Student involvement in assessment

The training of peer assessment skills

ico

Interuniversity Centre for Educational Research

www.ou.nl/otecresearch

ISBN: 90-52-78-345-2

Druk: Datawyse Maastricht
Omslag: Open Universiteit Nederland

© Dominique M.A. Sluijsmans, Heerlen, 2002

Alle rechten voorbehouden.
All rights reserved.

Student involvement in assessment

The training of peer assessment skills

PROEFSCHRIFT

Ter verkrijging van de graad van doctor aan de Open Universiteit Nederland op gezag van de voorzitter van het College voor Promoties prof. dr. W.M.G. Jochems

in het openbaar te verdedigen ten overstaan van het College voor Promoties te Heerlen op vrijdag 28 juni 2002 om 16.00 precies

door

Dominique Margarete Agnes Sluijsmans

geboren op 5 november 1973 te Kerkrade

Promotor:
Prof. dr. J.J.G. van Merriënboer, Open Universiteit Nederland

Toegevoegd promotor:
Dr. F.L.J.M. Brand-Gruwel, Open Universiteit Nederland

Leden beoordelingscommissie:
Prof. dr. J.H.H. van den Akker, Universiteit Twente
Prof. dr. H.M.J. van den Bosch, Open Universiteit Nederland
Dr. M. Elshout-Mohr, Universiteit van Amsterdam
Prof. dr. F.A.J. Korthagen, IVLOS, Universiteit Utrecht

‘It is not the actual methods or tools of assessing
which we believe should be changed in many case,
rather the underlying philosophy
and the aims of their use and application.’

From: Harris, D., & Bell, C. (1990). *Evaluating and assessing for learning*. Kogan Page: London. (p. 97)

Voorwoord

In de afgelopen vier jaar ben ik in de bevoorrechte positie geweest een groot deel van mijn tijd te mogen werken aan een proefschrift dat ik naar eigen inzichten kon invullen. Nu de afronding van deze fase is aangebroken, is het tijd even stil te staan bij deze periode en een aantal mensen te bedanken.

Filip Dochy, Paul Kirschner, Martin Valcke en Henk de Wolf, jullie zijn diegenen geweest die mij als eersten een gevoel van vertrouwen gaven toen ik in september 1997 bij het OTEC startte. Paul, ik kijk ernaar uit samen met jou verder te werken aan onderzoek naar peer assessment.

Wim Jochems en Freek Gastkemper, veel dank voor het bieden van een ideale werkomgeving.

In de periode dat ik aan het proefschrift heb gewerkt, heeft een aantal mensen in verschillende fasen van het traject een bijdrage aan de begeleiding geleverd. Hiervoor wil ik Filip Dochy, George Moerkerke, Rob Martens, Theo Bastiaens en Saskia Brand graag bedanken.

Jeroen, aan jou wil ik een bijzonder woord van dank richten. Toen jij in september 1998 bij OTEC als hoogleraar Onderzoek binnenkwam, heb jij mij pas echt goed bevraagd wat ik met mijn onderzoek wilde. Je hebt mij gemotiveerd een onderzoeksvoorstel te schrijven waarop je me kritisch bevroeg. Je hebt altijd vertrouwen in mijn werk gehad en mij de ruimte gegeven zelf de criteria voor het proefschrift te bepalen en met jou daarover in gesprek te gaan. Mede dankzij jou ben ik onderzoek gaan waarderen.

Ook dank aan alle mensen van de onderzoeksgroep. Bij deze wens ik alle AIO's en promovendi veel succes bij hun trajecten. Ik kijk uit naar jullie promoties!

Docenten en studenten van de Faculteit Onderwijs van de Hogeschool Zuyd, super dat ik bij jullie mijn experimenten mocht uitvoeren. Ik besef maar al te goed dat het niet niks was wat ik van jullie vroeg. Bert, aan jou een speciaal woord van dank. Mooi om te weten dat je het geloof in peer assessment met me deelt. Ik hoop dat we nog vaak zullen samenwerken in de toekomst en kunnen discussiëren over alles wat met onderwijs te maken heeft.

Andrea, Ronald en Wendy, geweldig hoe jullie je door de honderden assessmentreports van de studenten hebben geworsteld. Jullie zijn studentassistenten die hun taak serieus en nauwkeurig uitvoeren. Ik waardeer de inzet en de interesse in de resultaten enorm.

Jos Crombach en Coen Voogd, vele malen zadelde ik jullie op met repro-werk. Iedere keer leverden jullie alles perfect af. Vivian Rompelberg, bedankt voor de dynamische omslag!

Om goed te kunnen werken, is een fijne werkomgeving van groot belang. OTEC is zo'n omgeving, waarbij ik al vanaf de eerste dag dat ik er was voelde dat de deuren open stonden. Hoewel ik veel mensen bij OTEC en de OU een warm hart toedraag, wil ik Theo Bastiaens, Jo Boon, Evelijn Bos, Jan van Bruggen, Ronald Gossiau, Henry Hermans, Jose Janssen, Liesbeth Kester, Paul Kirschner, Rob Martens, Roos van der Meulen, Frans Prins, Jos Rikers, Peter Sloep en Huib Tabbers in het bijzonder noemen.

Huib en Liesbeth, jullie zijn toch wel m'n beste AIO-maatjes. Huib, veel gelachen om en met jou, jouw onuitputtelijke relativeringsvermogen werkte goed tegen mijn pessimistische buien. Ik hoop hiervan wat mee te nemen. Jammer dat je het zuiden verlaat. Liesbeth, met jou kan ik ook kletsen over de dingen die niks met onderzoek te maken hebben. Je hebt me bevestiging gegeven op momenten dat ik twijfelde over het een en ander. Ik ben trots op jou als collega en vriendin.

Ook buiten de OU zijn er lieve mensen bij wie ik m'n hart altijd mag luchten. Wendy, Susanne, Hortense, Milleke en Paul, Dorianne, Simone en Paul, Mieke en Jorrit, Brenda en Martijn, bedankt voor jullie blijvende belangstelling. Jullie vriendschap is mij veel waard.

Veel dank aan mijn ouders Léon en Marianne, zussen Daniëlle en Christiane en broer Luc, voor hun interesse in mijn werk en luisterend oor. Pap en Luc, echt te gek dat we naast familie ook collega's zijn. Mam, jij laat mij altijd weten dat ik vooral de dingen moet doen die ik graag wil doen. Pap, jij vooral bedankt voor het scheppen van de randvoorwaarden op de Faculteit Onderwijs voor mijn experimenten. Ik waardeer het dat je altijd op een gezonde afstand het onderzoek hebt gevolgd en "mij m'n ding hebt laten doen". Ik hoop nog lang met jou te kunnen mijmeren over het wel en wee in onderwijsland.

Tenslotte richt ik mij tot Roger. Ik besef maar al te goed dat jij mijn constante en stabiele begeleider bent geweest. Vooral de laatste maanden heb jij menig uurtje gestoken in het samen afronden van het proefschrift. Bedankt voor je kritische blik naar mijn werk en het stellen van de juiste vragen. Ik bewonder je inzicht en geduld. Ik kan nog veel van je leren.

Dominique

Kerkrade, mei 2002

Contents

Chapter 1

General introduction

Institutions of higher education are continuously confronted with a demand for more competency based learning, problem based learning, student centred learning, and project based learning. De Corte (1996) refers to the design of powerful learning environments. These learning environments have four main characteristics. First, they provide authentic, open problems and learning materials which have personal meaning for students and are presented in a variety of formats. Second, teaching methods are applied which arouse interest, activate prior knowledge, clarify meanings, and model appropriate learning strategies and reflective processes. In the third place, they initiate external regulation of specific learning strategies. At last, the monitoring of strategies and discussion of them in small groups is encouraged, whereby a classroom culture is achieved which encourages reflection on process (see also Entwistle, 2000).

Regarding these characteristics, one could situate a powerful learning environment in the constructivistic perspective, which emphasizes that learning means actively constructing knowledge and skills on the basis of prior knowledge, embedded in contexts that are authentic and offer ample opportunities for social interaction (Brown, Collins, & Duguid, 1989). Since the goals as well as the methods of instruction are oriented towards more complex curricular objectives, it is necessary for assessment practices to increasingly use various kinds of assessments in which students have to interpret, analyse and evaluate problems and explain their arguments. These assessments provide information about student progress and support students in selecting appropriate learning tasks. The traditional assessment approach concentrates mainly on the testing of basic skills, supposedly acquired through mainly drill and practice experiences. Such an assessment system is often referred to as a test culture (Birenbaum, 1996). A test measures knowledge or ability after a particular course unit, with the assumption that the product of learning will contain in itself all of the information that the evaluator needs to know about the learners and the quality of their thinking processes (Wiggins, 1989). Assessment that is performance-oriented on the contrary aims to measure not only the correctness of a response, but also the thought processes involved in arriving at the response. It encourages students to reflect on their own learning, and the belief is that instruction will be pushed into a more thoughtful, more reflexive, richer mode as well. Teachers who teach according to these kinds of alternative assessments will teach in ways that emphasize reflection, critical thinking, and personal investment in one's own learning.

Since assessment is broader than a set of standardized tests, one could say that a shift occurs from a test culture to an assessment culture, which strongly emphasizes integration of instruction, learning and assessment. It is suggested that the assessment culture fits in well with a learning environment based on the principles of powerful learning environments. Boud (1990) stresses that assessment practices in higher education have to be compatible with the curricular goals. Educational innovations, such as problem based education, new learning and competence based education, are more likely to succeed if they include new forms of assessment, whereby assessment and learning are strongly interconnected in the course materials. Assessment procedures then should not only serve as a tool for crediting students with recognised certificates, but also should be used to monitor progress and, if needed, to direct students to remedial learning activities. Assessment is now represented as a tool for learning (Arter, 1996; Dochy & McDowell, 1997), making an involvement of the students more and more desirable.

The compatibility between learning, instruction and assessment is the starting point of this dissertation, which will be further described within the theory of constructive alignment, introduced by Biggs (1996, 1999, 2001). Constructive alignment represents a marriage between constructivism and instructional design. Constructivism being used as a framework to guide decision-making at all stages in instructional design: in deriving curriculum objectives in terms of performances that represent a suitably high cognitive level, in deciding teaching/learning activities judged to elicit those performances, and to assess and summatively report student performance. When there is alignment between what teachers want to teach, how they teach and how they assess, teaching is likely to be more effective than when it is not. In this dissertation, it is argued that a well-designed course will hinge on a close relationship between the essential elements learning, instruction and assessment. A poorly designed course will not act to develop these close relationships and consequently it will be difficult for learners to achieve their desired learning goals.

From a scientific point of view, it is important to specify the theory of constructive alignment in a research goal. Main aim of this dissertation is to design a training which supports students who are active participants in performance-based assessments to become professionals in the skill to assess. It will be investigated what the effects of a training in assessment skills is on students' performance in their peer assessment skills and content based skills. Several issues that are of importance to understand the empirical studies are outlined in this introduction. First, the issue of student involvement in general is further elaborated in self-, peer-, and co-assessment. Because the majority of the studies focuses on the involvement of students as peer assessors in performance assessments, it is explained how the performance assessments and the peer assessments in the studies are designed. The characteristics of the assessment training are briefly introduced. Supporting students in developing their peer assessment skills involves activities in which students collaborate. The relation between peer assessment and collaborative learning is described. An integrated framework is presented, which illustrates the strong relation between course design, student involvement, performance assessment, skill acquisition and collaborative learning. Particularly the studies reported in Chapters 4, 5 and 6 are based on this integrated framework. The research context is outlined, in which the studies took place. At the end of this chapter the structure of the dissertation and the corresponding research questions will be presented.

Student involvement

Self- peer- and co-assessment

A conceptual shift which has occurred in recent times in higher education has been from a perspective which focused on the teacher, to a perspective in which student involvement is more central (Boud, 1995). As is stated in the prior section, it is highly supported that students make an active contribution to their own learning. Within assessment procedures, students can be involved by means of self-, peer- and co-assessments. Self assessment refers to the involvement of learners in making judgments about their own learning, particularly about their achievements and the outcomes of their learning (Boud & Falchikov, 1989). Peer assessment is the process whereby individuals evaluate their peer(s) (Falchikov, 1995; Freeman, 1995). Evaluating the work of peers is a social activity, especially when the peer assessment is non-anonymous. Co-assessment, the participation of the students together with the teacher in the summative assessment process, is a way of providing an opportunity for students to assess themselves while allowing the staff to maintain the necessary control over the final assessments (Hall, 1995). Self-, peer- and co-assessments should not be viewed as assessment methods. They only indicate that the student is involved as one of the assessors.

Required skills, competencies and the criteria for mastery or achievement can be identified by students themselves, and their own judgement of success can be incorporated in the final grade. Assessment of students' own work and that of peers is relevant to many disciplines and provides for reflection on the context of the course and the community of learners. Encouraging students to assess each other's contributions to discussion and discourse is further exposing them to the skills of critical reflection and analysis (Birenbaum, 1996; Sambell & McDowell, 1998). Therefore, self-, peer-, and co-assessment are regarded as *learning tools* that may have positive effects on other skills. According to many authors (e.g., Glaser, 1990; Glaser & De Corte, 1992; Dochy, 1992; Martens & Hermans, 2000) assessment should not only be used on a formal base. Involving students in the assessment creates possibilities for formative assessment during learning processes. Self and peer assessment are closely related and in many cases applied in combination. An extensive literature review on student involvement in assessment is provided in Chapter 2.

Peer assessment in performance assessment

In the majority of the studies presented in this dissertation, the focus lies on peer assessment. This is decided based on the specific character of peer assessment that focuses on social interaction and collaborative learning (see for further elaboration Chapter 1). As already addressed in the prior section, peer assessment indicates that students are active participants in the assessment. To pursue the theory of constructive alignment, it is decided to involve students in assessments that are further considered as performance assessments. Performance assessments 'represent a set of strategies for the . . . application of knowledge, skills, and work habits through the performance of tasks that are meaningful and engaging to students' (Hibbard, et al., 1996, p. 5). Performance assessment can be based on multiple products or processes, for example essays, reflection papers, oral assessments, simulations, process-analyses, group-products, and work-samples.

The performance based assessments are integrated into the instructional process to provide additional learning experiences for students. The term performance is chosen, since the assessment task is described in terms of a certain performance, that is perceived as worthwhile and relevant to the student. This performance may or may not represent an authentic situation (Wiggins, 1989). Performance assessment focuses on the ability to use combinations of acquired skills and knowledge, and therefore fits in well with the theory of constructive alignment and powerful learning environments (Linn, Baker, & Dunbar, 1991). Performance assessments require individuals to apply these knowledge and skills in context, not merely completing a task on cue. Students are observed while they are performing; products they create are examined, and the level of proficiency demonstrated is judged. Judgments are made about the level of achievement attained by comparing student performance to predetermined standards. All students have the opportunity to attain the standards, whereby they can play a crucial role in making judgments about the performance of their peers and defining appropriate criteria for these performances.

Designing performance assessments

In the last decade, considerable effort was put into the organisation of powerful learning environments in higher education. From the learner's perspective, assessment is usually the element of highest importance. Unfortunately, assessment often is the suppositious child in innovations. Yet, there are two issues that stress the impact of assessment in curricula. First, the influence of tests on what is taught is potentially great. Frederiksen (1984) refers in this scope to the 'real test bias'. There is evidence that tests do influence teacher and student performance and that for example multiple-choice tests tend *not* to measure the more complex cognitive abilities. It is suggested that other formats such as performance assessments encourage the teaching of higher level cognitive skills and provide practice with feedback.

Second, there is a distinction between 'what is meant to happen', that is, the curriculum stated officially by the educational system or institution, and what teachers and learners actually do and experience 'on the ground', a kind of de facto curriculum. Snyder (1971) calls this the hidden curriculum. Assessment is regarded as the element of educational practice which most powerfully determines the 'hidden curriculum' (Lockwood, 1995; Portelli, 1993; Sambell & Mc Dowell, 1998). In a lab researchers can ask students to read texts but in 'real life' students have there own hidden curriculum, 'adopting ploys and strategies to survive in the system.' (Lockwood, 1995, p. 197). Preparation for assessment may even hinder students' efforts towards genuine understanding of course material (Entwistle & Entwistle, 1997).

Commonly, teachers are not educated as assessors. Teachers therefore occasionally doubt the fairness of the way they currently assess students' performance (Airasian, 1991). A common error in designing a course or unit of study is to leave the development of the assessment tasks as a final activity. The basis of the effective application of performance assessment methodology is thoroughly trained raters relying on sound performance criteria to observe and evaluate student responses to quality exercises (Stiggins, 1994). Training in designing performance assessments - or at least a systematic verification of qualifications to rate performance - appeared to be essential in all contexts in which quality assessment results are the goal.

To design sound performance assessments, some guidelines are provided by Stiggins (1987). For the studies reported in Chapters 4, 5 and 6, it is decided to use a combination of Biggs' (1996) theory of constructive alignment and Stiggins' approach (1987) to design existing courses and performance assessments. The following four steps were taken to design courses in which instruction and asessments are completely aligned:

Step 1. Defining the purpose of the performance assessment. First, teachers must have a clearly defined purpose of a course. Several important questions are in order: what concept, skill, or knowledge is assessed? What should the students know? At what level should the students be performing? (Stiggins, 1987) By answering these questions, it is decided what type of activity best suits the assessment needs. This for example results in a skill decomposition in which the relevant skills are ordered hierarchal (Chapter 4), or in which they are organized in a concept map (Chapter 5).

Step 2. Choosing the assessment task. After the purpose of the assessment is defined, decisions are made concerning the assessment task. Issues that must be taken into account are time constraints, availability of resources, and how much data is necessary in order to make an informed decision about the quality of a student's performance. The literature distinguishes between two types of performance based assessment activities that can be implemented, informal and formal (Airasian, 1991; Mehrens, Popham, & Ryan, 1998; Stiggins, 1987). In the first, the student does not know that the assessment is taking place. Students who are being formally assessed are aware that they are evaluated.

Step 3. Defining performance criteria. After the assessment task is determined, it needs to be defined which elements of the task determine the measure of success of the student's performance. Sometimes, these can be found in so-called job-profiles. Although these resources may prove to be very useful, they often include lists of criteria that may include too many skills or concepts or may not fit exactly. Most of the time, teachers must develop their own criteria. The teacher has to analyse skills or products to identify performance criteria upon which to judge achievement. Quellmalz (1991) offers a set of specific guidelines for the development of quality performance criteria. Criteria should be significant, specifying important performance components, represent standards that would apply naturally to determine the quality of performance when it typically occurs. The criteria must be communicated clearly to and be able to be understood by all involved. Communicating information about performance criteria provides a basis for the improvement of that performance.

Step 4. Creating assessment forms. As opposed to most traditional forms of testing, performance assessments do not provide clear-cut right or wrong answers. The performance is evaluated in a way that allows for informative scoring on multiple criteria. This is accomplished by creating assessment forms. In these forms teachers determine at what level of proficiency a student is able to perform a task or display knowledge of a concept. For example, the different levels of proficiency for each criterion can be defined. Using the information of the assessment form, feedback is given on a student's performance either in the form of a narrative report or a grade. A criterion-referenced qualitative approach is desirable, whereby the assessment will be carried out against the previously specified performance criteria. An analytic or holistic judgment then is given on the basis of the standard the student has achieved on each of the criteria.

Designing a peer assessment

In the previous sections, the concepts of student involvement and performance assessment were described. To design quality performance assessments, a design procedure was presented. When a teacher has passed through this procedure, study tasks can be designed in which students are prepared for the performance assessment. These study tasks are directly related to the performance assessment task at the end of the course.

