt;b&lt;/sup&gt;</td>
<td>2.07</td>
<td>2.55</td>
<td>4.19</td>
</tr>
</tbody>
</table>

<sup>a</sup>p < .001

<sup>b</sup>maximum score = 19

<sup>b</sup>maximum score = 6
Results of the calculated frequencies on the dichotomous item "Does the student mark the reflection paper as satisfactory or unsatisfactory?" show that 28.2% of the students marked the paper as unsatisfactory in the intake versus 66.4% in the outtake. Results of the Chi-square tests (McNemar) reveal that this difference is significant ($p=.000$). The students are more negative about the anonymous example of the reflection paper in the outtake than in the intake.

In Table 3 the means and standard deviations after each course are reported for the variables that showed significant results in Table 2 from intake to outtake. Significant main effects were found on all variables for the factor Time of Testing. Univariate ANOVAs reveal which increase or decrease of the means presented in Table 2 was significant. We tested in advance if the conditions for this analysis were met (Stevens, 1996). To correct for a violation of sphericity, the numerator and the denominator degrees of freedom were multiplied by the Greenhouse-Geisser Epsilon.

**Defining Criteria**

On the variable *use of criteria* a significant decrease in the number of criteria used can be observed from the first course ($M = 9.25$) to the second course which was dedicated to the subject of feedback ($M = 6.35$), $F(1,102) = 29.59$, $MSE = 31.13$, $p < .001$). The use of criteria increases again from the second course to the third course ($M = 10.45$), in which students were trained in writing an assessment report ($F(1,102) = 45.84$, $MSE = 39.71$, $p < .001$)).

**Giving Feedback**

The number of positive comments increases significantly from the second ($M = 3.31$) to the third course ($M = 6.08$), $F(1,105) = 54.00$, $MSE = 15.72$, $p < .001$). On the variable *number of negative comments* a significant decrease is observed from the first course ($M = 4.28$) to the second course on feedback ($M = 2.50$), $F(1,104) = 26.19$, $MSE = 14.84$, $p < .001$). The number of negative comments increases again from the second to the third course ($M = 4.04$), in which students were trained in giving feedback, $F(1,104) = 18.77$, $MSE = 13.48$, $p < .001$). The number of constructive comments rises significantly from the first course ($M = 2.05$) to training in giving feedback ($M = 2.28$), $F(1,105) = 30.21$, $MSE = 5.04$, $p < .001$). There is a significant increase of the number of posed questions from the second course ($M = 1.92$) to the third course ($M = 3.56$), $F(1,105) = 10.64$, $MSE = 12.34$, $p < .01$).

**Writing an Assessment Report**

*Structure* increases significantly from the first ($M=2.07$) to the second course ($M = 2.55$), $F(1,105) = 13.72$, $MSE = 1.71$, $p < .001$), and after the third course ($M = 4.19$), $F(1,105) = 116.42$, $MSE = 2.42$, $p < .001$).
Figure 1: (a) Development in Defining of Criteria, (b) Development in Giving Feedback, (c) Development in Writing a Structured Assessment Report
The patterns over time of the six variables which represent the three skills are visualized in Figure 1a, b, and c. Figure 1a illustrates how students develop in the skill use of criteria. The changes of the four variables related to giving feedback are presented in Figure 1b. Figure 1c indicates how students acquire the skill of writing an assessment report.

Frequencies were calculated for the two remaining dichotomous items of the rating form. On the item "Does the student elaborate on the criteria?" it was found that 63.6% did this after the training dedicated to defining criteria (first course), and that this percentage dropped, after the second course, to 28.2%. Chi-square tests (McNemar) reveal that this decrease was significant ($p = .000$). The percentage of students that elaborate on the criteria rose again significantly after the third course to 69.1% ($p = .000$). On the item "Does the student mention own learning experiences?", a confirming answer could be given in 3.6% of the cases after the first course, 41.8% after the second course ($p = .000$), and 56.4% after the third course ($p = .036$).

**Effect on the Reflection Papers**

The reflection paper students wrote in the intake session and the final reflection paper were marked by the mathematics teacher. The average score of their final reflection papers was 70.77 ($SD = 7.44$) compared to an average score of the reflection papers written in the intake session of 46.80 ($SD = 13.19$), $t(1.96) = -16.59, p < .001$. Thus, the training and peer feedback did lead to better reflection papers.

