Routledge
Taylor & Francis Group

Innovations in Education and Teaching International

ISSN: 1470-3297 (Print) 1470-3300 (Online) Journal homepage: http://www.tandfonline.com/loi/riie20

Training teachers in peer-assessment skills: effects on performance and perceptions

Dominique M. A. Sluijsmans , Saskia Brand-Gruwel , Jeroen J. G. van Merriënboer & Rob L. Martens

To cite this article: Dominique M. A. Sluijsmans , Saskia Brand-Gruwel , Jeroen J. G. van Merriënboer & Rob L. Martens (2004) Training teachers in peer-assessment skills: effects on performance and perceptions, Innovations in Education and Teaching International, 41:1, 59-78, DOI: 10.1080/1470329032000172720

To link to this article: https://doi.org/10.1080/1470329032000172720

Innovations in Education and Teaching International,
Vol. 41, No. 1, February 2004

Routledge
Taylor & Francis Group

Training teachers in peer-assessment skills: effects on performance and perceptions

Dominique M. A. Sluijsmans*, Saskia Brand-Gruwel,
Jeroen J. G. van Merriënboer & Rob L. Martens
Open University of The Netherlands, The Netherlands

This paper focuses on two increasingly important issues in teacher education: the design of more skill-based education and the involvement of students by means of peer assessment. Ninety-three student teachers were trained in one important peer-assessment skill, namely 'defining performance criteria'. This training, which consisted of four peer-assessment tasks, was integrated in an existing course. Half of the group was trained in the skill of 'defining criteria' (experimental groups) and the other half was not (control group). By working on the peer-assessment tasks, student teachers in the experimental group learned to define performance criteria for a course content-related product. The effects of the training on students' ability to define criteria and the effects on the content-related skill were examined. Findings show that the student teachers from the experimental group scored significantly higher on the use of criteria, but did not surpass the control group on the content-related task performance.

Introduction

As is the case in many other countries, there is a growing awareness in The Netherlands that the curricula in higher vocational education should be based on the development and acquisition of skills (Tillema *et al.*, 2000). Skill-based learning is an ongoing issue, especially in the domain of Teacher Education (Kremer-Hayon & Tillema, 1999; Darling-Hammond & Snyder, 2000; James, 2000; Willems *et al.*, 2000). In the last few years, politicians have invested much time in redefining the image of primary school teachers. Instead of placing the primary-school teacher in the role of 'the king in the castle', teachers are encouraged to become a member of a learning organization. A number of Teacher Training Colleges collaboratively formulated a broad scale of skills that student teachers need to develop. The skill requirements of a primary school teacher are reported in a vocational training profile (LPC, 1995), which identifies 41 skills that are categorized into 10 domains. The skills represent the overall accepted knowledge, proficiency and attitudes a primary school teacher needs to

*Corresponding author: Educational Technology Expertise Centre, Open University of The Netherlands, P.O. Box 2960, 6401 DL Heerlen, The Netherlands. Email: dominique.sluijsmans@ou.nl

ISSN 1470–3297 (print)/ISSN 1470–3300 (online)/41/010059–20

DOI: 10.1080/1470329032000172720

acquire. Because these skills are nationally determined and integrated in the curriculum, the risk of educating teachers that set very different standards and values decreases. The goal is to ensure that student teachers meet the criteria of each skill. These criteria have to be the same as those used in the practice setting.

To establish an environment in which student teachers can develop their skills, a change is required on two fronts: in the preparatory (pre-service) education of teachers and in the continuing (in-service) education of those already in the educational profession. Both groups need assistance and support in how to apply skill-based learning. The present study is focused on the first front, the education of student teachers.

The importance of peer assessment in teacher education

Within the scope of training student teachers, the development of a specific skill of the vocational training profile of primary school teachers, namely 'the skill to assess the work of peers', is further elaborated (LPC, 1995). There are three reasons why this skill is important for the domain of teacher education.

First, the importance of communication between teachers in schools has been endorsed by many researchers (Johnson *et al.*, 1992; Cohen, 1994; Sharan & Sharan, 1994; Slavin, 1995). Teachers have to work together, learn from each other and become a member of a learning organization (Verloop & Wubbels, 2000). But within this collaborative and skill-based framework, student teachers have to be provided with procedures, tools and job aids that help them to structure their own working process. One of the main aspects is developing a professional attitude towards the work and ideas of other teachers in the school. This requires training in skills that transcend the basic know-how of a certain content domain. Peer assessment is one skill.

Second, as prospective teachers of children in primary schools, it is advisable to train student teachers in how to make critical judgements about the performance of peers, and later on about the performance of children. The student teachers will be assessors in their own classroom. They will have to design assessments.

A third reason is that after students leave higher education, they are likely to be heavily reliant on the judgement of their peers to estimate how effective their performances in the school are (Brown *et al.*, 1994). Training in peer-assessment skills stimulates this mutual influence to take place at a professional level.

Training in peer-assessment skills

The reasons mentioned above convinced those in the field of teacher education that being able to interpret the work of colleagues and peers is a necessary prerequisite for professional development and for improving one's own functioning (Verloop & Wubbels, 2000). Assessing the work of peers is a skill that needs to be developed (Birenbaum, 1996; Reilly Freese, 1999; Sluijsmans *et al.*, 2001). Students who are novices in assessing are insecure about their ability to assess and indicate that they need more guidance on the marking criteria (Cheng & Warren, 1997; Woolhouse, 1999). The importance of the negotiation about criteria has already been stressed in several studies (Boud, 1995; Orsmond *et al.*, 1996, 1997, 2000). However, there is little known about how teachers try to develop this peer-assessment skill with student teachers. That

teachers should be capable of critical reflection and that teachers at Teacher Training Colleges should contribute to the development of this skill is by now a generally accepted truth (Boud & Falcikov, 1989; Reilly Freese, 1999; Kremer-Hayon & Tillema, 1999; Korthagen & Wubbels, 2000), but training student teachers in assessment skills is an ill-defined area. Teachers are unfamiliar with ways to involve students in the assessment process through peer assessment.

A peer-assessment model

Several authors (Birenbaum, 1996; Fallows & Chandramohan, 2001; Hanrahan & Isaacs, 2001) have recommended training in assessment skills. In order to understand the use of peer assessment in courses and ways to train this type of skill, a peer-assessment model was developed and revised by a number of assessment experts from different countries (Sluijsmans & Van Merriënboer, 2000).