Peer assessment is defined as an activity to involve students in assessments. This implies that students evaluate the performances of peers at the end of a course unit and provide constructive feedback. Nevertheless, conducting a peer assessment is - like any other skill - a skill in which students need to be trained. It cannot be assumed beforehand that students are experienced in peer assessment practices. If peer assessment is considered as a complex skill in which students should be supported, it is advisable to design a peer assessment for students in which they are assessed on their assessment skills. A basic decision in this dissertation was to embed training in peer assessment skills in courses in which students are expected to criticize each other on their course-content related performances. It was decided to use the Stiggins' approach (1987) to design a peer assessment in which students have to practice their skills to assess:

Step 1. The purpose of a peer assessment. The purpose of a peer assessment is that students are able to assess a peer. This requires that the student is capable of making arrangements in which he or she negotiates with students of similar status about the design and appropriate criteria of specific study tasks and performances. The student takes the responsibility to make critical judgements about the performances of a peer applying the appropriate criteria.

Step 2. Choosing the format of peer assessment. To describe the skill to assess, the method of skill decomposition is applied, to identify constituent skills (Van Merriënboer, 1997). The task to peer assess is broken down into separate skills and these skills are practiced one at a time, before being recombined and practiced as a complete task. In Figure 1 the skill of peer assessment is modelled. Each constituent skill of the peer assessment is further described (see Appendix 1). Data for this decomposition were gathered through literature review and feedback from experts in the area of peer assessment.

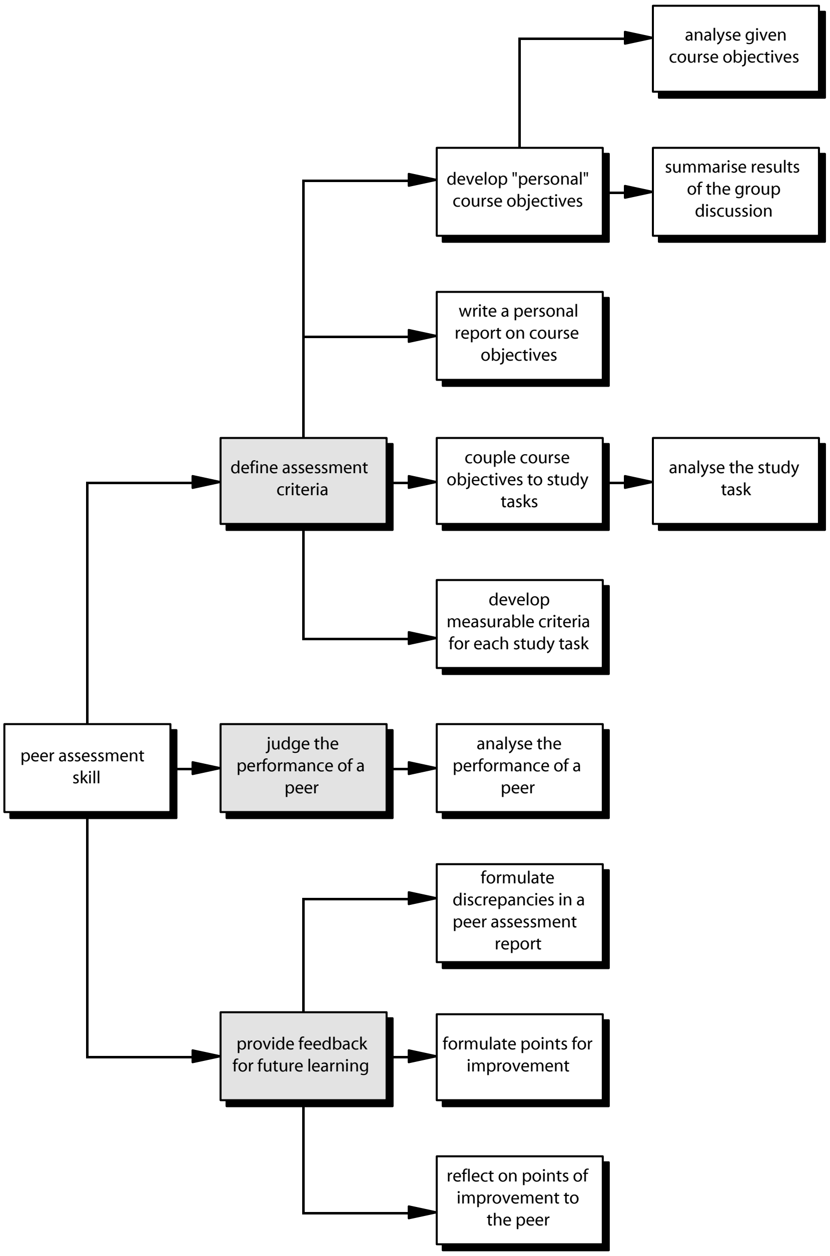

analyse given course objectives
develop "personal" course objectives
summarise results of the group discussion
write a personal report on course objectives
define assessment criteria
couple course objectives to study tasks
analyse the study task
develop measurable criteria for each study task
peer assessment skill
judge the performance of a peer
analyse the performance of a peer
formulate discrepancies in a peer assessment report
provide feedback for future learning
formulate points for improvement
reflect on points of improvement to the peer

Figure 1. Skill decomposition peer assessment

The horizontal relationship illustrates which more specific skills are necessary in order to be able to perform the skill under consideration. The vertical relationship illustrates which other skills are necessary to be able to perform the peer assessment skill. Based upon this, peer assessment tasks are designed that are derived from the skills presented in the model. The peer assessment tasks together build up the training. These tasks are directly coupled to the study tasks in a course. In this way, students are simultaneously and systematically guided in their assessment skills and content related skills.

The task that is object for analysis in our research is writing an assessment report about the performance of a peer at the end of the course. Writing the assessment report is the final peer assessment task, while the embedded peer assessment tasks have a more supportive function in developing the skills that are conditional for conducting a peer assessment. In the empirical studies reported in Chapters 4, 5 and 6, the researchers assess the quality of the assessment reports. The teachers assess the performance assessments that were directly related to the course content.

Step 3. Defining peer assessment criteria. Based on the skills presented in the model, criteria are defined for a good assessment report. Expert assessments are used to determine these criteria. After intense discussion, criteria are determined regarding the use of adequate criteria, giving feedback and the style of a written assessment report. Because peer assessment is addressed as an individual activity for which students individually receive study points, it is expected that certain rating errors like friendship marking, resulting in over-marking; collusive marking, resulting in a lack of differentiation within groups; decibel marking, where individuals who dominate groups get the highest marks; and parasite marking, where students fail to contribute but benefit from group marks, will stay out.

Step 4. Creating a peer assessment rating form. In the experiments, students have to write a qualitative assessment report about a performance of one or more peers on a blank peer assessment form. A rating form is developed by the researchers to analyse the quality of the peer assessments that were written by the students. This rating form is designed based on the criteria for a good assessment report. Research assistants rated each peer assessment in a quantitative way. Six variables are included in the rating form consisting of 36 dichotomous items. These variables are: the use of criteria (1), positive comments (2), negative comments (3), constructive comments (4), posing questions (5), and structure (6). Teachers did not use this rating form. It was only designed for research purposes in the empirical studies.

Skill acquisition and the training of assessment skills

Skill acquisition is a complex issue on its own. Most theorists agree that there are three distinct stages in learning a skill. The specific labels for these stages vary from author to author (Anderson & Lebière, 1998; Fitts, 1964; Lane, 1987; Van Merriënboer, 1997). In the first phase, labelled as the cognitive phase by Fitts (1964) or the declarative phase by Anderson and Lebière (1998), the student collects facts, background information, and general rules related to the skill. Performance is slow and effortful, requiring the student's full attention. During this phase the student is very dependent on external cues, prompts, feedback and information about the skill. The student leaves this phase with a basic understanding of the task requirements and a set of strategies. These are not yet fully elaborated or integrated. In the second stage, Fitts associative phase and Anderson's knowledge compilation phase, the skill is refined or smoothed out (Anderson, 1982; Fitts, 1964; Lane, 1987). Only in the third stage does the student achieve skilled performance.

Given the complexity of the skill to assess, it is almost impracticable to achieve fully skilled performance after one course in which peer assessment tasks are provided. The training in the empirical studies focuses therefore on the first phase. In the previous, it is illustrated which steps need to be taken in designing a task in which students perform a peer assessment. Developing a training for peer assessment implies the design of peer assessment tasks based on the constituent skills presented in the peer assessment model. The activities in these peer assessment tasks attempt to support students in their ability to apply all the skills presented in the model. Hereby it is important to secure integration of all the constituent skills involved (Van Merriënboer, 1997). This implies that students undertake a sequence of peer assessment tasks in which they are confronted with the constituent peer assessment skills defining criteria, judge the performance of a peer, and provide feedback for future learning. From the start of the training, students are notified that these skills are interrelated and necessary for whole task practice.

The peer assessment tasks in the training have several characteristics. First, each peer assessment task is directed at practising at least one constituent skill, without ignoring the whole complex skill. Second, the tasks are embedded in regular study tasks of the course, to foster integration between instruction and assessment and to increase the authenticity of the peer assessment. The regular study tasks are part of a course. This course encloses a period varying of six to eight weeks, whereby the peer assessment tasks take approximately four to eight hours a week. Third, the training is aimed at the development from novice to less novice in assessment. As is said previously, it would be too ambitious to focus on expert assessors after one course. At last, the activities are characterised by interactive discussions among students and between students and teacher. This requires an attitude towards learning tasks in which collaboration plays an important role. The next section particularly addresses the relationship between collaborative learning and peer assessment.

Collaborative learning and peer assessment training

With regard to the training of peer assessment skills, the peer assessment tasks must be authentic and are best learned through cognitive apprenticeship on the part of the student in a rich environment. In this dissertation, collaborative learning is an important aspect of the training of peer assessment skills. It is intrinsic to the skills in the model presented in Figure 1 that peer assessment is learned by collaborative activities. It is therefore useful to identify the related characteristics of collaborative learning that are of importance in learning to assess peers.

The concept of collaborative learning, the grouping and pairing of students for the purpose of achieving an academic goal, has been widely researched and advocated throughout the professional literature. The term ‘collaborative learning’ refers to an instructional method in which students work together in small groups toward a common goal. The students are responsible for each other’s learning as well as their own. Thus, the success of one student helps other students to be successful. Proponents of collaborative learning claim that the active exchange of ideas within small groups not only increases interest among the participants but also promotes critical thinking. According to Johnson and Johnson (1986), there is persuasive evidence that cooperative teams achieve higher levels of thought and retain information longer than students who work quietly as individuals. The shared learning gives students an opportunity to engage in discussion, take responsibility for their own learning, and thus become critical thinkers (Totten, Sills, Digby, & Russ, 1991).

For effective collaborative learning in peer assessment activities, three issues are important, social interaction, individual accountability, and positive interdependency (Slavin, 1989).

Social interaction

In peer assessment tasks, students will negotiate about criteria for performance. To establish grounded criteria, interactions with others are desirable where multiple perspectives on reality can be made more explicit. This includes orally explaining how to solve problems, teaching one's knowledge to others, checking for understanding, discussing concepts being learned, and connecting present with past learning. Each of those activities can be structured into group task directions and procedures. Doing so helps ensure that cooperative learning groups are both an academic support system (every student has one or more peers who help(s) him or her learn) and a personal support system (every student has one or more peers who is/are committed to him or her as a person). It is through promoting each other’s learning that members become personally committed to each other as well as to their mutual goals.

Individual accountability

In order for the group to be successful as a whole, members need to understand that they are each individually accountable for at least one aspect of the task at hand. In peer assessment, students are made individually responsible for an active contribution to group discussions. Individual accountability is the factor that shows that students are learning better cooperatively. After participating in a cooperative study task, students should accomplish the same kind of tasks by themselves. They learn to do something together so that they can do it better when they are alone.

Positive interdependency

When the group's task is to ensure that every group member has learned something, in our case to conduct a peer assessment, it is in the interest of every group member to spend time explaining concepts to their peers. Positive interdependence is successfully structured when group members perceive that they are linked with each other in a way that one cannot succeed unless everyone succeeds. It is the 'glue' that holds the members together.

When positive interdependency is solidly structured, it highlights that (a) each group member's efforts are required and indispensable for group success and (b) each group member has a unique contribution to make to the joint effort because of his or her resources and/or role and task responsibilities. Doing so creates a commitment to the success of group members as well as one's own and is at the heart of collaborative learning.

In peer assessment, as it is defined in this dissertation, it is a question of role interdependence, where the specific roles of assessor and assesee are assigned to the students. The interdependence occurs when one student receives feedback from a peer and that the student is responsible for giving feedback to another peer. In this situation a win-win relationship can be established. If a student fails in giving feedback, the assessee will be the one who suffers. The way to ensure that positive interdependence occurs lies with the promotive interactions between group members, in which students have to report their feedback to the peers in the group.

An integrated framework for training assessment skills

In the previous sections, it was stressed that several concepts are important for the empirical studies. In Figure 2, it is illustrated how these concepts are integrated in a framework that underlies particularly the design of the studies reported in Chapters 4, 5, and 6.

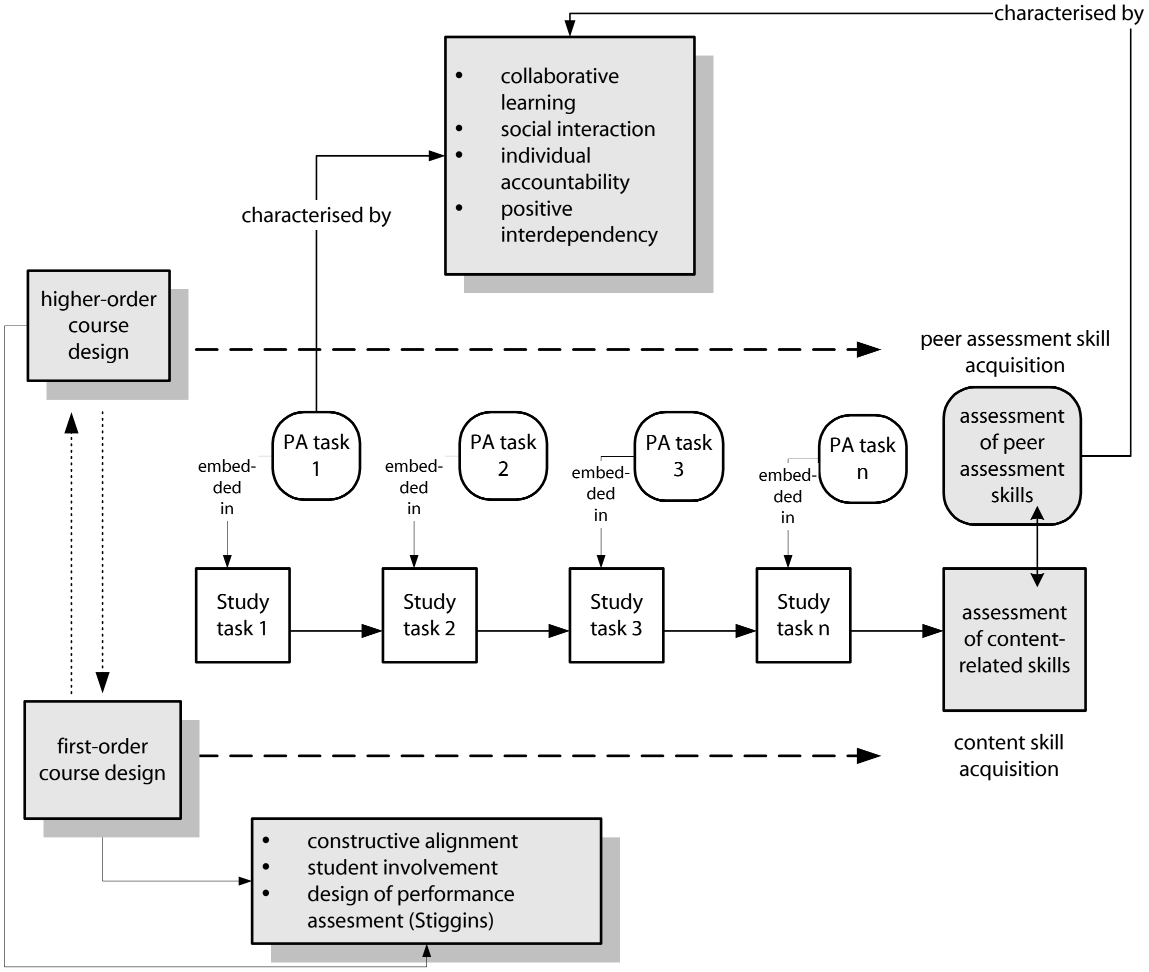

characterised by
collaborative learning
social interaction
individual accountability
positive interdependency
characterised by
higher-order course design
peer assessment skill acquisition
PA task 1
PA task 2
PA task 3
PA task n
embed-ded in
assessment of peer assessment skills
Study task 1
Study task 2
Study task 3
Study task n
assessment of content-related skills
first-order course design
content skill acquisition
constructive alignment
student involvement
design of performance assesment (Stiggins)

Figure 2. Student involvement and course design for powerful learning environments – an integrated framework

Overviewing the concepts discussed in the prior sections, it can be concluded that there are two parallel paths, illustrated by the shaded arrows. In the 'first-order course design path' students are guided in the acquisition of content related skills through study tasks, with the aim to meet the criteria for the content based performance assessment. The second path is the 'higher-order course design path', in which students are supported in the acquisition of peer assessment skills, by means of peer assessment tasks (PA-tasks). These peer assessment tasks, which are superposed on the regular study tasks, are characterised by collaborative learning, more specific by social interaction, individual accountability and positive interdependence. Students thus are guided towards two assessments: a peer assessment (peer assessment skill acquisition) and a content-related assessment (content skill acquisition).

The two paths are integrated (see the two dotted arrows), in other words, the peer assessment tasks are completely embedded in the study tasks of the course, because the content of the study tasks provide input for the peer assessment tasks. The first-order and higher-order course design are the basic elements of the framework, and are defined from the theory of student involvement, the constructive alignment theory and the design principles of Stiggins (1987). At the end of a course, students have to carry out a performance assessment, which is subsequently object of the peer assessment. Both the quality of the peer assessment skill and the content related skill are examined in the studies.

Summarizing the theoretical framework that is made operational in the an integrated framework presented in Figure 2, two basic assumptions were directive for particularly Chapters 4, 5 and 6. The first assumption is that conducting a peer assessment is a complex skill, in which students have to be guided from the moment they enrol in higher education. Before putting students into the role of assessor, it is a prerequisite that students understand which skills are involved while making a judgment of themselves or a peer. Students need explicit training in assessment techniques, to make reliable and acceptable assessment reports (Boud, 1990; Hanrahan & Isaacs, 2001). A peer assessment model provides the basis for designing a training in assessment skills. In this model peer assessment is interpreted as a *learning activity*, instead of only a scoring or ranking tool. This learning activities are partly characterised by collaborative learning in which social interaction, individual accountability and interdependency are guaranteed.

A second 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, Popham, & Ryan, 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 a training of 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. Or as Stiggins stated: "Once students internalise performance criteria and see how those criteria come into play in their own and each other's performance, students often become better performers" (1991, p. 38).

Research context: Teacher Education

The four empirical studies presented in Chapters 3, 4, 5, and 6, are conducted within the context of teacher education. Although performance assessment and peer assessment is of importance in higher education in general, the context of teacher training offers a well-grounded rationale for implementing performance assessment and peer assessment.