**Table 4**: Means and Standard Deviations of the Students Questionnaire Results at the Intake and Outtake on a five-point Likert-scale (N=110)

<table>
<thead>
<tr>
<th></th>
<th>Intake</th>
<th>Outtake</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Role of the mathematics teacher</td>
<td>$M = 3.32$</td>
<td>$M = 3.59$</td>
<td>$-4.76^*$</td>
</tr>
<tr>
<td></td>
<td>$SD = .46$</td>
<td>$SD = .47$</td>
<td></td>
</tr>
<tr>
<td>Self esteem of the student</td>
<td>$M = 2.30$</td>
<td>$M = 2.41$</td>
<td>$-1.87$</td>
</tr>
<tr>
<td></td>
<td>$SD = .48$</td>
<td>$SD = .49$</td>
<td></td>
</tr>
<tr>
<td>Vision on mathematical didactics</td>
<td>$M = 2.35$</td>
<td>$M = 3.14$</td>
<td>$-13.38^*$</td>
</tr>
<tr>
<td></td>
<td>$SD = .40$</td>
<td>$SD = .48$</td>
<td></td>
</tr>
<tr>
<td>Mathematics skills</td>
<td>$M = 3.33$</td>
<td>$M = 3.59$</td>
<td>$-2.58^*$</td>
</tr>
<tr>
<td></td>
<td>$SD = .84$</td>
<td>$SD = .74$</td>
<td></td>
</tr>
<tr>
<td>Fear for assessment</td>
<td>$M = 2.98$</td>
<td>$M = 2.58$</td>
<td>$3.80^*$</td>
</tr>
<tr>
<td></td>
<td>$SD = .88$</td>
<td>$SD = .84$</td>
<td></td>
</tr>
<tr>
<td>Obtrusiveness assessment</td>
<td>$M = 3.35$</td>
<td>$M = 3.56$</td>
<td>$-1.70$</td>
</tr>
<tr>
<td></td>
<td>$SD = .67$</td>
<td>$SD = 1.21$</td>
<td></td>
</tr>
<tr>
<td>Predictivity of assessment</td>
<td>$M = 2.73$</td>
<td>$M = 3.04$</td>
<td>$-2.77^*$</td>
</tr>
<tr>
<td></td>
<td>$SD = .85$</td>
<td>$SD = .90$</td>
<td></td>
</tr>
<tr>
<td>Involvement in assessment</td>
<td>$M = 3.14$</td>
<td>$M = 3.26$</td>
<td>$-1.29$</td>
</tr>
<tr>
<td></td>
<td>$SD = .74$</td>
<td>$SD = .68$</td>
<td></td>
</tr>
<tr>
<td>Group behavior</td>
<td>$M = 4.22$</td>
<td>$M = 4.37$</td>
<td>$-2.42$</td>
</tr>
<tr>
<td></td>
<td>$SD = .52$</td>
<td>$SD = .48$</td>
<td></td>
</tr>
<tr>
<td>Collaborative learning</td>
<td>$M = 3.40$</td>
<td>$M = 3.25$</td>
<td>$1.67$</td>
</tr>
<tr>
<td></td>
<td>$SD = .72$</td>
<td>$SD = .67$</td>
<td></td>
</tr>
<tr>
<td>Peer assessment</td>
<td>$M = 4.23$</td>
<td>$M = 4.27$</td>
<td>$-1.77$</td>
</tr>
<tr>
<td></td>
<td>$SD = .42$</td>
<td>$SD = .42$</td>
<td></td>
</tr>
<tr>
<td>Assessment skill</td>
<td>$M = 3.62$</td>
<td>$M = 3.87$</td>
<td>$-4.61^*$</td>
</tr>
<tr>
<td></td>
<td>$SD = .43$</td>
<td>$SD = .41$</td>
<td></td>
</tr>
</tbody>
</table>

$^*p < .001$
Research Question II: Change in Perceptions

In Table 4 the means and standard deviations for the twelve variables of the student questionnaire are presented.

Results of paired-sample t-tests show that the students changed in their perception on ten of the twelve variables after the training. After the training, they were more positive about the role of the teacher and had a more positive attitude to mathematics and didactics, higher self-esteem as well as feeling more skilled in mathematics. Their fear of assessment decreased significantly, they evaluated the assessment as more open and predictive. Student's perceptions towards the group's behavior changed positively. They also indicated that they were more capable in assessment than before the training.