In the peer-assessment model the underlying constituent skills of the complex skill to assess were identified. The model is based on several sources. First, the literature on peer assessment was analysed (Sluijsmans *et al.*, 1999). There seem to be several ways in which students can be involved in assessment on their courses: students can have a role in the choice of assessment tasks, in setting assessment tasks and in discussing assessment criteria. Based on these findings, a first draft of the constituent skills was constructed. Second, literature concerning the integration of assessment and instruction was analysed in relation to the role of the student. In the end, three levels were distinguished in the decomposition of the peer-assessment skill. At the first level, three main skills have been determined. These are (1) defining assessment criteria: thinking about what is required and referring to the product or process; (2) judging the performance of a peer: reflecting upon and identifying the strengths and weaknesses in a peer's product; and (3) providing feedback for future learning: giving constructive feedback about the product of a peer. At the second and third level another 11 constituent skills were defined (see Figure 1). The defined skills are the basis for the design of the training in peer assessment.

Because the peer-assessment skill is too complex to be covered in only one course (Van Merriënboer, 1997), for this study it was decided to train the students in the first main constituent skill: *defining criteria*.

The design of peer-assessment training: two basic assumptions

Designing training in peer assessment is based on two assumptions. The first assumption is that the training of assessment skills might have positive effects on the development of content-related skills, if the training is embedded in the existing course material which is designed according to a performance-based approach (Mehrens *et al.*, 1998). In this view, the assessment skill is not trained as an isolated skill, but is directly linked to course content. If a teacher, for example, integrates training in the assessment skill 'defining criteria' in his or her 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 performance in giving presentations, thus the assessment training will support students' development of their presentation skills. On that line of argument, student teachers will always be guided in at least two skills: the skill to assess the work of peers and a content-related skill, which contains the object

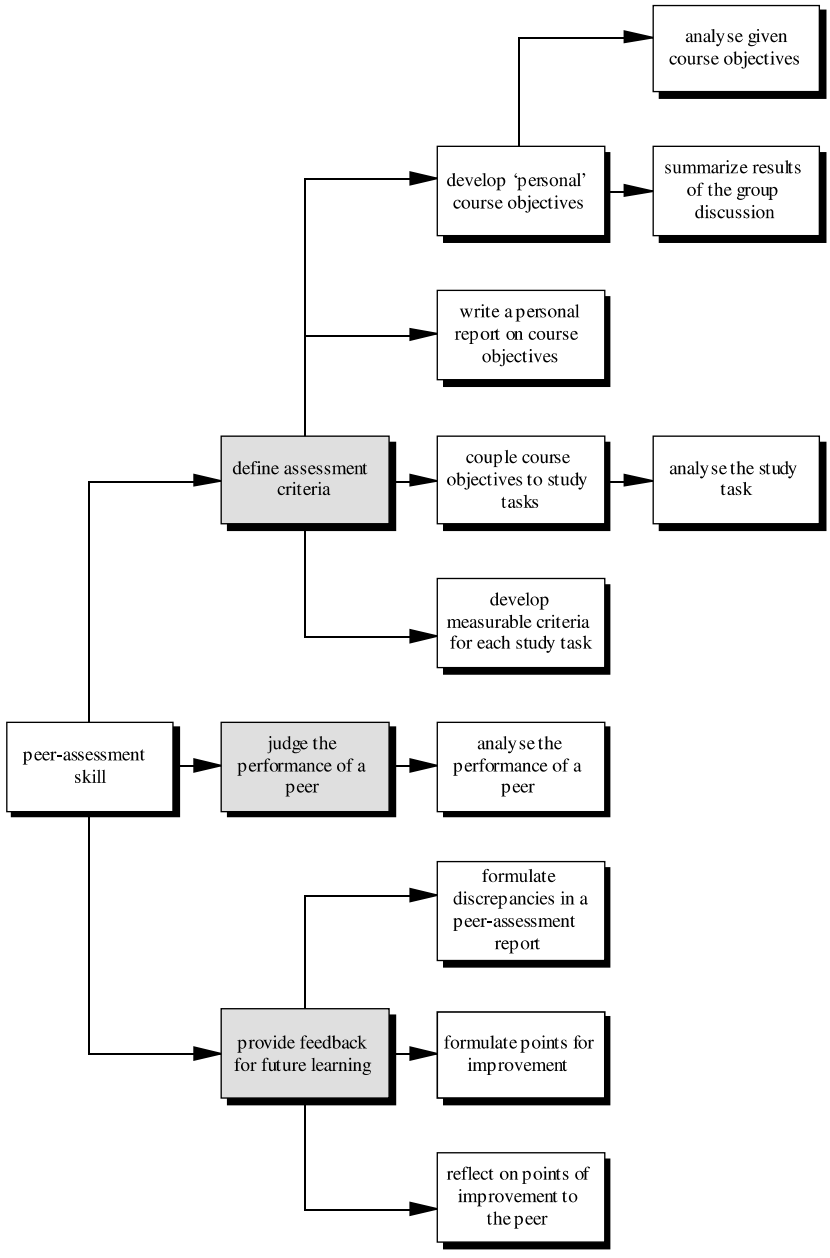

Figure 1. Peer-assessment model

of assessment. This leads to the hypothesis that if student teachers develop their skill to assess the performance of peers, this should also lead to a general improvement in their task performance in the domain of the course. It is assumed that knowing the criteria of a product and observing the work of peers leads to a higher understanding of the quality of one's own work (Falchikov, 1995; Freeman, 1995).

The second assumption is that training students in skills has consequences for the design of the courses. Within the framework of skill-based curriculum design, the educational material is no longer defined from the perspective of the content domain, but from the perspective of the skills (Tillema *et al.*, 2000). This means that skills are taught in the context of different content domains. This simultaneous change in both course design and the role of students is often experienced as being very complex by teachers in higher education, due to the lack of procedures and job aids regarding curriculum design (Verloop & Wubbels, 2000). Courses that are designed from the perspective of isolated content units will be affected by thorough revision, in order to make them skill-based. Ways in which the skills can be developed within existing courses should be considered. In this study, attention was also given to the consequences of such redesign of courses and to potential effects of content domains and/or teachers. We therefore investigated how students responded to a change in course design.