The Dutch Primary Teacher Training Colleges have been using a vocational training profile (LPC, 1995) for several years for the design of their curricula. This profile consists of 41 skills that are categorised in 10 domains. The skills represent the overall accepted knowledge, proficiency and attitudes a primary school teacher needs to acquire. To make judgments about the level of development of these skills, performance assessments are very useful, especially assessments which represent authentic situations from the field.

The skill to assess the work of peers is a specific skill of the vocational training profile of primary school teachers, and is important for the teaching domain for several reasons. First, peer assessment fosters reflection. Discussion about reflection is an ongoing issue in teacher education (e.g., Korthagen, 1985; Newman, 1996; Reilly Freese, 1999; Richert, 1999). Reflection skills are conditional for making reliable judgments about peers' work, but peer assessment can also be helpful in fostering reflection skills. Second, teachers have to work together, learn from each other and become a member of a learning organisation (Verloop & Wubbels, 2000). 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. The peer assessment skill is one of such skills. Third, as prospective teachers of children in primary schools, student teachers have to learn how to make critical judgements about the performance of peers, and later on about performances of children. The student teachers will be assessors in their own classroom. They have to be able to design assessments. A last 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, Rust, & Gibbs, 1994). Training in peer assessment skills stimulates that this mutual influence takes place at a professional level.

Structure of the dissertation and research questions

Within the framework of student involvement, constructive alignment, and instructional design, the main aim of this dissertation is to investigate the effects of an embedded training in peer assessment skills on students' performance in their peer assessment skills and content based skills. The training of these skills is embedded in courses that are systematically designed according to the constructive alignment approach, which embraces elements of powerful learning environments.

This dissertation contains two parts. The first part is labelled preliminary research and encloses Chapters 2 and 3. The second part reports the empirical studies that are designed according to the integrated framework. Three experimental studies are conducted, that are reported in Chapters 4, 5, and 6.

The preliminary research consists of a literature review and a pilot study. Chapter 2 reports a literature review in which 62 self-, peer,- and co-assessment studies are analysed. The goal of this study is to make an overview of ways students are involved in assessment procedures. The literature study focuses on the following two research questions:

1. How are self-, peer- and co-assessment applied in higher education?
2. What are the effects of the use of these forms of assessment on the quality of the learning environment?

In Chapter 2 a pilot study is reported, that was conducted within a problem based learning context. This pilot study included two substudies. In the first study the population consisted of 27 university students subscribed to a fourth year course in educational sciences using problem based learning. The population in the second study consisted of 51 fourth year students of a Primary Teacher Training College, also using problem based learning. In both studies, the students worked collaboratively in groups. In the pilot study, a quantitative approach of peer assessment was chosen, which means that students assessed the peers in their group based on a scoring method. They assessed their peers on several criteria concerning working in a group after a course period. Three research questions were addressed:

1. Are peer ratings in problem based learning groups reliable?
2. Do students have idiosyncratic (i.e., personal) strategies in peer assessment?
3. What are students' experiences with peer assessment and problem based learning?

In order to examine whether the peer ratings in problem based learning groups are reliable, the data of the peer assessments were analysed. The reliability of the ratings was estimated within the framework of generalizability theory (Brennan, 1983). To investigate if raters have personal strategies a Q-analysis was applied. A two-part evaluation questionnaire was developed to measure students' experiences.

Regarding the literature review and the results of the pilot study reported in the first part of the dissertation, it was decided to design studies in which the assumptions described in the prior sections were directive. A peer assessment model was designed and a training in assessment skills was developed. This training was intended to be fully integrated in existing courses. A redesign of these courses appeared to be necessary, to explicitate the course objectives and the criteria for the performance. Five research questions were explored in depth in three studies that are reported in Chapters 4, 5 and 6:

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

Chapter 4 reports an experiment that was carried out with 93 second-year student teachers. Half of the group was trained in peer assessment skills (experimental groups) and the other half was not (control groups). By working on the peer assessment tasks that were embedded in regular study tasks, student teachers in the experimental group had to define performance criteria within groups of ten students. The training was integrated in a redesigned course. The quality of the peer assessments written by the students were analysed with rating forms. The effects of the performance assessments were also measured. A student questionnaire was developed to measure students perceptions before and after the course. It was hypothesized that content domain influences these perceptions (research question 4). Interviews with students and teachers provided additional qualitative data (research question 4 and 5).

Chapter 5 reports a similar experiment, but in this study students were trained in several assessment skills, instead of only the skill of defining criteria. The experimental groups were trained in defining performance criteria, giving feedback and writing assessment reports. This was again established through peer assessment tasks that were embedded in a redesigned course. The research questions 1, 2 and 3 were measured similar to the analyses in the study reported in Chapter 4.

The findings of the third and fourth study set the design and goals of the fifth and final study, which is reported in Chapter 6. In this study, 110 first-year student teachers were longitudinally trained in peer assessment skills within three courses on mathematics. After each course, students wrote a reflection paper, which was assessed by a peer. Students were allowed to improve their first reflection paper after the second and third course, based on peer feedback. The teacher assessed the reflection paper not until after the third course. In the two prior studies students had no opportunity to improve their performance based on peer feedback. Research questions 1, 2, and 3 were studied.

In Chapter 7, the results of the five studies are integrated and discussed. Four important considerations for implementing peer assessment are presented. The limitations of the studies are outlined, and new ideas for future research are provided. Some important guidelines for higher education practices in general end this dissertation.

The studies that are presented in Chapter 2 to 6, have been published in international journals or have been submitted for publication. These chapters can be read separately, but are successive based on the findings of each study.

References

Airasian, P.W. (1991). *Classroom assessment*. New York : McGraw-Hill.

Anderson, J. R., & Lebière, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Arter, J. (1996). Using assessment as a tool for learning. In R. Blum, & J. Arter (Eds.), *Student performance assessment in an era of restructuring* (pp. 1–6). Alexandria, VA: Association for Supervision and Curriculum Development.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*, 347-364.

Biggs, J.B. (1999). *What the student does: Teaching for quality learning at university*. Buckingham: Open University Press.

Biggs, J.B. (2001). The reflective institution: Assuring and enhancing the quality of teaching and learning. *Higher Education, 14*, 221-238.

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* (pp. 3-29). Boston: Kluwer Academic Press.

Boud, D. (1990). Assessment and the promotion of academic values. Studies in *Higher Education, 15,* 101–111.

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

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

Brennan, R.L. (1983). *Elements of generalizability theory*. Iowa City, IA: American College Testing Program.

Brown. J.S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*, 32-42.

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

De Corte, E. (1995). Fostering cognitive growth: a perspective from research on mathematics. *Educational Psychologist, 30,* 37-46.

De Corte, E. (1996). Actief leren binnen krachtige leeromgevingen. [Active learning within powerful learning environments]. *Impuls, 26*, 145-156.

Dochy, F.J.R.C. (1992). *Assessment of prior knowledge as a determinant for future learning*. Utrecht/London: LEMMA, Jessica Kingsley.

Dochy, F.J.R.C., & McDowell, L. (1997). Assessment as a tool for learning. *Studies in Educational Evaluation, 23*, 279–298.

Entwistle, N. (2000). *Promoting deep learning through teaching and assessment: conceptual frameworks and educational contexts*. Paper presented at the TLRP Conference, Leicester, November, 2000.

Entwistle, N. J., & Entwistle, A. C. (1997). Revision and the experience of understanding. In F. Marton, D. J., Hounsell, & N. J. Entwistle (Eds.), *The Experience of Learning* (2nd ed.) (pp. 145-158). Edinburgh: Scottish Academic Press.

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

Fitts, P.M. (1964). Perceptual-motor skill learning. In A.W. Melton (Ed), *Categories of human learning*. New York, NY: Academic Press.

Frederiksen, N. (1984). The real test bias: influences of testing on teaching and learning. *American Psychologist, 3*, 193-202.

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

Glaser, R. (1990). Toward new models for assessment. *International journal of educational research, 5*, 457-483.

Glaser, R., & De Corte, E. (1992). Preface. In: F. Dochy (Ed.) Dochy, F.J.R.C. (1992). *Assessment of prior knowledge as a determinant for future learning*. Utrecht/London: LEMMA, Jessica Kingsley.

Hall, K. (1995, February). *Co-assessment: participation of students with staff in the assessment process: a report of work in progress*. Invited paper presented at the Second European Electronic Conference on Assessment and Evaluation. EARLI-AE list European Academic & Research Network (EARN) (EARLI_AE on Listserv@nic.surfnet.nl).

Hanrahan, S., & Isaacs, G. (2001). Assessing self- and peer assessment: the students' views. *Higher Education Research and Development, 20*, 1, 53-70.

Hibbard, K. M., Michael, K., et al. (1996). *A teacher's guide to performance-based learning and assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Johnson, R. T., & Johnson, D. W. (1986). Action research: Cooperative learning in the science classroom. *Science and Children, 24*, 31-32.

Korthagen, F.A.J. (1985). Reflective teaching and pre-service education in the Netherlands. *Journal of Teacher Education, 36*, 11-15.

Lane, N.E. (1987). *Skill acquisition rates and patterns: Issues and training implications*. New York: Springer-Verlag.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*, 5-21.

Lockwood, F. (1995). Students' perception of, and response to, formative and summative assessment material. In F. Lockwood (Ed.), *Open and Distance Learning Today* (pp. 197-207). London: Routledge.

LPC (1995). *Beroep in beweging. Beroepsprofiel leraar primair onderwijs* [Profession in action. Vocational training profile for the primary school teacher]. Utrecht: Forum Vitaal Leraarschap.

Martens, R., & Hermans, H. (2000). Internet based formative prior knowledge assessment. *Studies in educational evaluation, 26*, 245-258.

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

Newman, S.J. (1996). Reflection and teacher education. *Journal of Education of Teaching, 22*, 297-310.

Portelli, J. P. (1993). Exposing the hidden curriculum. *Journal of Curriculum Studies, 25*, 343-358.

Quellmalz, E.S. (1991). Developing criteria for performance assessments: The missing link. *Applied Measurement in Education, 4*, 319-332.

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

Richert, A.E. (1999). Teaching teachers to reflect: a consideration of programme structure. *Journal of Curriculum Studies, 22*, 509-527.

Sambell, K., & McDowell, L. (1998). The construction of the hidden curriculum: messages and meanings in the assessment of student learning. *Assessment and Evaluation in Higher Education, 23*, 391-402.

Sambell, K., McDowell, L., & Brown, S. (1997). "But is it fair?": an exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation, 23*, 349–371.

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

Snyder, B. R. (1971). *The hidden curriculum*. New York: Alfred A Knopf.

Stiggins, R. (1987). Design and development of performance assessment. *Educational Measurement: Issues and Practice, 6,* 33-42.

Stiggins, R. (1991). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice, 10*, 7-12.

Stiggins, R.J. (1994). *Student-centered classroom assessment*. Columbus, OH: Macmillan.

Totten, S., Sills, T., Digby, A., & Russ, P. (1991). *Cooperative learning: A guide to research*. New York: Garland.

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* (pp. 19-32). Leuven-Apeldoorn: Garant.

Wiggins, G. (1989). Teaching to the (authentic) test. *Educational Leadership, 46*, 41-47.

Wiggins, G. (1990). *The case for authentic assessment*. ERIC Digest. Washington, DC: ERIC Clearinghouse on Tests, Measurement, and Evaluation.

Chapter 2

Creating a learning environment by using self-, peer-, and co-assessment: a literature study[1]

Abstract

To develop the skills and competencies required in professional organizations, students have to reflect on their own behaviour. Many current assessment practices in higher education do not answer this need. The recent interest in new assessment forms, such as self-, peer-, and co-assessment, can be seen as a means to tackle this problem. In this chapter, a review of the literature provides answers to two questions: (1) How are self-, peer-, and co-assessment applied in higher education? and (2) What are the effects of the use of these forms of assessment on the quality of the learning environment? Analyses of 62 studies showed that self-, peer-, and co-assessment can be effective tools in developing competencies needed as a professional. These forms of assessment are often used in combination. Implementation of these forms of assessment accelerates the development of a curriculum based on competencies (knowledge as a tool) rather than knowledge (as a goal) and leads towards the integration of instruction and assessment in higher education. As such, this development of a learning environment contributes to the education of responsible and reflective professionals.

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

Introduction

It is widely recognised that the main goal of professional higher education is to help students to develop into 'reflective practitioners' who are able to reflect critically upon their own professional practice (Falchikov & Boud, 1989; Kwan & Leung, 1996; Schön, 1987). Students in modern organisations should be able to analyse information, to improve their problem-solving skills and communication, and to reflect on their own role in the learning process. The need for lifelong learning in modern society will increase (Sambell & McDowell, 1998) as it becomes more recognised that the acquisition of knowledge and skills cannot be restricted to the phase of initial education; rather, it has to be a process continuing throughout one's entire working life. Traditional testing methods do not fit goals like lifelong learning, reflective thinking, being critical, evaluating oneself and problem solving (Dochy & Moerkerke, 1997). Alternatives in assessment have received much attention in the last decade and several forms of more authentic assessment have been introduced in higher education (Birenbaum & Dochy, 1996). The skills of self-, peer-, and co-assessment are important in the development of autonomous, responsible and reflective individuals (Sambell & McDowell, 1998). Assessment procedures should not only serve as a tool for crediting students with recognised certificates, but should also be used to monitor progress and, if needed, direct students to remedial learning activities. Research showed that the nature of assessment tasks influences the approaches to learning which students adopt (Beckwith, 1991). The existing assessment approaches can have effects contrary to those desired. Eisner (in Boud, 1995) identified the features of the new assessment in education:

- Assessment tasks need to reflect the tasks that students will encounter in the world outside schools, not merely those limited to the schools themselves;
- Assessment tasks should not be restricted to the solutions that students formulate, but also reveal how students go about solving a problem;
- Assessment tasks should reflect the values of the intellectual community from which tasks are derived;
- Assessment tasks need not be limited to a solo performance;
- Assessment tasks should have more than one acceptable solution to a problem and more than one acceptable answer to a question;
- Assessment tasks should have curricular relevance, but not be limited to the curriculum as taught;
- Assessment tasks should permit the student to select a form of representation that he or she chooses to display what has been learned.

The view that the assessment of students' achievements is something which happens at the end of a process of learning is no longer widespread. Assessment is now represented as a tool for learning (Arter, 1996; Dochy & McDowell, 1997). The present study focuses on one new dimension of assessment innovation, namely, the changing place and function of the assessor. Alternative assessment also questions if the assessor must be the teacher, and whether students can be introduced as assessors in different settings. This study provides a literature review which focuses on forms of self-, peer-, and co-assessment from the points of view of their applicability in higher education and their effect on the learning environment.

This study specifically aims at finding out how self-, peer-, and co-assessment are used in research and educational practice. Two research questions are addressed.

1. How are self-, peer-, and co-assessment applied in higher education?
2. What are the effects of the use of these forms of assessment on the quality of the learning environment?

Method

In order to answer the research questions, a literature search was conducted using the following sources:

- The database of the Educational Resources Information Center (ERIC). This database was searched, according to the key words 'self assessment', 'peer assessment' and 'co-assessment'. The search was conducted for publications from 1987 to 1998.
- The database of Current Contents on Disk. The years 1996 to 1998 were searched in order to find recent studies which had not yet been included in ERIC. The same key words were used as in the ERIC search.
- Through the so-called snowball method, the references in all the above materials were checked for other studies.

From these sources, all abstracts of the hits were printed and their content was analysed. Three criteria were used to determine whether literature would be included in our study. First, the assessment form had to be predominantly self-, peer-, or co-assessment. Portfolio assessment and performance assessment, for example, were not central themes, although there was often a strong relationship with self-, peer-, and co-assessment. Second, the literature about the relationship between assessment and learning/reflection was included in this search to find answers to the second research question about the learning environment. Third, the subjects of study had to be students in higher education. For instance, studies dealing with peer assessment of university personnel were excluded.

The search of ERIC resulted in 191 articles. The articles reporting the application of assessment in higher education were selected for this study by analysing titles and abstracts. This resulted in 34 articles. The search of Current Contents on Disks resulted in 53 articles. After selection based on the above criteria, 10 articles were retained. The snowball method led to the selection of another 18 articles. In total, 62 articles were selected for further analysis.

The 62 studies were analysed in detail and recorded in schemes according to the authors and the year reported, the title, the aim of the study, the type and number of subjects, the advantages (of the assessment form discussed), the disadvantages (of the assessment form discussed), the treatment/instrument, the assessment criteria, the value of statistics reported and the conclusions of the study (see Appendix 2 for an example of a scheme).

The studies first were classified into empirical and non-empirical reports. In most of the articles, no statistics were reported. A study was considered empirical when quantitative data were gathered or statistical analyses were reported. Non-empirical studies were included because the information in these reports was too valuable to exclude them from the search.

Then, the studies were classified into self-, peer-, and co-assessment. However, because these forms of assessment are sometimes strongly interrelated, it was decided to divide the studies into five categories: studies of self assessment, studies of peer assessment, studies using self and peer assessment, studies using self-, peer-, and co-assessment, and more general studies about the relationship between assessment and learning or reflection on learning processes. The studies were coded by the three authors independently. Comparing the coding schemes revealed that there were no differences in classifying the studies and showed that there were: 14 studies about self assessment, 10 studies about peer assessment, 7 studies about self/peer assessment, 19 studies about self/peer/co-assessment, and 12 studies about the relationship between assessment and learning/reflection.

In the present study, a narrative review of the literature is used. This form of conventional literature review implies careful reading of separate studies and integrating them. Of course, because patterns in the results are sought, this integration is an intuitive undertaking (Knoors et al., 1995; Slavin, 1986). Only one of the selected studies included a control group and an experimental group. As a consequence, a statistical meta-analysis could not be done.

How are self-, peer-, and co-assessment applied in higher education?

In this section, the results related to the first research question are described using four separate subsections for the different combinations of self-, peer-, and co-assessment. These subsections begin with a definition. Then, the main findings are presented and the way in which the assessment forms are used in practice is outlined. Finally, a conclusion ends each subsection.

Self assessment

Self assessment refers to the involvement of learners in making judgements about their own learning, particularly about their achievements and the outcomes of their learning (Boud & Falchikov, 1989). Self assessment is not a new technique, but a way of increasing the role of students as active participants in their own learning (Boud, 1995), and is mostly used for formative assessment in order to foster reflection on one's own learning processes and results (Sluijsmans, Dochy, & Moerkerke, 1998).

Boud and Falchikov (1989) classified the literature of self assessment under three headings: conceptual, practical qualitative, and quantitative. One of the most important parts in the conceptual framework is the literature about the reflective practitioner (Schön, 1987). The practical qualitative group includes the processes involved in introducing and using self assessment in different situations. The quantitative group focuses on studies of student self-ratings compared to the ratings of students by teachers. Boud and Falchikov (1989) analysed studies from 1932 to 1988 and reported the over-rating and the under-rating of students. They related these findings to the different abilities of students. The finding was that good students tended to under-rate themselves and that weaker students over-rated themselves. Students in higher levels of classes could better predict their performance than students in lower levels of classes.

Griffee (1995) also investigated the question whether there is a difference in student self assessment between first-year, second-year and third-year classes in a university department. The general answer to this question was that there was no difference. All classes tended to rate themselves lower at the beginning of the school year and higher at the end of the year. As the semester progressed, students gained more confidence in their ability to perform. Another explanation for the fact that there was no difference between the self assessments of the three classes was the teacher intervention during the year.