However, the students were more negative towards collaborative learning.

Conclusion and Discussion

This study was conducted according to a longitudinal within-subjects design in order to investigate the development of students' assessment skills and their skills in writing reflection papers as an effect of assessment training that was integrated in three mathematics courses. The study also measured how students' views on several aspects of mathematics and assessment changed in the course of the training period. It was expected that students would develop their assessment skill and their skill to write reflection papers as a result of the training. Below the results found will be summarized and discussed. The findings of the teacher interview, field notes and observations of the feedback sessions, and some students' quotations from the final reflection papers will be introduced to substantiate the quantitative data.

The most interesting part in this study was the investigation of the effects of training on the assessment skill, because this was the main goal of the assessment training. Analysis of the peer assessments from the intake and outtake data revealed significant effects. This progress was significant for most variables. The time to write an assessment in the outtake session was very limited, while students had the opportunity to write the assessments at the testing moments between the intake and outtake students were enabled to write their assessments at home at their own pace. This could be an explanation for the fact that students used only five criteria in the outtake, while a maximum of 19 could be gained. The same was the case for structure, on which only a score of 2.46 out of the six possible points was reached. The decrease that occurs from the testing moment after the third course and the outtake can be contributed to the different settings in which the assessments were written.

Students appeared to be more negative about the same reflection paper in the outtake than in the intake. This indicates a more critical attitude after the training period. While writing the assessment report of the reflection paper which they also assessed in the intake, students made comments like: "This student did not use the criteria that we defined for a reflection paper".

The figures illustrated the students' development on the three subskills in the period between the intake and outtake. The training effects are clearly visible. Students use the
criteria more after the first part of training dedicated to defining criteria. However, this decreased after the second course. An explanation could be that the criteria "faded" for some period, because attention was focused on the second part of the assessment training, which considered giving feedback. In conclusion, students reached a higher level on all skills after the third course than they showed at the time of the first course. Besides having more assessment skills, students also wrote better reflection papers.

Regarding the second research question about the changes in students' perceptions during the training period, positive effects were observed. The results of the questionnaire showed a positive change of perspective towards different aspects of instruction and assessment. Comments from the students in their reflection papers implied that they felt less intimidated about conducting the peer assessment after each course. An important finding is that their test anxiety decreased. Students who developed a negative attitude towards mathematics changed their perceptions through reflection and peer assessment.

Students' attitude towards mathematics often differs from their attitude towards other domains, like for example pedagogy or history. First-year student teachers' perception on mathematics is highly influenced by their experiences in secondary education. While some of these students enjoy mathematics, many have less positive experiences (Ashcraft, Kirk, & Hopko, 1998; Fennema & Sherman, 1978). Student teachers with negative feelings about mathematics are unlikely to change their views (Sullivan, 1989). This is very surprising, since prospective primary school teachers are aware that they have chosen a profession in which teaching mathematics to children is a basic skill (Kelly & Tomhave, 1985; Trujillo & Hadfield, 1999; Watson, 1987). The feedback sessions could have played a helpful role in decreasing mathematics anxiety and increasing mathematical skills.

Another explanation for the observed decrease in test-anxiety could be that students were not immediately assessed by the teacher after the first course. The peer feedback was a form of formative assessment, which was less threatening to the students. Although the teacher chaired the feedback sessions, students felt that they did not have to succeed on the first attempt and that they had opportunities to improve their initial performance. The main bottleneck in structural changes in instruction and assessment is often that students still study to the test and that teachers teach to the test. Madaus (1988) illustrates this phenomenon on the level of curriculum: "It is testing, not the 'official' stated curriculum, that is increasingly determining what is taught, how it is taught, what is learned and how it is learned" (p. 83). When formative assessment is used as a learning and feedback tool, test-driven behavior is reduced. Dalziel (1998) adds to this that it is important that students are informed about the assessment to be used and that they are involved in the decision making about instructional issues. In other words, assessment is broader defined than as just a paper and pencil test.