In summary, the work described in this paper served three goals: (1) investigating the effects of peer-assessment training on the development of the assessment skill; (2) the effect of the training on task performance in the domain of the course; and (3) guidelines for designing courses that are suitable for training in professional skills. Based on the presented theoretical framework, the following research questions are elaborated:

1. Does training in peer assessment lead to the development of the skill to assess the work of peers?
2. Does training in peer assessment lead to an improved task performance in the domain of the course?
3. What are the perceptions of students regarding the redesigned course and does the content domain influence these perceptions?
4. What are the perceptions of the teachers in the Teacher Training College about the peer-assessment training and the redesigned course?

Method

Participants

The sample consisted of 93 second-year students from a Teacher Training College in The Netherlands (19 male, 74 female) with an average age of 20.7 years (SD = 1.6). Students were randomly assigned to experimental groups which received peer-assessment training ($n = 43$) and control groups ($n = 50$). The Teacher Training College offers a broad education leading to the qualification to teach every subject taught in primary schools, to pupils in the 4–12 age range. Five teachers of the Teacher Training College participated in this study. Each teacher was responsible for one content domain in the selected course for the study. These domains were pedagogy, physics, mathematics, philosophy and music.

Materials

Course. A second-year course on discovery learning was selected for redesign. The former version of the course was designed from the perspective of the content domain. A problem of this course was that students felt that discovery learning was basically linked to the physics domain, although four other domains were also involved. Another problem was that students

worked on several course objectives that led to a high workload, without thoughtful consideration of why they had to work on specifically those products. To solve these problems, the existing course was redesigned from a skill-based perspective for the purposes of the present study.

After discussion, it was decided that the new course objective was that students were trained in their skill to design a lesson plan on discovery learning in the context of one of the five content domains. In operational terms, at the end of the course students had to deliver a lesson plan that was related to one of the five content domains. Therefore, the 93 student teachers were randomly distributed amongst the pedagogy domain ($n = 20$), the physics domain ($n = 21$), the philosophy domain ($n = 21$), the mathematics domain ($n = 21$) and the music domain ($n = 10$).

Before the design of the concrete study tasks, the involved teachers decomposed the skill of designing a lesson plan on discovery learning similar to the way the skill to assess was analysed (Van Merriënboer, 1997). This resulted in four main sub-skills students had to acquire with regard to the design of a lesson plan for discovery learning: (1) introducing a problem in a classroom with pupils; (2) posing the right questions to the pupils in relation to the introduced problem; (3) analysing the problem with pupils, and (4) solving the problem with pupils. A study task was designed for each of the four skills in each of the five content domains.

The whole course enclosed six classes of an hour and a half each in a period of four weeks: an introductory class, four regular course classes and one class in which the students peer-assessed the end-product of peers. In the four regular classes, the content-related study tasks regarding discovery learning were taught, based on the four skills. For example, the study task for the physics groups focused on introducing, questioning, analysing and solving a *physics* problem in a classroom with pupils. For the mathematics groups, the study tasks focused on introducing, questioning, analysing and solving a *mathematical* problem.

Peer-assessment training. In this study, students were allocated to control and experimental groups. The teachers of the domains (pedagogy, physics, philosophy and mathematics) taught both a control group and an experimental group; the music teacher taught only a control group. This meant that in total there were nine groups of students, four experimental groups (three groups of 11 students, one of 10 students; $n = 43$) and five control groups (groups of 10 students each; $n = 50$).

Based on the redesigned course in which the course objective and content-related skills were defined, a peer-assessment training for only the *experimental* groups was developed. This training consisted of four so-called peer-assessment tasks, which were derived from the skill 'defining criteria'. In the four peer-assessment tasks that were embedded in the four course classes of the course 'Designing Discovery Learning Lesson Plans', students had to define measurable criteria that were related to each of the four skills for designing a discovery learning lesson plan. For this, the teacher presented examples of valid and invalid criteria. Each peer-assessment task was characterized by interactive discussions between the students to foster collaborative learning and focused on the skills that are related for defining criteria. Students were encouraged to think about 'personal' course objectives and the relation between course objectives and the study tasks (see Figure 1). Table 1 shows how the peer-assessment tasks are embedded in the regular study tasks.

Peer-assessment form. At the end of the course, all students had to assess the lesson plan on discovery learning of four peer dyads on a blank peer-assessment form.

Table 1. The peer-assessment tasks embedded in the study tasks

Classes (followed by the control groups and the experimental groups)	Embedded peer-assessment task (followed by the experimental groups)
1 Introductory class	—
2 Introducing a problem	Defining criteria for introducing a problem to pupils in the classroom
3 Posing the right questions related to the problem	Defining criteria for posing good questions to pupils
4 Analysing the problem	Defining criteria for the analysis of a problem with pupils
5 Solving the problem	Defining criteria for an adequate solution with pupils
6 Presentation of end-products and peer assessment	—

Rating form. To analyse the quality of the peer assessments that were written by the students, a rating form was developed. It was decided that the following eight variables—deduced from the output of the peer-assessment tasks—were important to determine the quality of the peer assessments: the use of criteria, naive word use, consequent structure, being critical, giving a conclusion, posing questions, giving a mark and giving points for improvement. For the first variable 'use of criteria', the 10 criteria developed by the students for well-designed discovery learning lesson plans were included in the rating form. Research assistants scored the valid criteria with one point. Because each student wrote four peer assessments, the maximum score that could be gained for this variable was 40. For the other seven variables a maximum score of four could be gained per variable, because each variable consisted of only one item (e.g. if the student gave a conclusion, one point was given). The maximum score that could be gained for these seven variables was 28. In total, students could gain 68 points for their peer assessments.

Although the scores on the variable 'use of criteria' were particularly important, because this skill was taught in the peer-assessment tasks, data were gathered for all seven variables because students indirectly discussed these variables in the peer-assessment tasks.

Three independent research assistants scored the peer-assessment forms using the rating criteria. For each variable the inter-rater reliabilities were calculated. These reliabilities were acceptable for all variables (Cohen's Kappa > .95).

Examinations. To measure the effect of the peer-assessment training on the performance of students, the marks on the discovery learning lesson plans of the students given by the teacher were analysed. The score could range from 0 to 100.