Several studies obviously show that the ability of students to rate themselves improves in the light of feedback or development over time (Birenbaum & Dochy, 1996; Boud & Falchikov, 1989; Griffee, 1995). Moreover, students' interpretations are not just dependent on the form of the assessment process, but on how these tasks are embedded within the total context of the subject and within their total experience of educational life.

In educational practice, different instruments are used for self assessment. Harrington (1995) used three different self assessment instruments. One was simply a listing of abilities with definitions and directions to indicate those areas that you feel are your best or strongest. A second approach is to apply a Likert scale to a group of designated abilities. For example, "in comparison to others of the same age, my art ability is excellent, above average, average, below average, or poor". Another approach is, for each ability, to provide different examples of the ability's applications so that individuals rate their performance level from high to low, and subsequently these are summed to obtain a total score. The self assessment forms that Harrington described are cheaper and less time intrusive than traditional ways of assessing students (Nevo, 1995).

An electronic interactive advice system for self assessment is provided by Gentle (1994). The aim of this system is to see how accurately students are able to assess their own work without the involvement of their supervisor. The system is based on question-and-answer screens for 38 skills. These skills are arranged into the four sections of (1) approach to the project — effort, time management, etcetera., (2) quality of day-to-day work, (3) quality of the description of the work and (4) quality of presentation. The procedure is as follows. "The user moves a cursor on a continuous scale of performance on that aspect of the work. The middle and end points on the scale are picked out by written statements to help the user and there is also a full advice screen available for each question. This feature makes this system much more than just an assessment program, since it includes large tranches of practical assistance, useful at any point in the project work. The output also provides much more than a mark; the five best and the five weakest points, selected by their weighted contribution to the mark, are extracted and displayed" (Gentle, 1994, p. 1159). Results of the use of the system show that students can assess themselves to within five percentage points. Students become more aware of the quality of their own work. They can predict their own mark and, while they are doing this, they reflect on their behaviour (reflective practitioner). Because the students reflect more often on their work, this will lead to a higher quality of the products. According to Gentle, the system is less time consuming than the conventional self assessment because the supervisor has a minor part in the assessment.

In research conducted by Hassmén et al. (1996), 128 women learned the correct answers on a specific task by either performing or observing. Participants took either a performance or a written test, with or without making self assessments about how sure they were that their selected answer was correct. Findings of the research support the hypothesis that those participants who engage in overt self assessment while learning obtain a higher percentage of correct responses during learning trials than those who learn without self assessments.

This is also illustrated in a study reporting successful language learning. McNamara and Deane (1995) designed a variety of activities that foster self assessment. Three of them are writing letters to the teacher, keeping a daily language learning log, and preparing an English portfolio. These activities can help students to identify their strengths and weaknesses in English, to document their progress, and to identify effective language learning strategies and materials. They also become aware of the language learning contexts that work best for them, and they establish goals for future independent learning. The idea of self assessment for use with portfolios is described by Keith (1996), who suggests self assessment assignments which ask students to report on their own learning. Assignments include sharing preconceptions about teaching and learning, comparing goals, creating a community of learners, generating student explanations and improving communication, group quizzes, challenging thinking dispositions, post-test evaluations and collaborative assessing. The roots of all the described assignments lie in collaborative learning. Keith finds that the most influential variable for effective learning is the amount of meaningful energy that the students put in. The assignments have to encourage students to feel responsible for their own learning.

Anderson and Freiberg (1995) used an audiotape self assessment instrument for student teachers to reflect on their teaching. This instrument - called the Low Inference Self Assessment Measure (LISAM) - has been developed to let student teachers analyse their instruction. Ten secondary student teachers completed four stages in the study. In the first stage, students learned to record themselves during a lesson. In the second stage, students were trained to analyse their own audiotapes. In the third stage, findings and suggestions for effective use of the LISAM were discussed. The students set goals for future use of the self assessment instrument. In the last stage, there was an interview with every student teacher. Anderson and Freiberg describe three reasons why the LISAM is practical and effective: (1) the use of LISAM makes student teachers more independent, provides feedback and stimulates them to reflect on their own teaching; (2) student teachers can practice LISAM immediately; and (3) the LISAM teaching behaviours are observable and alterable.

Generally, next to addressing the instruments used for self assessment, the content could be addressed. At the content level, it is striking that self assessments are mostly used to foster skills and abilities (in addition to knowledge) and that assessments are used in a formative or diagnostic way (Birenbaum & Dochy, 1996). For example, students at Alverno College have to develop problem solving skills as one of the eight abilities in order to graduate (Loacker & Jensen, 1988). At the heart of the educational process at Alverno stands assessment, which is seen as a natural part of encouraging, directing and providing for development of abilities. Because self assessment is required to be integrated with students' problem-solving process, students show increasing understanding of inter-relationships of ability, content and context. Students take responsibility for their learning as a dynamic, continuing process. They gradually internalise their practice of both problem solving and self assessment abilities.

Overall, it can be concluded that research reports positive findings concerning the use of self assessment in educational practice. Students in higher education are well able to self-assess accurately (see Gentle, 1994), and this ability improves with feedback and development over time. Moreover, students who engage in self assessment tend to score higher on tests. Self assessment, used in most cases to promote the learning of skills and abilities, leads to more reflection on one's own work, higher quality of products, responsibility for one's own learning, and increasing understanding of problem solving. Instruments for self assessment vary from Likert scales, ability listings and written tests to portfolios, audiotape assessments or electronic interactive systems.

Peer assessment

Falchikov (1995) defines peer assessment as the process whereby individuals rate their peer(s). This exercise might entail previous discussion or agreement over criteria. It could involve the use of rating instruments or checklists, which have been designed by others, before the peer assessment exercise, or be designed by the user group to meet their particular needs.

More theoretical studies found in the search present goals and forms of peer assessment. Somervell (1993) found that peer assessment engages students in making judgements about the work or the performance of other students. At one end of the spectrum, it could involve them giving feedback of a qualitative nature or, at the other, it might involve them in marking. The assessment could be formative or summative and could form part of a larger scheme in which peer feedback is given prior to self assessment by the recipient. Peer assessment is not only a grading procedure, but also part of a learning process in which skills are developed. Peer assessment can be seen as a part of the self assessment process and serves to inform self assessment. The contribution of other students can be a very useful input into the self assessment process. Students have an opportunity to observe their peers throughout the learning process and often have more detailed knowledge of the work of others than their teachers do (Somervell, 1993). Keaten and Richardson (1992) reported that peer assessment can foster high levels of responsibility among students, but that the students must be fair and accurate with the judgements which they make regarding their peers.

Peer evaluation is also an alternative term to peer assessment (Weaver & Cotrell, 1986). Peer evaluation "emphasizes skills, encourages involvement, focuses on learning, establishes a reference, promotes excellence, provides increased feedback, fosters attendance, and teaches responsibility" (Weaver & Cotrell, 1986, p. 25). Dancer and Dancer (1992) indicate that research studies have not shown the validity of peer rating. Peers are prone to produce ratings based on uniformity, race and friendship if there is no extensive training in peer rating. Based on this assumption, it is sometimes important to determine an individual's contribution to a group project.

Different forms of assessment are distinguished by Kane and Lawler (1978). *Peer ranking* involves each group member rank all of the others from best to worst on one of more factors. *Peer nomination* involves each group member being the highest in the group on a particular characteristic or dimension of performance. *Peer rating* involves each group member rating each other group member on a given set of performance or personal characteristics, using any one of several kinds of rating scales.

The more practically oriented studies focus on one of the main advantages of peer assessment, namely, fairness (construction of assessment criteria, more objective assessment by more subjects, assessment of output and process such as assessing student contribution). Conway et al. (1993) indicate that students found group projects more interesting than traditional methods of teaching. Because the fairness of the assessment was found to be the only negative aspect of this type of working, peer assessment was introduced. First, each group's presentation was assessed by the other members of the group. Secondly, the students assessed the contribution of their fellow group members to the work of the project. The aim of the study was to examine ways in which students can be awarded individual marks, which reflect personal effort, for group projects. Conway et al. found good elements in the scheme of Goldfinch and Raeside (1990) and simplified this scheme by combining the elements which could be very effective. The method which Conway et al. used is not outlined here, but the results showed that students felt that peer assessment is a good method and fair enough. Students felt that they should play a part in the assessment in order to make it more objective.

Cutler and Price (1995) describe an investigation in which presentations and seminars, built into each of the three years of the geography program, are peer assessed against a set of criteria. Self-appraisal forms are also a part of the assessment procedure. The majority of the students were happy and confident about being assessed by their peers. Half of the students felt that their assessment of their peers was accurate. A third of the students thought that they had improved in confidence, organisation of materials and use of voice.

Boud (1995) investigated students' ability to make assessments of themselves and their peers in a first year class entitled 'The Legal System' taught by the second author at the University of New South Wales. A series of instructions was provided for each student to give guidelines for formulating assessment criteria. Once the students had constructed the criteria that they would use, they made an assessment of both their own performance in class and that of the other students in the class. There were two methods of scaling used. Method A involves a scale of 1, 2, 3 ... 10 with 5 as the pass mark. Method B involves a scale of –2, –1, 0, 1, 2 with the rule that there should be equal numbers of students above and below the mean of 0. At the end of the exercise, the self-mark, the peer-mark and the teacher-mark were available for each student on each criterion using both methods. Students found constructing the assessment criteria very useful, they rated themselves more highly than they were rated by their peers, and students rated themselves less highly than the teacher using method A but higher using method B. In general, there was a very high level of agreement between the marks given by peers and those given by the teacher.

Experience from peer assessment revealed that, as a formative assessment method and as a part of the learning process, it can be seen as valuable. Students are more involved, both in the learning and in the assessment process. They find peer assessment fair enough and accurate. However, peer assessment can also involve: friendship marking (resulting in over-marking); collusive marking (resulting in a lack of differentiation within groups); decibel marking (where individuals dominate groups and get the highest marks); and parasite marking (where students fail to contribute but benefit from group marks) (Pond et al., 1995). These problems can be prevented by combining peer assessment with self assessment or co-assessment. This is exactly the reason why the majority of studies involved these combinations of assessment forms. The following sections illustrate this.

Experiences revealed that peer assessment, as a formative assessment method and as a part of the learning process, can be valuable because students are more involved both in learning and in the assessment process and because they find it fair and accurate. Disadvantages of peer assessment, such as friendship marking and decibel marking, are mostly solved by using combinations of peer assessment with self- and co-assessment, as further explained later in this Chapter.

Self and peer assessment

Self and peer assessment are combined when students are assessing peers but the student himself or herself is also included as a member of the group and must be assessed. This combination fosters reflection on one's own learning process and one's learning activities compared to the other members in the group or class.

Because of the disadvantages of peer assessment stated above, almost all studies found on combinations of assessment forms were practically oriented and sought more proof on validity, inter-rater (or inter-peer) reliability, and positive involvement of students in the process.

In a study described by Burnett and Cavaye (1980), fifth-year medical students assessed their peers as part of the examination. They also were asked to assess their own performance. Peer assessment highly correlated with the final grade ($r = 0.99$) and staff-assessment ($r = 0.93$), and self assessments highly correlated with the results of peer assessments ($r = 0.99$). This is a replicated finding (Birenbaum & Dochy, 1996; Falchikov, 1991; McDowell, 1995). As a consequence, friendship ratings should not be taken as too large a problem. Nevertheless, the experience is that in the research on peer assessment the problem lies more in the weakest students who over-rate themselves and are not able to judge their peers correctly. Such scores often behave as statistical outliers. In our investigations therefore, the highest and the lowest peer assessment scores are excluded for each individual in order to calculate the mean scores.

Strachan and Wilcox (1996) describe a peer and self assessment strategy in a third- year microclimatology course. Thirty students were asked to form groups of three persons to do a seminar presentation. The students were informed that they would undergo a peer assessment and that they could develop their own criteria. After this 'brainstorming', a 'group project peer and self assessment sheet' was established. It was also a self assessment sheet because every student had to include himself or herself in the assessment. Each member of each group completed the sheet after handing in their papers. The students also were asked to give written feedback on this way of assessment. Some students preferred written evaluative comments to number ratings and some students felt that this way of assessment was not necessary because you always have a certain responsibility in a group. Strachan and Wilcox recommend, however, that it is important to give the student an active role in the development of assessment criteria. The process is thereby equally important as the product for the quality of learning.

Warkentin et al. (1995) investigated self and peer assessment in a study with 83 undergraduate educational psychology students. Warkentin et al. hypothesized that students taking tests using individual and group assessments would perform better on a post-test based on educational psychology course concepts than students who took the traditional tests (individual examinations). The effects on student knowledge structure representations were examined. The results indicated that there were no significant differences between the two groups on achievement and knowledge structure. Warkentin et al. (1995), however, found that the reactions to the assessment procedure which they used were positive. The students did like the group assessment and thought that it contributed to their learning as they discussed and debated test items.

Sambell and McDowell (1998) studied six cases which included peer and/or self assessment. Students were generally positive towards involvement in the assessment process. Students' awareness that self and peer assessment helped them to develop important skills (e.g. problem solving) was high.

Under- and over-marking in self and peer assessment were investigated by Falchikov (1991). The process of working together on a small-group project was assessed by the group members, namely, seven developmental psychology students. In the study, the development of a self/peer group process assessment checklist was carried out and was designed to compare the assessments of task and maintenance functions (Falchikov, 1991). Students worked on a piece of coursework. The checklist contained 16 task functions and eight maintenance functions. This list was developed with the co-operation of the students; this enabled them to become familiar with the assessment checklist. After finishing the coursework, the students had to rate their peers and themselves on the checklist. They rated the level of activity (high, medium, low) to which each group member including themselves had carried out the 16 functions (group activities). The results showed that there was no consistency in over- or under-marking when comparing self-ratings with peer-ratings. There was also a high level of agreement between peers. Falchikov states that "the process of working in a group is a simple and effective way of moving towards a solution to the problem of evaluation of group work" (p. 15).

Self and peer assessment related to co-assessment

In the prior sections, the use of self assessment, peer assessment and a combination of these two forms was described. One step closer to the current mainstream in traditional educational practice is the assessment procedure in which the tutor plays a significant role in the process. The term 'co-assessment' refers to this situation.

Co-assessment, the participation of the students with the staff in the assessment process, is a way of providing an opportunity for students to assess themselves while allowing the staff to maintain the necessary control over the final assessments (Hall, 1995). Synonyms for co-assessment are 'collaborative assessment' and 'cooperative assessment'. Co-assessment can be used for summative purposes, while self and peer assessment are used in a formative way. Somervell (1993) sees collaborative assessment as a teaching and learning process in which the student and instructor meet to clarify objectives and standards. In this case, the student is not necessarily responsible for the assessment, but the student collaborates in the process of determining what will be assessed and perhaps by whom. It is a true collaboration in so far as both parties work on the shared goal of providing a mutually agreed assessment of the student's knowledge. This entails both parties negotiating details of the assessment and discussing any misunderstandings that exist.

Co-assessment is often related to forms of self and peer assessment. In a single study, combination of self- and co-assessment was found. In Hall's study (1995), the students and staff set the criteria. The process involved a double-sided face-sheet. On the back of this sheet, the students had the opportunity to give their own self assessment of the piece of work and then hand it to the staff member. The staff member used the outside of the sheet to record his or her assessment of the student's work. Then, the staff member turned it over to see whether or not students had chosen to offer their own assessment on the other side. The findings showed that generally the staff member's grade was higher than the student's grade. The experiment which Hall described identifies three purposes of co-assessment to assist student teachers in making role changes from being students to being teachers; to provide insights into the assessment process which may be of use to them in assessing their own students; and to provide a skill-development step towards self assessment.

Many other studies combine self-, peer-, and co-assessment in some way. In the studies, there is a majority of positive experiences for assessing essays and assignments and, to a smaller extent, assessing presentations. Falchikov (1986) and Stefani (1992, 1994) also described studies with collaborative self and peer assessment. The study of Falchikov aimed to implement and evaluate a method of collaborative self and peer assessment. First the tutors set criteria which were ranked in terms of their relative importance. Then students set criteria and tutor-student criteria comparisons were made. An essay marking schedule was drawn up. Students marked their own essays and then each group member and the tutor marked the essays. Self-, peer- and tutor marks were compared. Results show that collaborative and self assessment does appear to be comparable to traditional tutor methods of assessment, while collaborative and peer assessment corresponds less well with either tutor- or self-grading. Stefani (1992) carried out an experiment in collaborative self- and peer assessment involving a first-year undergraduate biochemistry laboratory practical experiment. The students themselves defined the marking schedule for a scientific report. The results show that students have realistic perceptions of their own abilities and can make rational judgements about the achievements of their peers. Many tutors expressed their fears in handing over the assessment to the student. Concerning the evaluation of the learning benefits, almost every student said that the scheme made them think more, learn more and was challenging.

Freeman (1995) conducted a peer assessment experiment with 210 final-year undergraduate business students who were divided into 41 teams, with each team completing two of the four assessable tasks. The presentation, one of the two tasks, was chosen by staff to experiment with a peer assessment worth 25% of the overall grade. In the first week of semester, each student was given the presentation marking and feedback sheet with 22 items, with eight items related to the content and 14 related to the presentation, weighted 60% and 40%, respectively. In the results of the team presentations rated by staff and peers, Freeman found that the quality of the presentations was very high. There was no statistically significant difference between the average staff ratings and average peer ratings. However, students tended to under-mark the good presentations and over-mark the poor presentations.

Longhurst and Norton (1997) designed a study to investigate how accurately 67 second-year psychology students would be able to assess their own essays and thereby ascertain whether or not they understand what taking a deep approach in their essays actually means. Student grades were compared with tutor grades. The students were asked to rate themselves on tutor-specified criteria which were designed to measure a deep approach, essay grade and level of motivation in completing one specified essay. The tutor did not see these self assessments because the self assessment sheet was removed from each essay. The tutors also marked the essays on the deep-processing criteria. The tutor grade for the essay highly correlated with each of the five criteria. There also was a positive correlation between student and tutor grades ($r = 0.43$). Overall, students were very accurate in grading their own essays, but less accurate in assessing their own deep processing. Less motivated and weaker students appeared to be less clear on understanding the individual criteria.

Oldfield and Macalpine (1995) investigated the competence of students in making assessments. The peer assessment was designed in steps from individual tasks to group assignments. Each task was assessed by the peer group and compared with the assessment of the lecturer. High correlations were evident between student marks and lecturers marks for individual essays and presentations. The self assessment which Oldfield and Macalpine describe is also an approach in achievable steps, with the first being a comparison of contributions to group activities excluding oneself, then including oneself, and finally a self assessment of individual work. The students first make a peer assessment of all of the groups' achievements. To train self assessment skill, students also had to do this for their own group. The same procedure takes place within the group: assessing the group members and then assessing one's own contribution. Orpen and Macalpine found that this assessment procedure strengthens the confidence of students to assess the work of others and of themselves.

Orpen (1982) describes a study with 21 students in an organisational behaviour course and 21 students in a political philosophy course. These students had to write an essay. The students were informed that "their papers would be marked by five lecturers later in the year, and that their final grade would be the average of the marks they received from their fellow-students and from the lecturers" (p. 568). The marks were given according to the criteria of (1) coverage of the relevant material, (2) coherence and strength of the underlying argument, and (3) fluency and clarity of expression. There was no difference between the lecturers and students in their average marks, in the variation of their marks, in the extent to which their marks agreed with each other, and in the relationship between their marks and the writer's performance in end-of-course examinations.

Rushton et al. (1993) developed a computer-based peer assessment tool. A group of 32 computer science undergraduates was asked to write an essay on the viability of peer assessment. They typed their essays on the subject of peer assessment into the system. The class was split into groups of three or four students. Each group member used the peer assessment window to mark each other's work. Contrary to expectations, the marks awarded by peers were remarkably similar to those awarded by tutors; so, peer and teacher assessment seem to be equally reliable forms of assessment.