During the feedback sessions, which took place three times, students conveyed their written assessment reports orally to their designated peer. In the first feedback session, in which the students had to present their assessment report to the rest of the group for the first time, students felt very insecure about conveying their findings to the assessed peer. Some students were very nervous and showed physical signs like sweating, stuttering and blushing. It was obvious that students were not accustomed to giving critical feedback. In the second and third feedback sessions, these symptoms diminished. As one student wrote in his reflection paper: "At first it was weird to give your feedback to another student,
especially when it was negative, but gradually it became easier to do". Students were more able to address their comments directly to the assessed person in the second and third feedback session, and were less dependent on the teacher. It was surprising that all students took their responsibility and wrote extensive assessment reports each time, especially so given that they did not receive grades for their reports. The teacher explained this as follows:

They worked very seriously on their tasks, I don't exactly know why. Colleagues often complain about the attitude of students. Maybe it turned out well because I conveyed a sense of confidence to the students, let them know that their input was worthwhile. Another explanation could be that they were interdependent. If a student did not write an assessment or did not make a constructive contribution in the group, this was noticed and criticized by the other peers. As far as this is concerned, it was nice to observe that my former role now shifted to the students.

Although the majority of the findings supported the hypotheses, some limitations of this study should be mentioned. One important limitation is that in the long training period, the students also carried out a number of other activities that may have contributed to their skill of writing a reflection paper and their assessment skill. A prior study however showed that a short training period leads to only small differences in development between trained groups and control groups (Sluijsmans, et al., 2002). Due to the length of the study, it is possible too that changes in perceptions were caused by other external factors. The students attended other courses in the same period and gained teaching experience at elementary schools.

A second limitation lies in the chosen approach. Peer assessment in this study featured mostly as a learning tool, not as a marking tool. This was however not equally beneficial to every student. A student who, for example, was already skilled in the writing of reflection papers, did not benefit from assessing a paper that was of lesser quality than his own. The opposite also occurred: a student who had little knowledge about the meaning of a reflection paper's criteria, was less capable of giving constructive feedback. Regarding this issue, the mathematics teacher's view was that a student may not benefit from feedback received from peers who have insufficient knowledge to give constructive comments, but the same student will learn by giving feedback to others. The question to what extent writing reliable assessments requires knowledge about the relevant criteria, is an interesting one and deserves further elaboration in future research.

A last comment on the design of this study concerns generalizability. Students learned how to write reflection papers in the mathematics domain but it was not investigated if the findings are generalizable to other subject matter domains. It would be interesting to examine if these students would also write better reflection papers in other domains.

Supporting the development of students' assessment skills can play an important role in the ongoing evolution of teacher education towards skill-based curricula. Results from the present study reveal that students can be trained in assessment skills and that this positively influences the quality of their reflection papers. The peer assessment tasks and
the feedback sessions encouraged the students to explore concerns involved in giving and getting feedback on reflection papers. Giving helpful feedback to others may also develop students' interpersonal skills and assist others in learning; accepting feedback from others improves performance on different skills. The more skilled students become in the peer assessment process, the easier it becomes to seek suggestions and learn from others - in the classroom, on the job, or in other areas of life.

In summary, this study showed promising results concerning the involvement of students in assessment procedures. Both the teacher and the students experienced the importance of developing criteria. The criteria of reflection papers are often not transparent and differ by teacher. That teachers should be capable of writing such papers and that teachers at teacher training colleges should contribute to the development of this skill is accepted (Boud, 1989; Korthagen, 2000; Reilly-Freese, 1999; Kremer-Hayon, & Tillema, 1999), but teachers still have multiple perspectives on the concept of reflection and the effective use of reflective activities. Implementing activities like a peer assessment training helps in making these perspectives explicit.

References


The Authors

DOMINIQUE SLUIJSMANS is an educational technologist at the Educational Technology Expertise Centre of the Open University of the Netherlands. Her main interests are student involvement in assessment, teacher education and curriculum design.
SASKIA BRAND-GRUWEL is an educational technologist at the Educational Technology Expertise Centre of the Open University of the Netherlands. Her areas of specialization are comprehensive reading, higher order skills and competency based education.

JEROEN J.G. VAN MERRIËNBOER is professor and head of the Research Department at the Educational Technology Expertise Center of the Open University of the Netherlands. He specializes in instructional design (ID) for complex learning, computer-based learning environments, and intelligent performance support for ID.

THEO J. BASTIAENS is an associate professor at the Educational Technology Expertise Centre of the Open University of the Netherlands. He has been working in the field of training and development since 1990. His special interest is in human resource development and e-learning.

Correspondence: <dominique.sluijsmans@ou.nl>