Student questionnaire and structured student interviews. Before and after the course, the students filled out a questionnaire about their perceptions on instruction and assessment. Ninety-two items were divided among 16 variables. Six variables were related to instruction, five variables were related to vision on instruction and assessment and another five were related to the role of the student in assessment. Because the students worked in smaller groups in the redesigned

course, the variable 'group atmosphere' was added in the post-test and not measured in the pre-test. The students had to answer the items on a five-point Likert scale, varying from 'I totally disagree' to 'I totally agree'. The pre-test was carried out to investigate the students' perceptions of prior courses that were comparable to the course on discovery learning. These prior courses were not designed in a skill-based way. The post-test concerned students' perceptions after the redesigned course. The clusters, variables, number of items, reliability coefficients and example items of the 16 variables are presented in Table 2.

After the course and the peer assessment, 16 students were interviewed (eight from the control group, eight from the experimental group). They had to give their answers on 11 questions about the peer-assessment tasks, the peer assessment and the course in general.

Table 2. Clusters, variables, number of items, reliability coefficients and example items of the 16 variables of the student questionnaire

Variable	No.	α	Example items
Cluster: Instruction			
Satisfaction classes	5	.75	The study tasks evoked interesting discussions
Transparency classes	4	.73	The course objectives were comprehensible
Learning access level	4	.79	I felt that I could distinguish main issues from side issues
Practical relevance	3	.78	The study tasks are practically oriented
Quality of the instruction	7	.80	The goals of the study tasks were instructed very clearly
Teacher involvement	3	.83	The teachers had an open mind for the opinions of the students
Cluster: Vision on instruction and assessment			
Relation instruction and assessment	4	.80	The study tasks and the assessment were interrelated
Assessment behaviour	4	.59	The first thing I do at the start of a course is find out what the assessment is
Fear for assessment	3	.73	I'm usually very nervous before taking an exam
Obtrusiveness assessments	5	.67	The questions on an exam have to be public to students before the exam is taken
Overall vision on assessment	2	.81	I support the way I am assessed
Cluster: Role of student in assessment			
Involvement in assessment	8	.69	I think that students should be more involved in the development of assessment criteria
Group behaviour	5	.64	I don't like it when students don't make an individual contribution to a group product
Collaborative learning	3	.67	I prefer to elaborate on problems with my peers
Assessment skill	18	.87	I'm able to analyse a product of a peer
Group atmosphere[a]	14	.89	I enjoyed working together on a study task as a group

[a] Cronbach's alpha calculated in post-test.

Teacher questionnaire and structured teacher interviews. Each teacher of the Teacher Training College who was involved in the course evaluated the four peer-assessment tasks by means of a short questionnaire. The questions concerned: (1) the invested time; (2) the desired output of each assessment task; and (3) transparency of the tasks. Besides the teacher questionnaire, teachers were asked several questions in an interview. The questions were related to two phases, the design phase of the course and the implementation phase. Regarding the design phase, questions were asked about their experiences with the redesign of the course and their co-operation with other colleagues. Questions relating to the implementation phase concerned the experiences with the instruction of the peer-assessment tasks and their vision on assessment and instruction, and the role of students and themselves.

Design and procedure

The experiment was set up according to a pre-test/post-test control group design. Before the start of the course, the students filled out the student questionnaire as a pre-test. Both the control groups and the experimental groups attended the regular classes. The experimental groups moreover followed the embedded peer-assessment tasks. The time students in the control groups spent on the regular classes was the same as the students in the experimental groups spent on the classes and the peer-assessment tasks together. Thus, the students in the control groups had relatively more time to discuss the content of the regular classes, because they did not receive the peer-assessment training.

In each peer-assessment task, a part of the whole criteria list for a lesson plan was developed (see also Table 1). This was done through constructive discussions guided by the teacher. The students were encouraged by the teacher to make their personal ideas explicit. At the end of the fourth and last peer-assessment task, the students had a list of 10 criteria. During the course, all students worked in dyads on the end-product. At the end of the course the dyads had to present their end-product to the rest of their group. The end-product involved the design of a lesson plan for an elementary school, which was based on the principles of discovery learning. The students designed a lesson plan for the domain they attended.

In the last class of the course, the students in both the control groups and the experimental groups were instructed to write a qualitative peer assessment with regard to the content of the lesson plan of the peer dyads. Each student wrote four peer assessments, because in each group there were four other dyads to assess. After the course, all students filled out the same questionnaire as in the pre-test. The teachers who taught the experimental groups filled out the teacher questionnaire after each peer-assessment task. In the two weeks after the course, the teachers and 16 students were interviewed.

Data analyses

Three independent research assistants analysed the 372 peer-assessment forms (93 students who wrote four assessments). These research assistants were instructed in the application of the rating form. A one-way analysis of variance (ANOVA) with the factor Group was applied to identify differences between the control and experimental groups on the eight variables of the rating form.

A one-way ANOVA with the factor Group was also applied to identify differences between the control and experimental groups on the task performance in the course domain, a lesson plan on discovery learning.

Means and standard deviations were calculated for the 15 variables of the student questionnaire for the control and experimental groups. The scores of each variable were analysed with a 2 (Groups) × 4 (Content Domains) × 2 (Time of Testing) analysis of variance with repeated measures on the last factor. An exception was the analysis of the variable *group atmosphere*, which was only measured in the post-test. A 2 (Groups) × 4 (Content Domains) was done, because it was only measured in the post-test. The students from the music domain were excluded, because the analysis requires data from the domains in which both control and experimental groups are represented.

The answers of the structured student interviews were categorized according to a code system. Frequencies were calculated.

Medians were calculated for the three variables of the teacher questionnaire. Because of the small number of student and teacher interviews, the answers are analysed qualitatively.

Results

Does training in peer assessment lead to the development of the skill to assess the work of peers?

Table 3 presents the means and standard deviations of the eight variables that were measured with the rating form for the experimental and control groups. Students could in total gain 68 points for their peer assessments. The average of the experimental groups was higher (mean = 16.77, SD = 9.65) than the average of the control groups (mean = 12.89, SD = 6.33). The difference between both groups was significant ($F(1,83) = 4.89$, MSE = 63.68, $p < .05$).