Zoller and Ben-Chaim (1997) investigated the self assessment ability of 71 biology majors enrolled in a four-year college program, with respect to higher-order cognitive skills and their confidence in self-assessing. A specially-designed self assessment questionnaire consisted of interdisciplinary science-technology-environment-society (STES) questions and Likert-type questions involving students' confidence. Students assessed their knowledge and understanding on this questionnaire. Students evaluated themselves as quite knowledgeable. The results further showed that 75% of the students thought that they were capable of self assessing and peer assessing. Zoller and Ben-Chaim found a discrepancy between the student assessment and the teacher assessment, and they accounted for this in terms of the lack of integration between assessment and learning in contemporary science teaching.

Kwan and Leung (1996) investigated tutor- and peer-group assessment of the performance of 96 students in a simulation exercise on hotel personnel training. The group was divided into five tutorial groups. Then students were paired and each student conducted a training session with the partner to an audience. The performance of each student was assessed by the tutor and the peers according to a checklist. Results show that there was some agreement between tutor and peer group markings, but somewhat less agreement than that reported by Falchikov (1986) and Stefani (1994). Arguments for this finding are that students aren't capable of assessing because this was the first time that they did it. Secondly, students made no contribution in identifying the criteria, and there was no negotiation between tutor and students in understanding the criteria.

Orsmond et al. (1996) describe an experiment in peer assessment for a first-year undergraduate animal physiology poster assignment. The 39 pairs of students completed a poster assignment. Students were informed about the poster requirements. At the end of the 12-week lecture course, the students were divided between two laboratories. Later, the students of each laboratory were asked to mark all posters of students in the other laboratory on five criteria. Each criterion had a grade of 0–4. Each poster was marked and a peer assessment grade for each criterion was calculated based on the maximum number of ticks (the modal value). After the students marked the posters, the tutor also marked the posters without seeing the marks which students had given. Orsmond et al. (1996) found that there was 18% agreement between students and tutor, with 56% of students over-marking and 26% of students under-marking. The correlation was 0.54. The students also filled in a questionnaire which showed that 76% of the students thought that "the peer assessment had made them think more, and work in a more structured way" (p. 243).

Fry (1990) describes a study in which the tutor introduced peer marking. The tutor first marked the students' scripts and then handed them back to the students. The tutor asked the students to mark each other's work (peer assessment) according to a marking scheme. The agreement between the tutor's marks and the students' marks were generally very good. Fry further found five strong aspects of peer marking, as discussed in a later section.

The relationships between self-, peer-, and co-assessments were examined by Horgan et al. (1997). They used the predictions of grades, actual grades, peer reviews, and reflective essays on self assessment of undergraduate teacher education students to analyse these relationships. The students were trained in self assessment. The students completed three multiple-choice examinations, the third of which was a cumulative final. Students predicted their grade and, after the examination, they reflected on their performance. The students also did a written analysis of a case study which was self-assessed and reviewed by three peers and the instructor based on five criteria. A third part in the assessment procedure described by Horgan et al. (1997) was an oral case analysis as part of a group. These presentations were also reviewed by peers. The final part was an essay about reflection on the self assessment activities. Results of the assessments described above showed (1) agreement across assessors, (2) little consistency of self assessment across tasks, (3) improvement in accuracy over the semester, (4) increased accuracy with increased performance, and (5) that better students used self assessments to guide work, while weaker students used feedback to find the errors.

The literature reviewed above illustrates how effective self-, peer-, and co-assessments are used in combination. Self and peer assessment can be used for summative purposes as part of the co-assessment by giving the tutor the power to make the final decision about a process or a product. In this way, the traditional assessment, in which the tutor makes an autonomous decision, is not comparable with co-assessment. The combination of self-, peer-, and co-assessment makes tutors and students work together in a constructive way and, as a result, they come to higher levels of understanding by negotiation. When students become teachers, this role-changing provides them with insights into the assessment process.

Several studies indicate that the marks given by the tutors and those given by the students are highly correlated (Freeman, 1995; Fry, 1990; Longhurst & Norton, 1997; Oldfield & Macalpine, 1995; Orpen, 1982). Only a few researchers found low correlations between student and tutor marks (Kwan & Leung, 1996; Rushton et al., 1993). Arguments for these findings were that the assessment and the learning weren't sufficiently integrated, or that students are not capable of assessing themselves when they first have to.

It can be concluded that the use of self-, peer- and collaborative-assessment is important to remove the student/tutor barrier, to develop enterprising competencies in students, and leads to greater motivation and 'deeper' learning (Somervell, 1993).

Where application of self assessment and peer assessment were mostly used for formative purposes, combinations of these forms with co-assessment work out well for summative assessments. Various applications show various possibilities, ranging from using the peer assessment as a contribution (e.g. 25%) to the overall score, to using peer assessment as a correction score for tutor-assessment.

Developments in this area clearly open possibilities for assessing skills and abilities in areas in higher education where traditionally there were problems in assessing or in the costs of assessing. If peer- and co-assessment indeed is a valid, fair and useful method for assessing essays and assignments, it could become a widespread assessment method in the near future.

What are the effects on the learning environment?

In order to find an answer to the second research question, regarding the effect of using self- peer- and co-assessment on the learning environment, we can refer to several previously-discussed studies which showed quantitative effects. Studies such as Hassmén et al. (1996) and Martens and Dochy (1997) show that students who use self assessment procedures get better scores on a final test. However, such studies mostly involved assessments as support devices embedded in the learning materials and used a traditional final test.

Some studies reviewed in earlier sections of this Chapter stress the agreement between self- or peer-marks and teacher-marks (e.g. Falchikov & Boud, 1989; Stefani, 1992, 1994). However, these findings don't give us much information about the impact which the form of assessment has on the learning process. Falchikov and Boud (1989) stated that "although we have focused on student-teacher agreement over-rating, we must not be distracted by the search to maximize congruence at all costs. Self assessment can be a valuable learning activity, even in the absence of significant agreement between student and teacher, and can provide potent feedback to the student about both learning and educational and professional standards" (p. 427). Above all, this is the main reason why these forms of assessment need to be integrated into curricula in higher education.

The studies in this literature review are more concerned with integrating learning and instruction, which means not only using assessment instruments to improve the learning process, but also adapting the assessment instruments (including the final examination) to the main goals of higher education. In this case, the focus is on the effects of new assessment forms on the quality of the learning process and the creation of a learning environment. Therefore, the studies primarily in the fifth category of our classification (studies about the relationship between assessment and learning/ reflection) were analysed.

An important concept that links assessment with the quality of learning is that of consequential validity (Boud, 1995, p. 41), which refers to the effects of assessment on learning and other educational matters. Assessment procedures of high consequential validity should be developed. Encouraging deep approaches to learning is one aspect which can be explored in considering consequences. Another is the impact which assessment has on the competencies and skills that students have in being able to assess themselves.

The importance of developing students into reflective practitioners is already pointed out. Boud (1992, 1995) developed a self assessment schedule to provide a comprehensive and analytical record of learning in situations where students have substantial responsibility for what they do. The main guidance is a handout which suggests the headings (goals, criteria, evidence, judgements and further actions) that students might use. Self assessment schedules are effective tools to use in enabling students to bring together a wide range of their learning, to reflect on their achievements and to examine the implications for further learning (Boud, 1992; Boud & Knights, 1994). Boud (1990) further recognised the gap between what is required of students in higher education and what happens in real life. He stresses the need for examination of assessment practices to see if they are compatible with the goals in higher education as described in the introduction of this Chapter. The two main goals of student assessment in his view are improvement of the quality of learning (e.g. by reflection) and the need to certify. Boud sees self and peer assessment as fundamental to learning. Existing assessment practices might be more defensible if they could bear some relationship to the ways in which academic and other professional work is assessed in actual working environments and the situation in which knowledge is used.

Adams and King (1995) investigated the perceptions of self assessment held by different student groups and the skills required for self-assessing. They also recognised that employment at a professional level usually requires specialist knowledge. An important part of this knowledge is the ability to have a continual knowledge of one's own capabilities and to be able to update weaknesses as appropriate. Adults, for example, like to learn and are capable of considerable amounts of high-quality learning of their own. Adams and King identified activities that can develop self assessment skills. A framework helps students to develop self assessment skills to be competent at self assessment. Adams and King identified three levels. At the first level, students work on having an idea about the assessment process. Students perform activities such as discussing good and bad characteristics of sample work, discussing what is required in an assessment, critical reviews on literature, etc. At the second level, students work on identifying important criteria for assessment. At the third level, students work towards playing an active part in identifying and agreeing on assessment criteria and being able to assess peers and themselves competently.

The relation between reflection and self assessment is also pointed out by Sobral (1997). Self assessment of self-directed learning support reflection and learning partnerships and are facilitated by discussions and exercises. Therefore, Longhurst and Norton (1997) claim that self assessment is clearly an important part of helping students to improve their own learning, as it focuses students' attention on the meta-cognitive aspects of their learning and teaches them to be more effective at monitoring their own performance.

In some studies, the perceptions of students towards innovative assessment and its impact on learning are investigated. Sambell et al. (1997), for example, investigated the perceptions of students towards different aspects of innovative assessment. When discussing innovative assessment, many students believed that success more fairly depended on consistent application and hard work, not on a last-minute burst of effort or sheer luck. Many students felt that openness and clarity were fundamental requirements of a fair and valid assessment system. Students were very positive about the effects of alternative assessment on their learning.

A small-scale study on the views of a group of newly-enrolled Open University students in London resulted in a mixed response to alternative methods of assessment (Peters, 1996). The majority of the students disagreed with self and peer assessment. This finding, however, did not mean that the students were totally committed to traditional forms of assessment. The possibility of being able to re-draft assignments after tutor feedback was viewed more favourably. Williams (1992) found that the majority of students (90%) see benefits in peer assessment. Benefits are seen in three main categories: in comparison of approaches; in comparison of standards; and in exchange of information. However, students found that criticising their friends was difficult (see also Strachan & Wilcox, 1996). Students also found peer assessment difficult or undesirable when guidelines for evaluation are not established first. The two major findings in the study of Williams (1992) were that (1) students like to have more say in how they approach their learning and its assessment and (2) students need guidance and training in new role behaviours before this can actually happen. Orsmond et al. (1996) found that students enjoyed carrying out the peer assessment and considered that it was beneficial to their learning. Keaten and Richardson (1992) also affirmed that peer assessment fosters an appreciation for internal awards and interpersonal relationships in the classroom. Cheng and Warren (1997) conducted research in the English department of the Hong Kong Polytechnic University to gauge students' attitudes prior to and after a peer assessment. The students and the teacher assessed each group seminar and oral presentation. Before and after the peer assessment, the students filled out a questionnaire with four items. The results of the questionnaire show that students were mostly positive towards the peer assessment, but that a few students thought that beginning students were unable to conduct the assessment in a fair and responsible manner. The same result was reported by Falchikov and Boud (1989). Further, the students were not entirely confident in their ability to assess their peers. However, there was a positive shift overall in both attitudes and confidence. Finally, Cheng and Warren concluded that there is a need to give students systematic and comprehensive training in how they can assess their peers and how to establish criteria (see also Williams, 1992).

Overall, self-, peer-, and co-assessment seem to improve different aspects of the quality of the learning environment and the learning of students. However, training in the skill to self assess or to peer assess is needed for an optimal impact. The prior section stresses the use of self-, peer-, and co-assessment as tools for learning (i.e. ways to develop reflective practitioners in higher education).

Discussion

Self-, peer-, and co-assessment in higher education can be used in different ways and can be reliable and valid to a large extent. Assessment as a tool for learning has considerable impact on the students' learning and development into reflective practitioners. In this final part of this Chapter, the main strengths and weaknesses of self-, peer-, and co-assessment that arise from the studies are discussed.

According to McDowell (1995), the main strengths are that (1) there is a development of evaluative and critical abilities, (2) there are opportunities for skill development, (3) knowledge is more integrated and (4) students collaborate, are motivated and are satisfied. The weaker points of innovative assessment lie sometimes in organisational issues, such as the occurrence of cheating, stress and time constraints. At last, there could be a mismatch between learning and marks without feedback. Self assessment improves the independence in students' learning, responsibility for decision making related to assignments, proactivity, and creativity in taking charge of their own work (Klenowski, 1995). Constraints on student self assessment were lack of time, the paucity of professional development and support for student self-evaluation, and the change process itself. Adams and King (1995) remind us of a real problem of perception. The idea that teachers do the teaching and marking is hard to change. Also there is the possibility that students take advantage of their role and that they become strategic in their approach to their studies. Once students experience the self assessment schedule described by Boud (1992, 1995), they see its value. The majority of students were initially supportive and became enthusiastic having been through the process of constructing a schedule (Boud, 1992, p. 191). There is much more concern about the self-grading aspect than there is about the qualitative assessment. If it is not possible to demonstrate that students can produce marks which are acceptable for teachers, the self assessment should be restricted to a purely learning role and as a skill to be developed (Boud, 1989). Until now, self assessment is regarded as a formative tool. Moreover, it should be clear that students have to know the criteria clearly and that peer assessment can be time consuming.

Strengths in using self and peer assessment (see also Brown & Dove, 1991) are that it (1) can foster students' feelings of ownership for their own learning, (2) can motivate students and encourage their active involvement in learning, (3) makes assessment a shared activity rather than a lone one (i.e. more objective), (4) promotes a genuine interchange of ideas, (5) leads to more directed and effective learning, (6) encourages students to become more autonomous in learning, (7) signals to students that their experiences are valued and their judgements are respected, (8) develops transferable personal skills, (9) produces a community of learning in which students feel that they have influence and involvement, (10) reduces the teacher's workload (Rushton et al., 1993), and (11) makes students think more deeply, see how others tackle problems, pick up points and learn to criticise constructively.

Considering these strengths and weaknesses, it may be concluded that probably the most difficult aspect of self-, peer-, and co-assessment is to determine the criteria. Criteria are the basis of evaluating student progress; they identify the critical aspects of a performance or a product that describe in specific terms what is involved in meeting the learning outcomes. It is necessary for the concept that the criteria are presented in operational terms with which all participants are familiar. Criteria should include information about the area to be assessed, the aims to be pursued and the standards to be reached (Boud, 1995). Boud and Falchikov (1989) identified two elements in any assessment decision: the identification of criteria or standards to be applied to one's work; and the making of judgements about the extent to which work meets these criteria. In self assessment, students judge their own performance and products against their own assessment criteria (Falchikov, 1986). Students also have to be trained in self and peer assessment. They have to learn and understand their role in the assessment process. Assessment should only be used in a "summative assessment system whose outcome is not a grade or label but a *profile* of the student to which all who are able to speak about him can contribute what they know — and in which conflicting assessments are highlighted rather than ironed out" (p. 297). Involvement in learning, including assessment, is vital to effective learning, and the teacher is the key person to help students to develop this learner autonomy (Dickinson, 1998).

An important shift is that the future labour market will play a distinguished part in the way in which the curriculum and the goals are revised (Pilot, 1997). The form of assessment determines whether the student achieves the skills required for the working field. This working field will thus also influence the content of the assessments (Moerkerke & Terlouw, 1998). Curricula nowadays are becoming more competency-based. This redesign of the curriculum requires a redesign of the assessment in order to create a learning environment that fosters learning. Self-, peer-, and co-assessment can discourage passive, reproductive forms of learning. By integrating these forms of assessment into the curriculum, students are likely to develop into competent persons and lifelong learners who reflect continuously on their behaviour and learning process (Moerkerke, 1996). These alternative forms of assessment should be a part of a process of change towards a student-centred learning environment. This change requires a shift in emphasis from the norm-referenced to the criterion-referenced testing, from purely summative to formative and summative assessment, from external to internal evaluation, and from the assessment of product to the assessment of process as well. In order to be successful, the following supporting factors seem to be necessary: pedagogical change; a shared value system between students and teachers; and an organisation-wide evaluation ethic.

References

Adams, C., & King, K. (1995). Towards a framework for student self assessment. *Innovations in Education and Training International, 32,* 336–343.

Anderson, J.B., & Freiberg, H.J. (1995). Using self assessment as a reflective tool to enhance the student teaching experience. *Teacher Education Quarterly, 22*, 77–91.

Arter, J. (1996). Using assessment as a tool for learning. In R. Blum, & J. Arter (Eds.) *Student performance assessment in an era of restructuring* (pp. 1–6). Alexandria, VA: Association for Supervision and Curriculum Development.

Beckwith, J.B. (1991). Approaches to learning, their context and relationship to assessment performance. *Higher* Education*, 22*, 17–30.

Birenbaum, M., & Dochy, F. (1996). *Alternatives in assessment of achievement, learning processes and prior knowledge*. Boston, MA: Kluwer.

Boud, D. (1989). The role of self assessment in student grading. *Assessment and Evaluation in Higher Education, 14*, 20–30.

Boud, D. (1990). Assessment and the promotion of academic values. *Studies in Higher Education, 15,* 101–111.

Boud, D. (1992). The use of self assessment schedules in negotiated learning. *Studies in Higher Education, 17*,185–200.

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

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

Boud, D., & Knights, S. (1994). Designing courses to promote reflective practice. *Research and Development in Higher Education*, *16*, 229–234.

Brown, S., & Dove, P. (1991). Opening mouths to change feet: some views on self- and peer assessment. In S. Brown, & P. Dove (Eds.), *Self- and peer assessment* (pp. 59–65). Birmingham, UK: Standing Conference on Educational Development.

Burnett, W., & Cavaye, G. (1980). Peer assessment by fifth year students of surgery. *Assessment and Evaluation in Higher Education, 5,* 273–278.

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

Conway, R., Kember, D., Sivan, A., & Wu, M. (1993). Peer assessment of an individual's contribution to a group project. *Assessment and Evaluation in Higher Education, 18*, 45–56.

Cutler, H., & Price, J. (1995). The development of skills through peer assessment. In A. Edwards, & P. Knight (Eds.), *Assessing competence in higher education* (pp. 150–159). Birmingham, UK: Staff and Educational Development Series.

Dancer, W.T., & Dancer, J. (1992). Peer rating in higher education. *Journal of Education for Business, 67*, 306–309.

Dickinson, L. (1988). Collaborative assessment: an interim account. In H. Holec (Ed.), *Autonomy and self-directed learning: present fields of application* (pp. 121–128). Strasbourg, France: Council of Europe.

Dochy, F., & McDowell, L. (1997). Assessment as a tool for learning. *Studies in Educational Evaluation, 23*, 279–298.

Dochy, F., & Moerkerke, G. (1997). The present, the past and the future of achievement testing and performance assessment. *International Journal of Educational Research, 27*, 415–432.

Falchikov, N. (1986). Product comparisons and process benefits of collaborative peer group and self assessments. *Assessment and Evaluation in Higher Education, 11*, 146–166.

Falchikov, N. (1991). Group process analysis: self and peer assessment of working together in a group. In S. Brown, & P. Dove (Eds.), *Self- and peer assessment* (pp. 15–27). Birmingham, UK: Standing Conference on Education Development.

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

Falchikov, N., & Boud, D. (1989). Student self assessment in higher education: a meta-analysis. *Review of Educational Research, 59*, 395–430.

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

Fry, S.A. (1990). Implementation and evaluation of peer marking in higher education. *Assessment and Evaluation in Higher Education, 15*, 177–189.

Gentle, C.R. (1994). Thesys: an expert system for assessing undergraduate projects. In M. Thomas, T. Sechrest, & N. Estes (Eds.), *Deciding our future: technological imperatives for education* (pp. 1158–1160). Austin, TX: The University of Texas.

Goldfinch, J., & Raeside, R. (1990). Development of a peer assessment technique for obtaining individual marks on a group project. *Assessment and Evaluation in Higher Education, 15*, 210–231.

Griffee, D.T. (1995). Criterion-referenced test construction and evaluation. In J.D. Browne, & S.O. Yamashita (Eds.), *Language testing in Japan* (pp. 20–28). Tokyo, Japan: The Japan Association for Japan Language Testing.

Hall, K. (1995, February). *Co-assessment: participation of students with staff in the assessment process: a report of work in progress*. Invited paper presented at the Second European Electronic Conference on Assessment and Evaluation. EARLI-AE list European Academic & Research Network (EARN) (EARLI_AE on Listserv@nic.surfnet.nl).

Harrington, T.F. (1995). *Assessment of abilities*. Greensboro, NC: ERIC Clearinghouse on Counseling and Student Services.

Hassmén, P., Sams, M.R., & Hunt, D.P. (1996). Self assessment responding and testing methods: effects on performers and observers. *Perceptual and Motor Skills, 83*, 1091–1104.

Horgan, D.D., Bol, L., & Hacker, D. (1997, August). *An examination of the relationships among self, peer, and instructor assessments*. Paper presented at the meeting of the European Association for Research on Learning and Instruction, Athens, Greece, 1997.

Kane, J.S., & Lawler III, E.E. (1978). Methods of peer assessment. *Psychological Bulletin, 85*, 555–586.

Keaten, J.A., & Richardson, M.E. (1992, February). *A field investigation of peer assessment as part of the student group grading process*. Paper presented at the Western Speech Communication Association Convention, Albuquerque, NM.

Keith, S.Z. (1996). Self-assessment materials for use in portfolios. *Primus, 6*, 178–192.

Klenowski, V. (1995, April). *Student self-evaluation processes: empowering students in learner-centered contexts.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Knoors, E., Dochy, F., & Moerkerke, G. (1995). *Factors influencing the use of prior knowledge and progress assessment: a review of literature and a drop out study*. Heerlen, The Netherlands: Open University.

Kwan, K., & Leung, R. (1996). Tutor versus peer group assessment of student performance in a simulation training exercise. *Assessment and Evaluation in Higher Education, 21*, 205–214.

Loacker, G., & Jensen, P. (1988). The power of performance in developing problem solving and self-assessment abilities. *Assessment and Evaluation in Higher Education, 13*, 128–150.

Longhurst, N., & Norton, L.S. (1997). Self-assessment in coursework essays. *Studies in Educational Evaluation, 23*, 319–330.

Martens, R., & Dochy, F. (1997). Assessment and feedback as student support devices. *Studies in Educational Evaluation, 23*, 257–275.

McDowell, L. (1995). The impact of innovative assessment on student learning. *Innovations in Education and Training International, 32*, 302–313.

McNamara, M.J., & Deane, D. (1995). Self-assessment activities: toward language autonomy in language learning. *TESOL Journal, 5*, 17–21.

Moerkerke, G. (1996). *Assessment for flexible learning*. Utrecht, The Netherlands: Lemma.

Moerkerke, G., & Terlouw, C. (1998). Herontwerp van toetsing [Redesign of assessment]. *Tijdschrijft voor Hoger onderwijs en Management, 1*, 19–25.

Nevo, D. (1995). School-based evaluation: a dialogue for school improvement. London: Pergamon Press.

Oldfield, K.A., & Macalpine, J.M.K. (1995). Peer and self-assessment at the tertiary level - an experiential report. *Assessment and Evaluation in Higher Education, 20*, 125-132.

Orpen, C. (1982). Student versus lecturer assessment of learning: a research note. *Higher Education*, 11, 567–572.

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.

Peters, M. (1996). Student attitudes to alternative forms of assessment and to openness. *Open Learning*, 11(3), 48–50.

Pilot, A. (1997, December). *Curricula voor een op informatie gerichte maatschappij*. [Curricula in information society]. Paper presented at the Consortium COO HBO and SUNCOO Conference, Utrecht, The Netherlands.

Pond, K., Ul-Haq, R., & Wade, W. (1995). Peer review: a precursor to peer assessment. *Innovations in Education and Training International, 32*, 314–323.

Rushton, C., Ramsey, P., & Rada, R. (1993). Peer assessment in a collaborative hypermedia environment: a case study. *Journal of Computer-Based Instruction, 20*, 75–80.

Sambell, K., & McDowell, L. (1998). The value of self and peer assessment to the developing lifelong learner. In C. Rust (Ed.), *Improving student learning — improving students as learners* (pp. 56–66). Oxford, UK: Oxford Centre for Staff and Learning Development.

Sambell, K., McDowell, L., & Brown, S. (1997). "But is it fair?": an exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation, 23*, 349–371.

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

Slavin, R. (1986). Best-evidence synthesis: an alternative to meta-analysis and traditional reviews. *Educational Researcher, 15*, 5–11.

Sluijsmans, D.M.A., Dochy, F., & Moerkerke, G. (1998). Alternatieve toetsmethoden in studentgericht onderwijs. [Alternative assessment in student centred education]. *Tijdschrift voor Hoger Onderwijs, 16*, 298-314.

Sobral, D.T. (1997). Improving learning skills: a self-help group approach. *Higher Education*, *33,* 39–50.

Somervell, H. (1993). Issues in assessment, enterprise and higher education: the case for self-, peer and collaborative assessment. *Assessment and Evaluation in Higher Education, 18,* 221–233.

Stefani, L.A.J. (1992). Comparison of collaborative, self, peer and tutor assessment in a biochemistry practical. *Biochemical Education*, *20,* 148–151.

Stefani, L.A.J. (1994). Peer, self- and tutor assessment: relative reliabilities. *Studies in Higher Education, 19*, 69–75.

Strachan, I.B., & Wilcox, S. (1996). Peer and self-assessment of group work: developing an effective response to increased enrolment in a third-year course in microclimatology. *Journal of Geography in Higher Education, 20*, 343–353.

Warkentin, R.W., Griffin, M.M., Quinn, G.P., & Griffin, B.W. (1995, April). *An exploration of the effects of cooperative assessment on student knowledge structure*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Weaver II, R., & Cotrell, H.W. (1986). Peer evaluation: a case study. *Innovative Higher Education, 11*, 25–39.

Williams, E. (1992). Student attitudes towards approaches to learning and assessment. *Assessment and Evaluation in Higher Education, 17*, 45–58.

Zoller, Z., & Ben-Chaim, D. (1997, August). *Student self-assessment in HOCS Science Examinations: is it compatible with that of teachers?* Paper presented at the meeting of the European Association for Research on Learning and Instruction, Athens, Greece.

Chapter 3

Peer assessment in problem based learning[1]

Abstract

Two studies are presented in this chapter that answer the following questions: 1) Are peer ratings in problem based learning groups reliable? 2) Do raters have idiosyncratic strategies? and 3) What are students' experiences with problem based learning and peer assessment? In the first study the population consisted of 27 university students subscribed to a fourth year course in educational sciences using problem based learning. The population in the second study consisted of 51 fourth year students of a Primary Teacher Training College, also using problem based learning. Generalizability studies prove that peer ratings are reliable in one of the two studies, but results from evaluation questionnaires reveal that students feel very uncomfortable when they have to make negative judgements without any prior instruction. Implications for the design of peer assessment in problem based learning environments are discussed.

[1] Based on: Sluijsmans, D.M.A., Moerkerke, G., Dochy, F., & Van Merriënboer, J. (2001). Peer assessment in problem based learning. *Studies in Educational Evaluation, 27*, 2, 153-173.

Introduction

In many institutions for higher education problem based learning has become the educational concept. The aim of problem based learning is to improve students' ability to work in a team to solve new, complex and ill-structured real-life problems, showing their co-ordinated abilities to access information and turn it into viable knowledge. Knowledge, then, would not be something possessed only for the learners' own sake, but rather something accessed and constructed when needed to solve a problem or design something useful (Segers & Dochy, 1999). These key elements should be transferred to the design of assessment. However, the tests in problem based learning are not always in line with the goals of problem based learning. Teachers often develop assessments that test content knowledge, rather than areas like self-directed learning, problem solving and skills as a group member. Progress tests for example, are applied in medical problem based learning. These are multiple-choice tests with true/false questions about all content areas of a specific profession (Van der Vleuten, Verwijnen, & Wijnen, 1996). Although Van der Vleuten et al. claim that progress tests are in line with the curricular goals of problem based learning, the danger still exists that students develop 'test behaviour': they only invest in what is required in the assessment (Lockwood, 1995). And if this is content knowledge, they only learn content knowledge. In these cases, assessment is not congruent with instruction, since the goal is to go beyond memorisation. New forms of assessment, such as overall tests (Segers, 1997) and peer- and co-assessment (Sluijsmans, Dochy, & Moerkerke, 1999) provide such possibilities.

It is important to plead for an assessment system that requires students to use higher-order thinking skills to solve and analyse problems instead of memorising facts and solving well structured, decontextualised problems. Two of these higher-order skills, which are important in professional organisations, are that students should be able to reflect on their own behaviour and that of their peers. The assumption that self and peer assessment are important skills in order to work on complex problems is widely acknowledged in education (e.g., Birenbaum & Dochy, 1996; Boud, 1995; Sambell & McDowell, 1998). Problem-based learning should moreover occur in a clear operationalisation of a constructivist learning environment, characterised by co-operative learning and self-directed learning. In such a learning environment the responsibility for the learning process is partly given to the student. In order to enlarge the educational congruence, students should also receive responsibility in the assessment. Peer-assessment provides such an opportunity (Dochy, Segers, & Sluijsmans, 1999).

In problem based learning students work in groups. The size of the group varies from 7 to 14. When students work together as a team on a particular problem, each student has to take his or her responsibility for a certain part of the task. Tutors often find it difficult to determine what each individual has contributed to the group product. Introducing peer assessment can be a way to force students to take the responsibility to make a judgement about the actual contribution of each of their peers in the group discussion.

Prior analysis of 62 studies showed that self and peer assessment can be effective tools to develop the skills needed in the working field (Sluijsmans, Dochy, & Moerkerke, 1999). But assessing one's process or product is not a simple task. Because students often are novices in assessing the work of a peer, rating errors can occur. A number of these errors can be identified, five of which are well documented. Although these errors are described in general terms, one can conclude that these errors are also applicable to peer assessment.

First, there are many personal differences among raters in their standards and their rating styles (Coffman, 1971; De Groot, 1975). Raters may differ in their severity or leniency. Some raters consistently tend to give high grades (lenient raters), while others consistently tend to give low grades (severe raters; see also Lunz, Wright, & Linacre, 1990). Second, raters differ in the extent to which they distribute grades on the score scale. Some raters tend to distribute scores closely around their average; others will spread scores much more widely. In other words, some raters avoid giving extreme grades while others prefer to use them. A third effect is the so-called halo effect. This is the tendency of human raters to base distinctive aspects of the rating on an overall impression created by one single dominating aspect. This may indicate that raters cannot differentiate among distinct aspects of one product or procedure (Borman, 1975). Fourth, the significant effect refers to the fact that raters may have different opinions about the rating tasks. According to Voss and Post (1990), this problem is not so much related to the divergent views of an individual, but rather to the diverging opinions of groups of individuals. Voss and Post argue that in particular in the assessment of 'soft' or less 'tangible' skills, objectivity is significantly decreased due to divergence of views among raters of different schools. The fifth and last rating error is caused by so-called evaluation policy. Judges differ in the ways they employ criteria (Sadler, 1983). Every assessor has his or her own evaluation policy. According to some the performance must achieve a minimum qualifying level on a number of criteria. Other judges act conjunctively: While the performance is excellent on one criterion, it is weak on the rest of the criteria. One could also judge compensatorily: Poor showings on some criteria could be balanced by high performance on others.

Based on the rationale for introducing peer assessment in a problem based learning context, two exploratory studies were conducted to find answers to the following research questions:

1. Are peer ratings in problem based learning groups reliable?
2. Do students have idiosyncratic (i.e., personal) strategies in peer assessment?
3. What are students' experiences with peer assessment and problem based learning?

Studies I and II below describe projects in which the students themselves assessed the work process of each of their peers, while the product was assessed by the tutor.

Study I

Method

Participants

The population consisted of 27 university students (9 male, 18 female) who were enrolled in a four-year course in educational sciences using problem based learning. The students were randomly distributed amongst two groups – Group I (n = 13) and Group II (n = 14). Twenty students graduated the first licentiate in educational sciences with satisfaction (approximately 'B') and seven graduated with distinction (approximately 'A'); 23 of the students entered university directly after secondary education, and 4 students had first been enrolled in higher vocational education before attending the university.

Materials

At the end of a predefined period, all students assessed the peers in their own group on four criteria, which were explained in detail on a peer assessment form. The criteria were defined by the students in negotiation with the tutor. These criteria were: (1) contribution to the group discussions, (2) quality of the contributions, (3) preparedness to be involved in tasks, and (4) actual contribution to the teamwork. Peers scored on a scale varying from *better than the group* (3), *mean of the group* (2), *slightly below the mean of the group* (1*), no help for the group* (0) to *hindrance for the group* (-1). This scaling was based on a comparable scaling method used by Boud (1995), with a positive contribution to the group yielding positive scores, and a negative contribution to the group yielding negative scores.

A two-part evaluation questionnaire was developed. The first part consisted of 28 closed items (5-point Likert scale) about different aspects of problem based learning, such as working in a team, problem solving, the learning process and the role of the tutor. These 28 items were reduced to four variables: the satisfaction of working in a group, the achievement of the goals of problem based learning, the instructional process and the role of the tutor. The second part included eleven items about peer assessment, seven yes/no-items and four open-ended questions.

Procedure

The two groups worked for four consecutive periods of six weeks. Each period had several specific, content-related goals, such as understanding different teaching and learning methods and being able to apply several alternative assessment tools. In each period the groups received one or more problem tasks which had to be solved in the group. At the end of each of the four periods, students had to report how they had solved the problem. Students shared the work load for this report and organized meetings in order to be able to make it a real group result.

The four peer assessment criteria were defined by the students in negotiation with the tutor in weekly two-hour discussion meetings. At the end of the fourth period, a session was organized for conducting the peer assessment and filling out the evaluation questionnaires. In Group I, for example, each student gave scores to his or her twelve peers.

In this study, the peer assessment score was a part of the final score. The tutor also rated each student on the four criteria.

Data analysis

In order to examine whether the peer ratings in problem based learning groups are reliable, the data of the peer assessments were analysed. The reliability of the ratings was estimated within the framework of generalizability theory. This theory provides a mechanism for disentangling the error term into multiple sources (Brennan, 1983). Through generalizability analysis, the relative magnitude of variance caused by persons, raters, criteria, and their interaction can be estimated. In contrast to classical test theory, which treats only one error source at a time (e.g., inter-rater reliability or test-retest reliability), generalizability recognizes that there may be multiple sources of error variance which determine how accurately observed scores allow us to generalize raters' behaviour in a universe of situations. The *person variance* is an estimate of the variance across person's mean scores, where the mean is taken across all criteria and raters. The *criteria component* is the estimated variance of criteria mean scores, where each mean is taken across all persons and raters. The *rater component* is the variance of rater mean scores, where each mean is taken across all persons and criteria.

Decisions about students will generally not be based on the results of a single scoring of a single task. Important individual scores of a student will be based on average scores over multiple criteria and/or raters. In a so-called decision study or D-study the reliability of the scores can be estimated on the basis of the variance components (Brennan, 1983; Feldt & Brennan, 1989). A decision study is designed to identify the number of raters that would be required to obtain acceptably small error variances or acceptably large reliability coefficients.

To investigate if raters have personal strategies a Q-analysis was applied. Using the so-called Q-analysis or profile analysis, it is possible to determine the similarities and differences among raters (McKeown & Thomas, 1988; Tucker, 1962). Our Q-analysis used inter-rater correlations as similarity measures. The aim of the analysis of ratings was to verify whether there were (groups of) students with idiosyncratic strategies on the peer assessment task. For each of the two groups, correlations were calculated. Each correlation matrix was analysed with Principal Component Analysis using SPSS.

The evaluation questionnaires were analysed to measure the third research question about students' experiences with peer assessment and problem based learning. Descriptives were calculated for the four variables concerning problem based learning. Frequencies were calculated for the seven yes/no-questions about the peer assessment. The answers to the four open-ended questions about peer assessment were analysed qualitatively and reduced to categories.

Results

The reliability of peer ratings

In Tables 1 and 2 the estimated variance components are shown for the student ratings for Group I and Group II. A positive sign regarding the results of Group I is the fact that the largest variance in the scores was related to the performance of *persons* (40%). The variance related to the *raters* was relatively large (*raters*: 5%; *persons***raters*: 11%; *raters***criteria*: 3%) compared to other studies.

Table 1. Estimation of variance components persons, raters and criteria of Group I

Source of variance	Sum of Squares	df	Means square	σ^2	% total variance
Persons (P)	178.531	13	13.733	0.232	40
Raters (R)	32.352	13	2.488	0.032	5
Criteria (C)	0.449	3	0.150	0.000	0
Persons*raters	80.077	169	0.474	0.065	11
Persons*criteria	18.122	39	0.465	0.018	3
Raters*criteria	17.872	39	0.458	0.017	3
Error	108.556	507	0.214	0.214	37

Table 2. Estimation of variance components persons, raters and criteria of Group II

Source of variance	Sum of Squares	df	Means square	σ^2	% total variance
Persons (P)	114.882	12	9.569	0.1429	15
Raters (R)	35.515	12	2.960	0.0399	4
Criteria (C)	2.822	3	0.941	0.000	0
Persons*raters	104.3313	144	0.725	0.127	13
Persons*criteria	57.947	36	1.610	0.196	20
Raters*criteria	13.562	36	0.377	0.227	24
Error	93.668	432	0.217	0.217	23

The results of Group II show that the variance in scores related to persons (15%) is not the largest component in this analysis. Moreover, variance components involving *raters* were high (*raters*: 4%; *persons*raters*: 13%; *raters*criteria*: 24%).

Table 3. Estimation of generalizability coefficient (ρ^2) for the pooled peer assessment procedure

Number of rating students	G-coefficient for group I	G-coefficient for group II
1	.653	.408
2	.785	.550
3	.841	.622
4	.872	.665
5	.892	.695
6	.906	.716
7	.916	.731
8	.923	.743
9	.929	.753
10	.934	.761
11	.938	.768
12	.942	.774
13	.944	.778

In Table 3 the estimates of the generalizability coefficients are given for both groups. Gronlund (1988) gives some rules of thumb for the acceptability of generalizability coefficients. He states that the generalizability coefficients for classroom assessments in education usually have values between 0.60 and 0.80; 0.60 is considered to be acceptable, but open to improvement, 0.80 is considered as a very reasonable value. The generalizability of the ratings in Group I is better than those of Group II. Within Group I, there is hardly any need for pooling the scores. The generalizability of scores based on the work of one rating student is already acceptable (0.653). When final scores are based on three rating students the generalizability is good (0.841). In Group I there is, from the perspective of generalizability, no need to base the scores for the peer assessment on more than three students. The generalizability coefficients for Group II show that the quality of scoring is low. A somewhat acceptable level is reached when the final score is based on ratings of at least three students. Although the maximum value looks acceptable, one should bear in mind that this value (0.778) is only reached when the final score is a composite of 13 scores.