Further analyses revealed that the experimental groups scored significantly higher on the variables 'use of criteria', $F(1,83) = 5.73$, MSE = 44.93, $p < .05$; 'consequent structure', $F(1,83) = 5.91$, MSE = 1.18, $p < .05$; and 'giving a mark', $F(1,83) = 4.32$, MSE = 1.26, $p < .05$. A contrary

Table 3. Means and standard deviations of the experimental and control groups on the peer-assessment forms at the post-test

		Experimental groups		Control groups	
Variable	Maximum score	Mean	SD	Mean	SD
Use of criteria*	40	13.95	8.31	10.45	5.05
Naive word use	4	1.32	0.55	1.46	0.43
Consequent structure*	4	0.79	1.42	0.27	0.80
Being critical	4	1.58	1.62	1.90	1.34
Giving a conclusion	4	0.29	0.96	0.49	0.96
Posing questions	4	0.71	1.01	0.59	0.86
Giving a mark*	4	0.66	1.48	0.19	0.81
Giving points for improvement**	4	0.11	0.31	0.46	0.73
Total score*	68	16.77	9.65	12.89	6.33

*$p <$ .05; **$p <$ .01.

effect was found on the variable 'giving points for improvement', where the control groups scored significantly higher than the experimental groups, $F(1,83) = 8.99$, MSE = 0.43, $p < .01$. Overall, the training had the expected effect, because the experimental groups used the criteria significantly more often than the control groups.

Does training lead to better performances/products?

At the end of the course, the students were responsible for one final product, a discovery learning lesson plan. The average score of the experimental groups was 70.31 (SD = 8.22); the average of the control groups was 68.71 (SD = 7.63). The difference between both groups was not significant.

What are the perceptions of students regarding the redesigned course and does the content domain influence these perceptions?

In Table 4 the means and standard deviations of the student questionnaire are given. The scores of each variable were analysed according to a 2 (Groups) × 4 (Content Domains) × 2 (Time of

Table 4. Means and standard deviations of the experimental and control groups' students questionnaire results at the pre-test and post-test on a five-point Likert scale

	Experimental groups				Control groups			
	Pre-test		Post-test		Pre-test		Post-test	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Cluster: Instruction								
Satisfaction classes	3.10	0.51	3.70	0.55	2.99	0.53	3.86	0.53
Transparency classes	3.15	0.63	3.70	0.53	3.28	0.54	3.89	0.56
Learning access level	3.45	0.81	3.75	0.70	3.49	0.80	3.73	0.84
Practical relevance	3.52	0.84	3.99	0.67	3.57	0.69	4.01	0.81
Quality of the instruction	2.81	0.75	3.88	0.61	2.91	0.50	4.02	0.50
Teacher involvement	3.28	0.76	3.97	0.68	3.33	0.67	4.12	0.59
Cluster: Vision on instruction and assessment								
Relation instruction and assessment	2.18	0.72	3.96	0.86	2.37	0.59	3.98	0.65
Assessment behaviour	1.86	0.61	2.46	0.48	1.80	0.59	2.38	0.67
Fear for assessment	2.70	0.91	2.58	0.53	2.88	0.99	2.66	0.52
Obtrusiveness assessment	2.41	0.56	3.35	0.79	2.49	0.57	3.35	0.67
Overall vision on assessment	2.80	1.02	3.91	0.86	3.17	0.88	3.82	0.92
Cluster: Role of student in assessment								
Involvement in assessment	3.20	0.45	3.87	0.42	3.13	0.44	3.23	0.66
Group behaviour	4.17	0.45	3.95	0.57	3.93	0.54	3.86	0.56
Collaborative learning	3.80	0.60	4.21	0.55	3.98	0.44	4.12	0.68
Assessment skill	3.82	0.41	3.89	0.36	3.69	0.37	3.85	0.36
Group atmosphere[a]	—	—	4.36	0.50	—	—	4.27	0.48

[a]Only measured in post-test.

Table 5. F-values in MANOVAs with repeated measures on scores on the 16 variables of the student questionnaire

Variable	Time of Testing	Groups	Content Domain	Time of Testing × Groups	Time of Testing × Content Domains	Time of Testing × Groups × Content Domain
Satisfaction classes	100.49***	0.15	2.58	4.81*	2.67	0.64
Transparency classes	26.89***	2.54	2.45	1.94	2.84	0.60
Learning access level	5.69*	0.06	1.67	1.22	4.14**	1.04
Practical relevance	26.36***	0.06	1.96	1.58	5.54**	3.46*
Quality of the instruction	146.65***	0.65	0.87	1.48	2.82*	1.03
Teacher involvement	75.89***	1.08	2.29	0.83	5.95**	1.64
Relation instruction and assessment	301.13***	1.21	2.43	0.02	5.07**	2.06
Assessment behaviour	55.02***	1.03	0.50	0.11	1.09	1.45
Fear for assessment	1.64	0.35	1.79	0.041	0.78	0.66
Obtrusiveness assessments	108.32***	0.00	1.24	0.62	6.51**	4.38**
Overall vision on assessment	34.81***	1.26	3.58*	0.69	2.31	1.13
Involvement in assessment	0.03	1.30	1.46	0.30	0.67	2.41
Group behaviour	6.20**	0.56	0.04	0.02	1.36	0.38
Collaborative learning	0.105	0.37	0.19	0.42	0.28	0.17
Assessment skill	6.40**	0.19	0.90	0.59	1.53	2.04
Group atmosphere[ab]	—	0.25	0.94	—	—	—

[a]This variable was only added in post-test, and therefore measured with a one-way ANOVA.

[b]For Group × Content Domain, $F(3,58) = 2.28$, MSE = 0.197, $p < .05$.

*$p < .05$; **$p < .01$; ***$p < .001$.

Testing) analysis of variance with repeated measures on the last factor. Table 5 presents the *F*-values for each of the MANOVAs on the scores of all the variables.

As indicated in Table 5, there were highly significant main effects for Time of Testing. For 12 of the 15 variables, students were more positive in the post-test than in the pre-test. One important significant effect was the effect on the variable 'assessment skill', because this variable concerned items that measured students' self-perception of their skill to assess.

There were no significant main effects for Groups. For Content Domain, a significant effect was found regarding the variable 'overall vision on assessment', $F(3,53) = 3.58$, MSE = 0.561, $p < 0.05$. Post-hoc tests (Tukey) revealed that there was only one significant effect between two content domains, namely mathematics and philosophy (mean difference = 0.62, $p < .05$). The effect though is not caused by the treatment, and is therefore less important.

For Time of Testing × Groups a significant interaction effect was found on the variable 'satisfaction classes', $F(3,52) = 4.81$, MSE = 0.173, $p < .05$. For the pre-test the experimental groups (mean = 3.11, SD = 0.52) were more positive than the control groups (mean = 2.99, SD = 0.54), while for the post-test the opposite pattern was shown (experimental groups: mean = 3.76, SD = 0.55; control groups: mean = 3.87, SD = 0.54).

For Time of Testing × Content Domains a significant interaction effect was found on the variables 'learning access level', 'practical relevance', 'quality of the instruction', 'teacher involvement', 'relation instruction and assessment' and 'obtrusiveness assessment'. Means and standard deviations are presented in Table 6. The means and post-hoc analysis (Tukey) indicated that for all six variables the mathematics group showed a much lower increase or even a decrease, from the pre-test to the post-test, than the three other groups.

Table 6. Means and standard deviations of the six variables on a five-point Likert scale

	Pre-test		Post-test	
Variable/Content domain	Mean	SD	Mean	SD
Learning access level				
Pedagogy	3.78	0.71	3.76	0.60
Physics	3.13	0.83	3.84	0.76
Mathematics	3.57	0.66	3.25	0.85
Philosophy	3.76	0.92	4.17	0.60
Practical relevance				
Pedagogy	3.68	0.58	4.22	0.51
Physics	3.33	1.01	4.12	0.62
Mathematics	3.62	0.61	3.45	0.89
Philosophy	3.61	0.78	4.22	0.64
Quality of the instruction				
Pedagogy	2.84	0.55	3.97	0.36
Physics	2.71	0.74	3.93	0.55
Mathematics	2.99	0.54	3.66	0.67
Philosophy	2.95	0.72	4.34	0.56
Teacher involvement				
Pedagogy	3.44	0.57	4.12	0.46
Physics	2.90	0.80	3.96	0.44
Mathematics	3.49	0.65	3.63	0.84
Philosophy	3.45	0.70	4.48	0.46
Relation instruction and assessment				
Pedagogy	2.30	0.65	4.04	0.54
Physics	1.98	0.72	4.17	0.57
Mathematics	2.39	0.63	3.44	1.02
Philosophy	2.49	0.59	4.29	0.59
Obtrusiveness assessment				
Pedagogy	2.38	0.49	3.44	0.59
Physics	2.35	0.56	3.41	0.75
Mathematics	2.55	0.67	2.92	0.69
Philosophy	2.56	0.57	3.69	0.75

Table 7. Means and standard deviations of the two involved variables on a five-point Likert scale

	Pre-test, experimental		Pre-test, control		Post-test, experimental		Post-test, control	
Variable/Content domain	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Practical relevance								
Pedagogy	3.73	0.43	3.62	0.73	4.52	0.41	3.88	0.40
Physics	3.42	0.62	3.22	0.62	3.89	0.50	4.53	0.65
Mathematics	3.56	0.37	3.39	0.81	3.33	0.69	3.56	1.04
Philosophy	3.42	0.92	3.81	0.57	4.07	0.64	4.44	0.62
Obtrusiveness assessment								
Pedagogy	2.28	0.39	2.47	0.57	3.63	0.55	3.25	0.60
Physics	2.19	0.56	2.53	0.52	3.46	0.76	3.31	0.82
Mathematics	2.68	0.57	2.41	0.77	2.55	0.73	3.21	0.53
Philosophy	2.54	0.65	2.59	0.49	3.63	0.73	3.76	0.84

For Time of Testing × Groups × Content Domains a significant interaction effect was found on the variables 'practical relevance' and 'obtrusiveness assessment'. Means and standard deviations were calculated for these variables and are presented in Table 7.

For 'practical relevance', the experimental mathematics group showed a decrease from the pre-test to the post-test while all other groups were more positive on the pre-test than on the post-test. For 'obtrusiveness assessment', a similar pattern was observed.

After the course, 16 students (eight from the experimental group, eight from the control group) were asked 11 questions about the redesigned course and the peer assessment they carried out at the end of the course. The calculated frequencies indicated that 94% of the students rated the extent to which they had to work independently as high. The same counts for the individual contribution in the group (75%). All students evaluated working in small groups as very positive. Seventy-five per cent of the students were satisfied with the peer-assessment procedure at the end of the course. That a learning effect occurred as a consequence of the peer assessment itself was subscribed by 75% of the students, whereby the students from the experimental group are in majority. All students stressed the importance of peer assessment for their role as professional teachers. The majority of the students (83%) did not feel capable of assessing a peer; 63.5% of the interviewed students stressed that it is still uncomfortable to assess a peer; 94% of the students indicated that they would like to receive more training in assessment skills. One student said:

> I would like to have more training in this type of skill … I never realized that assessing the work of a peer is so difficult … I think that this training is a step in the right direction … a first impression … but I like to know more about it.

As far as the peer-assessment tasks are concerned, seven of the eight interviewed students of the experimental group were satisfied with the instructions. Half of the students from the experimental group indicated that they learned from the peer-assessment tasks and their peers. One student described this relationship with the peers as follows:

Table 8. Medians of the variables of the teacher questionnaire on a five-point Likert scale

	Peer-assessment task			
Variables	1	2	3	4
Invested time	3.50	4.00	3.50	4.00
Transparency of the task	2.50	4.00	4.00	4.00
Desired output	3.50	4.00	4.00	5.00

> I think it is useful to pay attention to the development of assessment skills, because what you practise with peers, you can also use in the class environment with pupils. That also is the case when you observe lessons of a colleague. Such activities are very purposeful.

What are the perceptions of the teachers in the Teacher Training College about the peer-assessment training and the redesigned course?

The teacher questionnaire and the teacher interviews were analysed to investigate this fourth research question. In Table 8, the medians were calculated for the three variables of the teacher questionnaire: (1) invested time; (2) transparency of the task; and (3) desired output.

All teachers indicated that the peer-assessment tasks could be taught in the available time. The teachers in this study were more able to arrange their instruction time in the fourth peer-assessment task than in the first peer-assessment task. The means of the transparency of the task show that the students mostly understood the goals of each task. In the fourth peer-assessment task, all teachers achieved the desired output.

Implementing the training forced the teachers to discuss the content from an alternative angle. One teacher described this process as follows:

> We wanted to define clear goals regarding the design of discovery learning plans that were recognisable for each of the content domains. That is something that I always aimed at, but personal desires of individual teachers about the content obstructed this process. In the redesign-phase, teachers were forced to leave their own territory. And that is mostly a matter of attitude. The systematic approach, continuing reflection, and documenting several steps made the redesign successful.