Idiosyncratic strategies

The eigenvalues and percentages of explained variance of the first two principal components are presented in Table 4. The structures of the eigenvalues show, in both groups, a dominant first principal component. The amount of variance between respondents explained by the first factor is 69.3% for Group I and 63.7% for Group II. The amount of variance explained by the second factor is respectively 9.9% and 11.1%. The scree-test strongly suggests a final solution with two principal components. This means that the systematic variance in the peer ratings of the students could be accounted for by two latent variables.

Table 4. Eigenvalues and percentages of explained variance for the first three principal components for Groups I and II

Group	Principal component	Eigenvalue	% of explained variance
I	1	10.51	69.3
	2	1.48	9.9
	3	0.86	5.7
II	1	8.91	63.7
	2	1.56	11.1
	3	1.13	8.0

Figure 1 presents the component plot for Group I. The factor loadings on the first component for the students ranged from 0.72 to 0.92, while the factor loading for the tutor was 0.70. This means that all respondents had a substantial and positive correlation with the first principal component. This component may be interpreted as a mutual understanding of the quality. Factor loadings on the second component ranged from -0.53 to 0.48, for the students and was 0.56 for the tutor. So, the tutor is somewhat at the extreme pole of the second principal component. The second principal component can be interpreted as the deviation of the mutual understanding of the quality of learning.

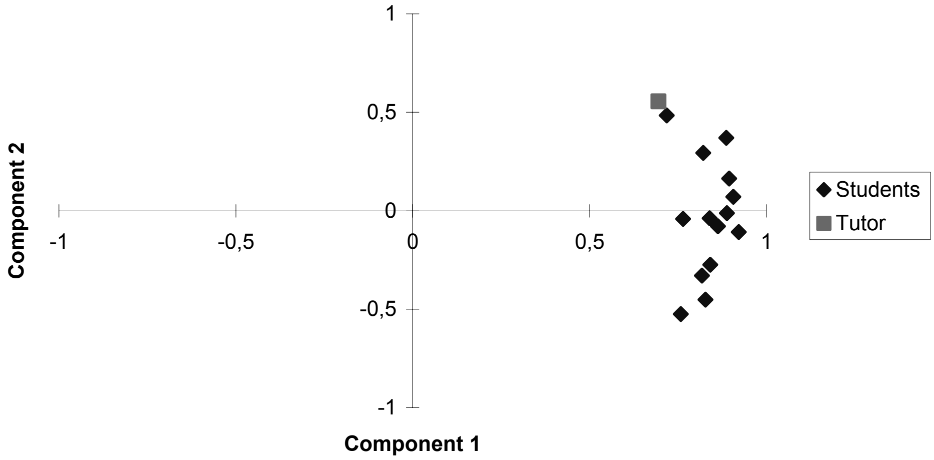


Figure 1. Component plot Group I

Figure 2 presents the component plot for Group II. For the students the factor loadings on the first component ranged from -0.22 to 0.90, while for the tutor this was 0.83. One rater had a negative factor loading. The other raters had a substantial and positive correlation with the first principal component. Factor loadings on the second component ranged from -0.36 to 0.51. The factor loading for the tutor is 0.13. Inspection of the component plots suggests that one of the students was using an idiosyncratic rating strategy.

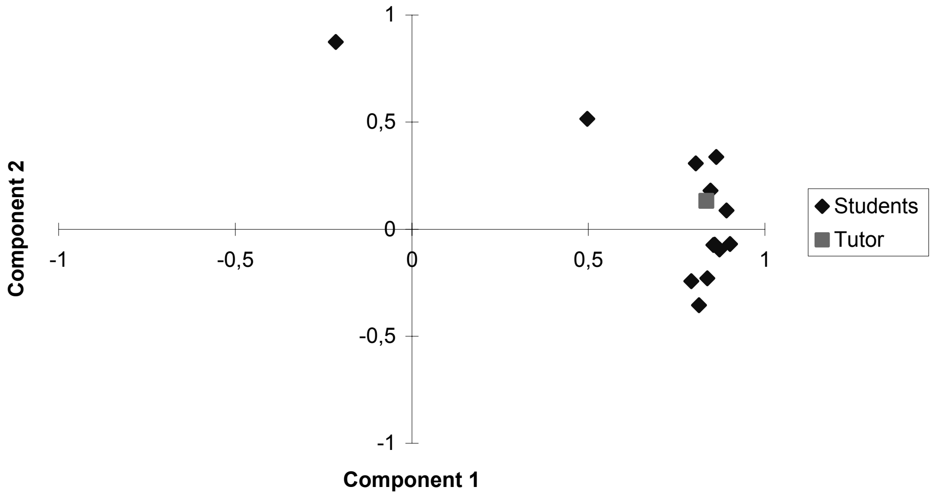


Figure 2. Component plot Group II

Students' experiences

Students were asked to give their opinion on various items in the following areas: teamwork, the goals of problem based learning, the instruction and the role of the tutor. In Table 5 the results on the problem based learning variables are presented.

Table 5. Descriptives of the items about problem based learning in Study I (*N* = 27)

Variables	Minimum	Maximum	Mean	SD
Working in a group	2.80	4.40	3.30	.39
Achieving goals of PBL	2.70	4.50	3.42	.42
Instruction process	2.82	4.09	3.47	.32
Role of the tutor	2.00	4.00	3.26	.51

Students were slightly positive about working in a group. Regarding the achievement of the goals of problem based learning, the majority of the students indicate that problem based learning contributed to the development of their problem solving skills and critical thinking abilities. Regarding the instruction process, the situation is not substantially different. In contrast to traditional lectures, problem based learning brings the content of the knowledge domain to the surface and makes the students responsible for their own learning process. Prior knowledge is activated and sometimes the relationship with other knowledge domains is discussed. The authentic character of the problem tasks stimulates active participation in discussion and in working towards problem solution. The search for and selection of relevant information, either independently or in the group, and the integration of different topics were considered as very useful. The perceived role of the tutor varied from quite negative to very positive, with a mean of 3.26, which can be regarded as a neutral attitude.

The students feel that working in a system of problem based learning is very intensive and invokes a high level of responsibility. The learning effects though are very positive because of the active participation in the group process. A high level of co-operation in the group is regarded as conditional for an optimal effect of problem based learning. Dominant roles of certain students occurred, which hindered other students to contribute their input. The students perceived a need for more attention for the development of communication skills. The students sometimes felt that they did not receive enough feedback during the course periods. Especially because they were not used to learning and working in a problem based way. In Table 6 the results on the peer assessment items are presented for Study I.

Table 6. Percentages of 'Yes' responses on the items peer assessment in Study I

Items peer assessment	%
Students are capable of assessing each other	44
Students are capable of assessing each other in a fair and responsible way	19
I feel comfortable when assessing peers	7
I knew what peer assessment was about	66
I am in favour of implementing peer assessment	74
Implementing peer assessment means a major change for our institution	82
Peer assessment can be used in other courses	26

In the open-ended questions, students were asked to write down their experiences with peer assessment. The most positive aspects were that the students had the opportunity to express their opinion about the contribution of each peer-student in the group. In this method, more persons make a judgement. The students felt that their scores could be helpful to the tutor. The involvement in the assessment was regarded as fair, although the majority of the students doubted the reliability of the method. Some students were stimulated to think critically about their own learning behaviour. Students indicated that peer assessment is not only product-evaluation but also process-evaluation.

A more negative aspect of the peer assessment was that the contribution of the peers differed every period while a score had to be given for the average contribution of four periods. Many students indicated, moreover, that working with a score only was too simple and ineffective. There was no room for feedback. The criteria appeared to be difficult to interpret. One student suggested to weight the criteria, i.e., to indicate which criteria are important and which are of less importance. The peer assessment was not introduced sufficiently well. Some students experienced it as difficult and felt uncomfortable, because they had no prior experience in peer assessment.

Conclusions

In Group I, the largest variance in the scores was related to the performance of *persons* (40%). This means that most of the variance in scores can be attributed to individual effort. The *raters'* variance components suggest differences in leniency. The interaction between *persons* and *raters* (*persons*raters*) indicates that interpersonal relationships are biasing the peer assessments. The magnitude of the variance components for these student ratings was in concordance with the pattern often found in studies on the generalizability of performance assessment (cf. Moerkerke, 1996).

The quality of ratings clearly differs between the two groups. This means that it is unclear if the peer assessment method applied leads to acceptable results. Probably, the robustness of the method can be improved (as outlined in the discussion below). The peer assessment score is based on the ratings of 12-14 students. Acceptable quality was reached for both groups with this number of raters: The generalizability coefficient was over .90 for Group I and over .75 for Group II.

When we look at the relative magnitude of the eigenvalue for the first component and the magnitude of each of the loadings on the first principal component it can be concluded that there is a high level of common strategy among the *raters* in Group I. The relative magnitude of the eigenvalue for the first component and the magnitude of each of the loadings on the first principal component also pointed to a high level of common strategy among the *raters* in Group II. The analysis of interrater correlations revealed that one rater can be regarded as odd.

The results of the problem based learning items in the questionnaire show that students stress cooperation, working on an authentic problem, constructing knowledge and skills and active participation as the best features of problem based learning. They generally feel that implementation of this system requires a lot of time and that the ability to work together in a group demands particular skills, which need to be trained and monitored. The most striking result is that the students felt very uncomfortable in assessing their peers. The implementation of peer assessment would be a major change for the institution.

Study II

Method

Participants

Besides a study in a university setting, an identical study was organized within the setting of higher vocational education. The population in the second study consisted of 51 fourth year students of a primary teacher training college (13 male, 38 female) also using problem based learning. The students were randomly distributed amongst four groups (*n1* = 12; *n2* = 13; *n3* = 13; *n4* = 13).

Materials

The peer assessment form used in Study I was used for the actual peer assessment. The same evaluation questionnaire was used to measure students' perception on problem based learning and peer assessment.

Procedure

During the course 'Developing School Plans', students had to develop a school plan that represented the ideal elementary school. To do this, students had to be aware of current developments in elementary education. In this particular course, students had to be willing to make an active contribution to a group product, to work independently and to be responsible for their own learning.

Each of the four groups worked for a period of six weeks. The groups received a problem task, which concerned the development of a school plan. At the end of the sixth week, all students assessed the peers from their group on the same four criteria as in Study I. In contrast to the first study, the scores were not used as part of the final score. The teacher of the class decided that the peer assessment was an independent activity in the course. The individual peer assessment score was translated into an absolute score ranging from 1 to 10. In this situation a non-passing peer assessment score (< 5.5) would lead to an additional study task. In this study the tutor was not one of the raters in the peer assessment.

Data analysis

The data-analyses were similar to the analyses in Study I.

Results

The reliability of peer ratings

In Tables 7-10, the estimated variance components are shown for the student ratings for Groups I to IV, respectively. To enable the interpretation of the quality of the method, the data of the four groups should be compared. The four groups manifest two patterns. The pattern of Groups I and II corresponds with the pattern of Group II in the first study. This pattern consists of relatively low *person* variance (about 20%) and thus relatively high error variance and interaction variance. *Rater* variance (about 5%) and *criteria* variance (about 1%) are low.

Table 7. Estimation of variance components persons, raters and criteria of Group I

Source of variance	Sum of Squares	df	Means square	σ^2	% total variance
Persons (P)	17.005	11	1.546	0.0278	14
Raters (R)	10.130	11	0.921	0.011	6
Criteria (C)	0.880	3	0.293	0.000	0
Persons*raters	17.973	121	0.149	0.006	3
Persons*criteria	6.223	33	0.189	0.005	3
Raters*criteria	12.432	33	0.377	0.021	2
Error	45.214	363	0.125	0.125	64

Table 8. Estimation of variance components persons, raters and criteria of Group II

Source of variance	Sum of Squares	df	Means square	σ^2	% total variance
Persons (P)	19.905	12	1.659	0.028	22
Raters (R)	6.231	12	0.518	0.006	5
Criteria (C)	2.071	3	0.690	0.002	2
Persons*raters	16.018	144	0.111	0.011	9
Persons*criteria	5.467	36	0.152	0.007	6
Raters*criteria	5.006	36	0.139	0.006	5
Error	28.456	432	0.066	0.066	52

Table 9. Estimation of variance components persons, raters and criteria of Group III

Source of variance	Sum of Squares	df	Means square	σ^2	% total variance
Persons (P)	76.485	12	6.374	0.117	44
Raters (R)	17.562	12	1.464	0.022	8
Criteria (C)	2.402	3	0.801	0.003	1
Persons*raters	27.015	144	0.188	0.027	10
Persons*criteria	5.751	36	0.160	0.006	2
Raters*criteria	7.598	36	0.211	0.010	4
Error	34.749	432	0.080	0.080	30

Table 10. Estimation of variance components persons, raters and criteria of Group IV

Source of variance	Sum of Squares	df	Means square	σ^2	% total variance
Persons (P)	45.524	12	3.794	0.069	43
Raters (R)	6.678	12	0.556	0.006	4
Criteria (C)	1.129	3	0.376	0.002	1
Persons*raters	27.899	144	0.194	0.038	24
Persons*criteria	2.121	36	0.059	0.001	1
Raters*criteria	3.121	36	0.087	0.004	3
Error	17.379	432	0.040	0.040	25

This pattern in variance leads to low generalizability coefficients as indicated in Table 11. A somewhat acceptable level of generalizability is reached when at least five ratings are considered. The second pattern can be found in Groups III and IV. This pattern corresponds with the pattern of group I in the first study. It consists of relatively high *person* variance (about 40%), low *rater* variance (about 6%) and low *criteria* variance (about 1%), and leads to acceptable generalizability coefficients. The mean score of one or two ratings lead to acceptable rating practice.

Table 11. Estimation of generalizability coefficient (ρ^2) for the pooled peer assessment procedure

Number of rating students	G-coefficient for group I	G-coefficient for group II	G-coefficient for group III	G-coefficient for group IV
1	.419	.488	.708	.585
2	.582	.643	.825	.737
3	.669	.720	.873	.807
4	.723	.766	.899	.847
5	.760	.796	.915	.873
6	.787	.817	.926	.891
7	.807	.833	.935	.904
8	.823	.846	.941	.915
9	.836	.856	.946	.923
10	.846	.864	.950	.930
11	.855	.870	.953	.935
12	.862	.876	.956	.940
13	.869	.881	.958	.944
14	.874	.885	.960	.947

Idiosyncratic strategies

In order to investigate this question the same approach was used as in Study I. Using the interrater correlations matrix of each of the four groups a principal component analysis was conducted. The explained variance of the first two principal components is presented in Table 12. Due to lack of variance the principal component analysis for Group II cannot be performed. The pattern of the component plot of Group I has the same structure as group II in Study I (see Figure 2). The patterns of the component plots of Groups III and IV are comparable with Group I in Study I (see Figure 1).

Table 12. Eigenvalues and percentages of explained variance for the first two principal components for Groups I, III and IV

Group	Principal component	Eigenvalue	% of explained variance
I	1	5.54	46.2
	2	2.33	19.4
III	1	11.04	84.9
	2	.95	7.3
IV	1	9.98	76.6
	2	1.29	9.9

Students' experiences

The students in this study also filled out the evaluation questionnaire. The descriptives of the four variables of the problem based learning environment (working in a group, achieving the goals, the instruction process, the role of the tutor) are presented in Table 13.

Table 13. Descriptives of the items about problem based learning in Study II (N = 57)

Variables	Minimum	Maximum	Mean	SD
Working in a group	2.92	4.62	3.99	.51
Achieving goals of PBL	2.70	4.50	3.67	.48
Instruction process	2.82	4.09	3.87	.42
Role of the tutor	1.80	4.20	2.56	1.08

The mean of the first variable indicates that problem based learning stimulates active contribution in a group. The students are also positive about the achievement of certain goals. It appeared that working according to the problem based approach is supportive in developing skills like defining instructional problems, analysing problems, critical thinking, leading a discussion, interacting with peers, using prior knowledge, giving argumentations and presenting reports. The instruction process was evaluated very positively, resulting in a mean score of 3.87. Finally, the students were less positive about the role of the tutor. In Table 14 the results on the peer assessment items are presented.

Table 14. Percentages of 'Yes' responses on the items peer assessment in Study II

Items peer assessment	%
Students are capable of assessing each other	73
Students are capable of assessing each other in a fair and responsible way	61
I feel comfortable when assessing peers	59
I knew what peer assessment was about	31
I am in favour of implementing peer assessment	71
Implementing peer assessment means a major change for our institution	53
Peer assessment can be used in other courses	50

Results show that the majority feels that they are capable of making assessments and in favour of implementation of peer assessment practices, although they were not very well informed as to what the peer assessment was about.

The most positive element of the peer assessment for the students was that they felt involved in the assessment procedure. They appreciated that their opinion was taken seriously. The majority of the students experienced peer assessment as a method to force them to think about the contribution of their peers in the group as well as about their own contribution. Some students stressed that a student at a primary teacher training college should be capable of giving critical comments on the work of peers. Peer assessment was regarded as an opportunity to express appreciation for the work done.

The most negative aspects of the peer assessment were that the students felt uncomfortable in making negative judgements. They stressed that the situation is too personal and that it is useless if there is no opportunity to give feedback. For certain students, receiving a negative score can have serious consequences. Students indicate that they should not have the power to give negative scores if there is no evaluation or argumentation afterwards.

Conclusions

The question *Are peer assessments in problem based learning groups reliable?* cannot be answered affirmatively. The patterns in the data are scattered. In two of the four groups the peer assessment method does lead to acceptable generalizability coefficients. However, in the other two groups this is not the case. This result can be explained by a low *person* variance. In such a case, the scores do not discriminate between skilled and non-skilled students. An in-depth look at the data of Groups III and IV in Study II revealed that those categories of the rating scales that indicated incompetent behaviour, were hardly used. When scales are not fully used, scores become homogeneous and thus non-informative.

Based on the relative magnitude of the eigenvalue for the first component and the magnitude of each of the loadings on the first principal component it can be concluded that there is a high level of common strategy among the *raters* in Groups III and IV. The relative magnitude of the eigenvalue for the first component and the magnitude of each of the loadings on the first principal component pointed to a lower level of common strategy among the *raters* in Group I of Study II. The component plot of the data of group I also showed one outlier.

The results of the questionnaire show that most students support the idea of problem based learning, but that they are less positive about the way the tutor functioned. Students are positive about the peer assessment, although they agree that filling out a piece of paper does not change the whole educational system. The method needs improvement; in this way the peer assessment is too subjective according to a lot of students. Many students agree that this is an innovation that needs further development, both for students and teachers.

Overall conclusions and discussion

Studies I and II describe the reliability of peer assessment in problem based learning groups. The generalizability of the ratings in Study I appeared to be better for Group I than for Group II. Within Group I, the generalizability is acceptable when final scores are based on three rating students (0.84). The generalizability coefficients for Group II led to the conclusion that the maximum value is acceptable, but that this value (0.78) is only reached when the final score is a composite of 13 scores. In Study II, an acceptable level of generalizability is reached in Groups I and II when at least five ratings are considered (0.76 and 0.80). In Groups III and IV an acceptable level of generalizability is reached when at least two ratings are considered (0.83 and 0.74). This pattern corresponds with the pattern of Group I in the first study. Although the groups are randomised, the result of the first study leads in one group to acceptable results and in the other group to unacceptable results. The peer assessment method does not lead to enough score variance. This is also the case in the second study, where only the results in Groups III and IV lead to acceptable results.

It is remarkable that there seems to be a cultural difference between the students in Study I (university education for educational scientists) and Study II (higher vocational education for primary school teachers). In Study I extremely negative scores were used more often. The comments of the students in Study II indicate that they were not satisfied with the method. They found it unacceptable to give negative scores without having the opportunity to give informative feedback. In a primary teacher training college students should be supported to give constructive comments and not a mere score.