The change in role definition was hard to accept for the physics teacher. This teacher indicated that after 30 years of teaching experience, the willingness to innovate decreased. The mathematics teacher was more positive about his 'new' role. One teacher expressed the following:

> My experience as a designer changed my view on what a teacher should be fundamentally. It became clear to me that my main task is not educating student teachers towards mathematicians, but towards educators of mathematics. The redesign of this course was definitely an eye-opener.

Conclusion and discussion

The objective of this study was to investigate the effects of peer-assessment training on the development of the peer-assessment skill and the effects on the performance of students. Peer

assessment in this study did not focus on scoring peers on a number of criteria, as in many peer-assessment studies (see Boud & Falchikov, 1989; Falchikov & Goldfinch, 2001), but on the quality of peer assessments of individual student teachers. It also explored what the effects were of the redesign of the course in a more skill-based way on several variables, based on students' perceptions. The influence of content domain was analysed. Teacher perceptions were taken into account regarding the process of the peer-assessment training and the redesign of the course.

First, the main results will be briefly summarized. Regarding the first question whether training in peer assessment leads to the development of the skill to assess the work of peers, the answer is positive. Results from the presented study reveal that the student teachers from the experimental groups were more capable in using the set criteria determined during the peer-assessment tasks than the student teachers of the control groups. This confirms our hypothesis that peer assessment is a skill that can be trained.

Some reservations are in order with regard to this result, because the results also show that the student teachers from the experimental groups still are novices in their assessment skills, especially in the use of the criteria. The means of the experimental groups are still low. The difference between the control groups and experimental groups are small. An explanation for this result may be caused by the short training period. Complex skills need to be trained during an extensive period of time in several contexts (Van Merriënboer, 1997). The fact that the training only focused on the use of criteria could be an explanation for the unexpected result that the student teachers in the control groups gave more points for improvement. These are aspects of the peer-assessment skill that were not trained. It might be interesting in a follow-up study to train both groups in giving feedback also, to see if then an effect occurs in both groups.

The second research question focused on the effect of training in peer-assessment skills on students' performance. A difference between the performance quality of the students from the control and from the experimental groups was not found. The small progress in the peer-assessment skill may be the reason that an effect on the quality of the end-product could not be recorded. It is possible that further training will eventually lead to an effect on the level of performance. A second explanation could be that the redesign of the course had an effect on the learning result of all students.

On the third research question, what are the perceptions of students regarding the redesigned course and does the content domain influence these perceptions, several results were found. The results showed a change of perception towards 12 aspects of instruction and assessment. The whole group was more positive about the instruction and the integration of assessment and instruction after they took the redesigned course. The renewed course led to an active participation of student teachers and the teachers of the Teacher Training College. It can be concluded that the student teachers changed positively in their views on aspects of learning and assessment. They are more satisfied about the classes, the criteria and goals are clearer. The role of teachers was also evaluated in a more positive way. The student teachers indicated that they are more capable in assessing than before the redesigned course. In the interviews though students also indicated that they did not feel an expert assessor after the training.

The relationship between the peer assessment and the role of a teacher was clear for the students. In additional comments, students indicated that it was sometimes hard to translate their thoughts about the work of a peer in writing. In this perspective, it is interesting to

study the differences between the quality of oral and written assessments in future research. Another student pointed out that students have to prove that they understand the criteria before they can assess a peer. They need to have an objective perspective and give constructive criticism.

The factor content domain had a high influence, mainly caused by the domain mathematics. It was not possible to determine whether these findings were the result of the content domain itself, or the teacher involved in the content domain. It is, however, remarkable that one domain causes the significant interaction effects. It may be due to the specific character of the mathematics domain.

With regard to the fourth research question, it can be concluded that major problems in teaching the peer-assessment tasks did not occur. Because of the small number of teachers the interviews are not structurally elaborated. The answers that teachers gave were illustrative of three assumptions that will be further explained.

A first one is that the metaphor 'the tail wags the dog' was underlined by the teachers: implementing the peer-assessment training led to a rethinking of the existing instructional material. To close the gap between instruction and assessment, a redesign of existing courses often seems to be inevitable, since the criteria of the products have to be operationalized. This was a consequence of the definition of the key outcomes desired at the end of the course. Clarity about these outcomes must be obtained before assessment activities are designed (Boud *et al.*, 1999). The chosen redesign led to the situation that the summative assessment was sufficiently related to the study material.

Second, in line with this first assumption, it can be argued that the *role of the teachers* was reconsidered. The teachers became more skilled in defining skills and designing effective study tasks, instead of only being an expert in a certain content domain. The teachers in the Teacher Training Colleges also have to become reflective practitioners (Schön, 1987).

A third assumption is that the teachers in the current Dutch educational system still spend most of the day separated from colleagues, with little time or opportunity to share problems encountered in the class environment. In contrast, teachers in other countries are given far more paid time for planning: Japanese teachers for example spend about 40% of their paid time on professional development and collaboration compared with about 20% for their Dutch counterparts (Web-based Education Commission, 2000). One teacher indicated that the training teachers do receive in skill development is usually too little, too basic and too generic to help them develop complex skills in their everyday teaching. Teachers need more than a quick course in skill development. They need guidance in using the best tools in the best ways to support the best kinds of instruction. Above all, they need time.

Some comments about the conducted research set up have to be made. The first one is that certain effects might have been masked by the fact that both the control groups and the experimental groups received a redesigned course. Secondly, the present study focuses on short-term effects. It is conceivable that peer-assessment training and more critical reflection about assessment might have a long-term effect for students, which was not taken into account in this study. Third, analysis of the dependent variables focused on a quantitative approach. No in-depth analyses were performed on, for instance, the quality of the criteria used by students. Another aspect of the analysis concerns the fact that the set-up of the current study makes it difficult to distinguish between teacher effects and domain effects. Finally, much emphasis was put on the

ecological validity of the study. This inevitably decreased the experimental control that would have been possible in a more laboratory-like approach. In this study, for instance, students might have exchanged ideas or guidelines between groups and, although we tried to control for this, teachers might have done so as well.

The results of this study, as well as certain design aspects of the study, put forward a need for further research. Studies that allow unravelling domain effects and teacher effects, as well as studies that take long-term effects into account, are required. Small-sized studies with more in-depth analysis of the student use of criteria, question posing, the development of student feedback and so on, could be combined with such studies. Future research might also allow an extension of the skills that were trained, going further than defining criteria for assessment, which was the principal skill in the training in the current experimental condition. To date, research is conducted that aims at the assessment of long-term effects and at the development of student feedback. With this type of research that is embedded in the everyday learning practice of students and teachers, it is possible to develop students who are not only able to analyse the work of peers, but also have structural involvement in the design of their own education.

Notes on contributors

Dominique Sluijsmans is an educational technologist at the Educational Technology Expertise Centre at the Open University of The Netherlands. Her main interests are student involvement in assessment, peer assessment, teacher education and curriculum design.

Saskia Brand-Gruwel is an educational technologist at the Educational Technology Expertise Centre at the Open University of The Netherlands. Her areas of specialization are comprehensive reading, higher-order skills and competency-based education.

Jeroen van Merriënboer is head of the research department at the Educational Technology Expertise Centre at 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.

Rob Martens is a senior educational technologist at the Educational Technology Expertise Centre at the Open University of The Netherlands. His areas of specialization are assessment, cognitive load, distance learning and electronic learning.

References

Birenbaum, M. (1996) Assessment 2000: towards a pluralistic approach to assessment, in: M. Birenbaum & F. Dochy (Eds) *Alternatives in assessment of achievements, learning processes and prior knowledge* (Boston, MA, Kluwer Academic Press).

Boud, D. (1995) *Enhancing learning through self-assessment* (London, Kogan Page).

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

Boud, D. & Falchikov, N. (1989) Quantitative studies of self-assessment in higher education: a critical analysis of findings, *Higher Education,* 18, 529–549.

Brown, S., Rust, C. & Gibbs, G. (1994) *Strategies for diversifying assessment* (Oxford, Oxford Centre for Staff Development).

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

Cohen, E. G. (1994) Restructuring the classroom: conditions for productive small groups, *Review of Educational Research,* 64, 1–35.

Darling-Hammond, L. & Snyder, J. (2000) Authentic assessment of teaching in context, *Teaching and Teacher Education,* 16, 523–545.

Falchikov, N. (1995) Peer feedback marking: developing peer-assessment, *Innovations in Education and Training International,* 32, 175–187.

Falchikov, N. & Goldfinch, J. (2001) Student peer-assessment in higher education: a meta-analysis comparing peer and teacher marks, *Review of Educational Research,* 70, 287–322.

Fallows, S. & Chandramohan, B. (2001) Multiple approaches to assessment: reflections on use of tutor, peer and self-assessment, *Teaching in Higher Education,* 6, 229–246.

Freeman, M. (1995) Peer-assessment by groups of group work, *Assessment and Evaluation in Higher Education,* 20, 289–300.

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

James, P. (2000) A blueprint for skills assessment in higher education, *Assessment and Evaluation in Higher Education,* 25, 353–367.

Johnson, D. W., Johnson, R. T. & Johnson-Holubec, E. (1992) *Advanced cooperative learning* (Edina: Interaction Book Company).

Korthagen, F. & Wubbels, T. (2000) Are reflective teachers better teachers?, in: G. M. Willems, J. H. J. Stakenborg & W. Veugelers (Eds) *Trends in teacher education* (Leuven-Apeldoorn, Garant).

Kremer-Hayon, L. & Tillema, H. H. (1999) Self-regulated learning in the context of teacher education, *Teaching and Teacher Education,* 15, 507–522.

LPC (1995) *Profession in action. Vocational training profile for the primary school teacher* (Utrecht, Forum Vitaal Leraarschap).

Mehrens, W. A., Popham, W. J. & Ryan, J. M. (1998) How to prepare students for performance assessment, *Educational Measurements: Issues and Practice,* 17, 18–22.

Orsmond, P., Merry, S. & Reiling, K. (1996) The importance of marking criteria in the use of peer-assessment, *Assessment and Evaluation in Higher Education,* 21, 239–249.

Orsmond, P., Merry, S. & Reiling, K. (1997) A study in self-assessment: tutor and students' perceptions of performance criteria, *Assessment and Evaluation in Higher Education,* 22, 357–369.

Orsmond, P., Merry, S. & Reiling, K. (2000) The use of student derived marking criteria in peer and self-assessment, *Assessment and Evaluation in Higher Education,* 25, 23–38.

Reilly Freese, A. (1999) The role of reflection on pre-service teachers' development in the context of a professional development school, *Teaching and Teacher Education,* 15, 895–909.

Schön, D. A. (1987) *Educating the reflective practitioner: towards a new design for teaching and learning in the professions* (San Francisco, CA, Jossey-Bass).

Sharan, Y. & Sharan, S. (1994) Group investigation in the cooperative classroom, in: S. Sharan (Ed.) *Handbook of cooperative learning methods* (Westport, CT, Praeger).

Slavin, R. E. (1995) *Cooperative learning: theory, research and practice* (Boston, MA, Allyn & Bacon).

Sluijsmans, D., Dochy, F. & Moerkerke, G. (1999) Creating a learning environment by using self-, peer- and co-assessment, *Learning Environments Research,* 1, 293–319.

Sluijsmans, D., Moerkerke, G., Dochy, F. & Van Merriënboer, J. J. G. (2001) Peer-assessment in problem based learning, *Studies in Educational Evaluation,* 27, 153–173.

Sluijsmans, D. & Van Merriënboer, J. J. G. (2000) *A peer-assessment model* (Heerlen, Open University of The Netherlands).

Tillema, H. H., Kessels, J. W. M. & Meijers, F. (2000) Competencies as building blocks for integrating assessment with instruction in vocational education: a case from the Netherlands, *Assessment and Evaluation in Higher Education,* 25, 265–278.

Van Merriënboer, J. J. G. (1997) *Training complex cognitive skills* (Englewood Cliffs, NJ, Educational Technology Publications).

Verloop, N. & Wubbels, T. (2000) Some major developments in teacher education in the Netherlands and their relationship with international trends, in: G. M. Willems, J. H. J. Stakenborg & W. Veugelers (Eds) *Trends in teacher education* (Leuven-Apeldoorn, Garant).

Web-based Education Commission (2000) Helping isolated teachers make new connections, in: *The power of the internet for learning*. Available online at: http://interact.hpcnet.org/webcommission/doc.htm (accessed 13 March 2001).

Willems, G. M., Stakenborg, J. H. J. & Veugelers, W. (Eds) (2000) *Trends in teacher education* (Leuven-Apeldoorn, Garant).

Woolhouse, M. (1999) Peer-assessment: the participants' perception of two activities on a further education teacher education course, *Journal of Further and Higher Education,* 23, 211–219.