The results of the questionnaire about problem based learning of Study I and II are quite similar, although two aspects need some discussion. First, the students in Study I are less positive about working in a group than the students in Study II. An explanation for this difference between the students of the two studies might be that students in Study I had no experience with problem based learning and working in teams, while students in Study II were more familiar with this kind of instruction. Second, students in the first study are less negative about the tutor than the students in the second study. This may be explained by the fact that the tutor in the second study was much less involved; students had to do the problem solving process all on their own.

The results of the peer assessment part of the questionnaire revealed that the students in the second study are more confident in their ability to assess than the students in the first study. They feel more comfortable about assessing than the group in the first study. An explanation could be that students in Study II have more experience with different kinds of instruction. The students in Study I had more prior knowledge, because the teacher informed the students better about the peer assessment. All students were positive about implementing peer assessment, but the students in the first study predicted many more implications for the institution. Half of all students see possibilities for peer assessment in other courses with problem based learning.

On the whole it can be concluded that the peer assessment method applied in these studies needs improvement. One improvement could be that not only processes but also products are evaluated in a peer assessment. In the current study only the process was subject to assessment.

Another major improvement would be that students are provided with a possibility to give informative feedback to benefit subsequent learning processes. These types of improvements should lead to a full use of the assessment scale.

Peer assessment the way it was conducted in the studies does not prevent rating errors like *friendship marking*, resulting in over-marking; *collusive marking*, resulting in a lack of differentiation within groups; *decibel marking*, where individuals who dominate groups get the highest marks; and *parasite marking*, where students fail to contribute but benefit from group marks (Pond, Ul-Haq, & Wade, 1995). The rating errors, outlined in the introduction of this chapter, seem not to be eliminated. Severity and leniency, respectively, lead to under-marking and over-marking of peers. The halo effect occurs when students find one criterion the most important one, thus slanting the objectivity towards the other criteria. Students have different views on the quality of a performance. This was outlined as the significant effect. And finally, evaluation policy means that all students have their own interpretations of the importance/ meaning of the established assessment criteria.

Giving students opportunities to carry out peer assessments by means of a scoring form seems not enough. Our results underpin the need for instruction in peer assessment, in order for students to make reliable judgements. While Arter (1996) and Perkins (1986) already stressed the need for training students in assessment skills, schools have only recently been paying attention to the development of this type of skills. Students, but also teachers, seldom get training and practice in the development of assessment skills. This was also the case in the school settings presented in the two studies. Moreover, little is known about training in assessment skills.

The basic goal in current and future research should be to control the strategies students use when they have to make a judgement about their own work or that of their peers. Novice-behaviour in rating is characterised by rating-errors or the naive strategies learners exhibit in using peer assessment. When there is no training in assessment skills, rating processes will stay subject to a variety of measurement errors. Hogarth (1981) already stated that the literature shows a depressing picture of human judgmental ability.

Despite the occurrence of rating errors, several studies show that the ability of students to rate themselves improves in the light of feedback or development over time (Birenbaum & Dochy, 1996; Boud & Falchikov, 1989; Griffee, 1995). Moreover, students' interpretations are not just dependent on the form of the assessment process, but on how these tasks are embedded within the total context of the subject and within their total experience of educational life.

An overall conclusion is that adequate training in peer assessment strategies is necessary to eliminate rating errors. This training has to be embedded in the course domain, in order to ingrate assessment and the instruction (Frederiksen, 1994). To begin a process of designing instruction in which peer assessment strategies are included, it is important to reflect on the learning environment in which such training can achieve an optimal effect. A clear definition of the context is required. Criteria have to be elaborated and discussed during a course period. Students have to learn how to give feedback and how to write a peer assessment report. Peer assessment is not only a tool to provide a peer with constructive feedback which is understood by the peer. Above all, peer assessment is a tool for the learner himself (Dochy & McDowell, 1997).

References

Arter, J. (1996). Using assessment as a tool for learning. In R. Blum, & J. Arter (Eds.), *Student performance assessment in an era of restructuring* (pp. 1-6). Alexandria, VA: Association for Supervision and Curriculum Development.

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* (pp. 3-29). Boston, MA: Kluwer.

Birenbaum, M., & Dochy, F. (Eds.) (1996). *Alternatives in assessment of achievement, learning processes and prior knowledge*. Boston, MA: Kluwer.

Borman, W.C. (1975). Effects of instruction to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology, 60*, 556-560.

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

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

Brennan, R.L. (1983). *Elements of generalizability theory*. Iowa City, IA: American College Testing Program.

Coffman, W.E. (1971). Essay examinations. In R.L. Thorndike (Ed.), *Educational measurement* (pp. 271-302). Washington, DC: American Council on Education.

De Groot, A.D. (1975). *Methodology* (9th ed.). 's-Gravenhage: Mouton.

Dochy, F., & McDowell, L. (1997). Assessment as a tool for learning. *Studies in Educational Evaluation, 23*, 279-298.

Dochy, F., & Moerkerke, G. (1997). The present, the past and the future of achievement testing and performance assessment. *International Journal of Educational Research, 2*, 415 - 432.

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

Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: Macmillan.

Frederiksen, N. (1994). The integration of testing with teaching: Applications of cognitive psychology in instruction. *American Journal of Education, 102*, 527-564.

Griffee, D.T. (1995). Criterion-referenced test construction and evaluation. In J.D. Browne & S.O. Yamashita (Eds.), *Language testing in Japan* (pp. 20-28). Tokyo, Japan: The Japan Association for Japan Language Testing.

Gronlund, N.E. (1988). *How to construct achievement tests.* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Lockwood, F. (1995). Students' perception of, and response to, formative and summative assessment material. In F. Lockwood (Ed.), *Open and distance learning today* (pp. 197-207). London: Routledge.

Lunz, M.E., Wright, B., & Linacre, M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*, 331-345.

McKeown, B., & Thomas, D. (1988). *Systematic data collection*. Newbury Park: Sage.

Pond, K., Ul-Haq, R., & Wade, W. (1995). Peer review: A precursor to peer assessment. *Innovations in Education and Training International, 32*, 314–323.

Sadler, D.R. (1983). Evaluation and the improvement of academic learning. *Journal of Higher Education, 54,* 60-79.

Sambell, K., & McDowell, L. (1998). The value of self and peer assessment to the developing lifelong learner. In C. Rust (Ed.), *Improving student learning — improving students as learners* (pp. 56–66). Oxford, UK: Oxford Centre for Staff and Learning Development.

Segers, M.S.R. (1997). An alternative for assessing problem-solving skills: The OverAll test. *Studies in Educational Evaluation, 23,* 373-398.

Segers, M., & Dochy, F. (1999). Een nieuw onderwijsmodel voor het Hoger Onderwijs in theorie en praktijk [A new educational model for higher education in theory and practice]. In M. Lacante, & P. De Boeck, *Handboek leerlingenbegeleiding [Handbook student counselling]* (pp. 153-180). Dordrecht: Kluwer.

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

Tucker, L.R. (1962). Factor analysis of relevance judgements: An approach to content validity. *Proceedings: 1961 Invitational Conference on Testing Problems* (pp. 29-38). Princeton, NJ: Educational Testing Service.

Van der Vleuten, C.P.M., Verwijnen, G.M., & Wijnen, W.H.F.W. (1996). Fifteen years of experience with progress testing in a problem based learning curriculum. *Medical Teacher, 18,* 103-109.

Voss, J.F., & Post, T.A. (1990). On the solving of ill-structured problems. In N. Frederiksen, R. Glaser, A. Lesgold & M.G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 261-285). Hillsdale, NJ: Erlbaum.

Chapter 4

Redesigning education for training peer assessment skills in teacher training: Effects on performance and perceptions[1]

Abstract

This chapter 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. Based on a theoretical framework about student teachers and peer assessment, a study was designed in which 93 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 a redesigned course, since the implementation of peer assessment in the chosen perspective on peer assessment implies reconsideration of existing courses. 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' skill to define criteria and the effects on the content related skill were examined. Students' perceptions on the redesigned course were also investigated. Both the peer assessment reports regarding the content related product that all students wrote at the end of the course, and the marks on the content related product were analysed. Questionnaire results provided answers regarding students' perceptions. 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. It was found that both groups were significantly more satisfied regarding the new design of the course compared to prior courses.

[1] Based on: Sluijsmans, D.M.A., Brand-Gruwel, S., Van Merriënboer, J.J.G. & Martens, R.L. (2001, submitted). *Redesigning education for training peer assessment skills in teacher education.* Presented as a paper at the biennial EARLI conference, Fribourg, Switzerland, August 28-September 1, 2001.

Introduction

Like 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, Kessels, & Meijers, 2000). Skill-based learning is an ongoing issue, especially in the domain of Teacher Education (Darling-Hammond & Snyder, 2000; James, 2000; Kremer-Hayon & Tillema, 1999; Willems, Stakenborg, & Veugelers, 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 organisation. A number of Teacher Training Colleges collaboratively formulated a broad scale of skills student teachers need to develop. These skills of a primary school teacher are reported in a vocational training profile (LPC, 1995), which consists of 41 skills that are categorised in 10 domains. The skills represent the overall accepted knowledge, proficiency and attitudes a primary school teacher needs to 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 the student teachers meet the criteria of each skill. These criteria have to be the same as those used in the practice setting (Hager & Butler, 1996).

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.

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 (Cohen, 1994; Johnson, Johnson, & Johnson-Holubec, 1992; Sharan & Sharan, 1994; Slavin, 1995). Teachers have to work together, learn from each other and become a member of a learning organisation (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. The peer assessment skill is one of such skills.

Second, as prospective teachers of children in primary schools, it is advisable to learn student teachers how to make critical judgements about the performance of peers, and later on about performances 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, Rust, & Gibbs, 1994). Training in peer assessment skills stimulates this mutual influence to take place at a professional level.

The reasons mentioned above convinced 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, Dochy, Moerkerke, & Van Merriënboer, 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, Merry, & Reiling, 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, 1989; Freese, 1999; Hayon-Kremer & Tillema, 1999; Korthagen, 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.

Several authors (Birenbaum, 1996; Fallows & Chandramohan, 2001; Hanrahan & Isaacs, 2001) recommend 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 are identified. The model is based on several sources. First, the literature on peer assessment is analysed (see also Sluijsmans, Dochy, & Moerkerke, 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 eleven constituent skills were defined (see Figure 1). The defined skills are the basis for the design of the training in peer assessment.

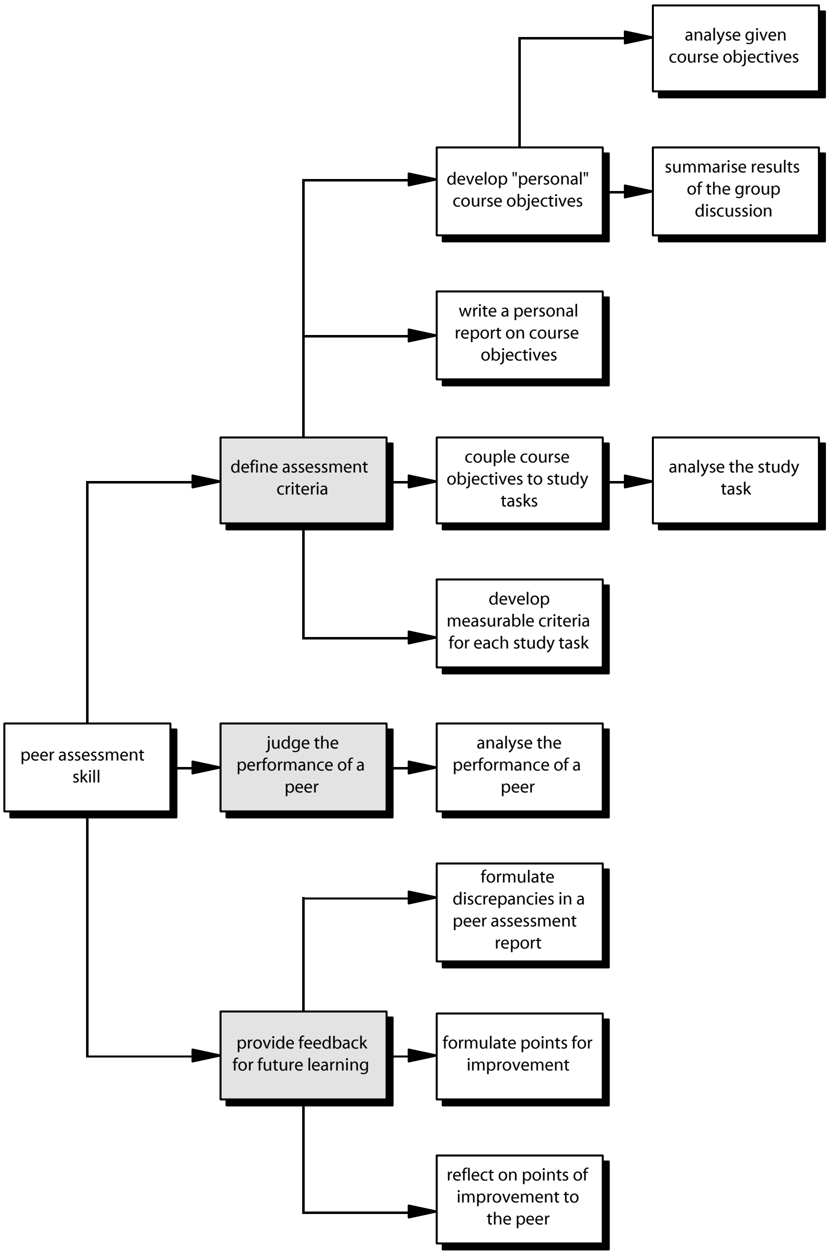

analyse given course objectives
develop "personal" course objectives
summarise results of the group discussion
write a personal report on course objectives
define assessment criteria
couple course objectives to study tasks
analyse the study task
develop measurable criteria for each study task
peer assessment skill
judge the performance of a peer
analyse the performance of a peer
formulate discrepancies in a peer assessment report
provide feedback for future learning
formulate points for improvement
reflect on points of improvement to the peer

Figure 1. Skill decomposition peer assessment

Because the peer assessment skill is too complex to be trained 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*.

Designing a training in peer assessment is based on two assumptions. A 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, Popham, & Ryan, 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 a training of 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 work of peers and a content-related skill, which contains the object 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).

Second, 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, Kessels & Meijers, 2000). This means that skills are trained in the context of different content domains. This simultaneous change on both course design and the role of students is often experienced as 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 will also be given to the consequences of such redesign of courses and to potential effects of content domains and/or teachers. It is investigated how students respond to a change in the course design.

In summary, the present study will serve three goals: investigating the effects of a peer assessment training on the development of the assessment skill, the effect of the training on task performance in the domain of the course, and guidelines for designing courses that are suitable for training professional skills. Based on the presented theoretical framework, the following research questions are elaborated:

1. Does following a training in peer assessment lead to the development of the skill to assess the work of peers?
2. Does following a 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 of 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 (*n* = 43), which received peer assessment training, 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 to 12 age range. Five teachers of the Teacher Training College participated in this study. Each teacher taught 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 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 subskills 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 instructed, based on the four skills. For example, the study tasks 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.

It was decided to rename the course into 'Designing Discovery Learning Lesson Plans'. A complete overview of the organisation of the course is given in Appendix 3.

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 regular 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 characterised by interactive discussions between the students to foster collaborative learning and paid attention to the skills that are related to 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 presents how the peer assessment tasks are embedded in the regular study tasks.

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 end products and peer assessment	-

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.

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, a 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 ten 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. On 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 trained 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 form. For each variable the interrater-reliabilities were calculated. These reliabilities were acceptable for all variables (Cohen's Kappa > .95).

Examinations. To measure an 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 are 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 on 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.

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

Variable	#	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 finding 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 a 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*	14	.89	I enjoyed it to work together on a study task as a group

* Cronbach's alpha calculated in post-test

After the course and the peer assessment, 16 students were interviewed (8 from the control group, 8 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.

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 the other colleagues. Questions related 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 as presented in Appendix 3. 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 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 ten 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, both the students of 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 4 assessments). These research assistants were instructed in the application of the rating form. One-way analyses of variance with the factor Groups were applied to identify differences between the control and experimental groups on the eight variables of the rating form.

One-way analyses of variance with the factor Groups were also applied to identify differences between the control and experimental groups on the task performance in the course domain, developing a lesson plan on discovery learning.

Means and standard deviations were calculated for the fifteen variables of the student questionnaire for the control and experimental groups. The scores of each variable were analysed with a 2 (Groups) x 4 (Content Domains) x 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 only was measured in the post-test. A 2 (Groups) x 4 (Content Domains) was done, because it was only measured in the post-test. The students from the music domain are 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 categorised 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

Effects on the assessment skill

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 (*M* = 16.77, *SD* = 9.65) than the average of the control groups (*M* = 12.89, *SD* = 6.33). The difference between both groups was significant (*F* (1,83) = 4.89, *MSE* = 63.68, *p* < .05).

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

Variable	max. score	experimental groups		control groups	
		M	*SD*	*M*	*SD*
use of criteria*	40	13.95	8.31	10.45	5.05
naive word use	4	01.32	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 < .01. * p < .05.

Further analyses reveal 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 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* = .43, *p* < .01. Overall, the training had the expected effect, because the experimental groups use the criteria significantly more often than the control groups.

Effects on the content related performances

At the end of the course, the students were responsible for one final products, a discovery learning lesson plan. The average 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.

Effects on students' perceptions and the influence of content domain

In Table 4 the means and standard deviations of the student questionnaire are given.

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

The scores of each variable were analysed according to a 2 (Groups) x 4 (Content Domains) x 2 (Time of Testing) analysis of variance with repeated measures on the last factor. Table 5 presents the *F*-values for each of the MANOVA's on the scores of all the variables.

Table 5. F-values in MANOVA's with repeated measures on scores on the fifteen variables of the student questionnaire

Variable	time of testing	groups	content domain	time of testing x groups	time of testing x content domains	time of testing x groups x content domain
Satisfaction classes	100.49***	.15	2.58	4.81*	2.67	.64
Transparency classes	26.89***	2.54	2.45	1.94	2.84	.60
Learning access level	5.69*	.06	1.67	1.22	4.14**	1.04
Practical relevance	26.36***	.06	1.96	1.58	5.54**	3.46*
Quality of the instruction	146.65***	.65	.87	1.48	2.82*	1.03
Teacher involvement	75.89***	1.08	2.29	.83	5.95**	1.64
Relation instruction and assessment	301.13***	1.21	2.43	.02	5.07**	2.06
Assessment behaviour	55.02***	1.03	.50	.11	1.09	1.45
Fear for assessment	1.64	.35	1.79	.041	.78	.66
Obtrusiveness assessments	108.32***	.00	1.24	.62	6.51**	4.38**
Overall vision on assessment	34.81***	1.26	3.58*	.69	2.31	1.13
Involvement in assessment	.03	1.30	1.46	.30	.67	2.41
Group behaviour	6.20**	.56	.04	.02	1.36	.38
Collaborative learning	.105	.37	.19	.42	.28	.17
Assessment skill	6.40**	.19	.90	.59	1.53	2.04
Group atmosphere[ab]	.	.25	.94	.	.	.

*** p < .001. ** p < .01. * p < .05.

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

[b] For Group x Content Domain, *F* (3, 58) = 2.28, *MSE* = .197, p < .05.

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 (see the means in Table 4). One important significant effect is the effect on the variable 'assessment skill', because this variable concerned items that measured students' self-perception on their skill to assess.

There were no significant main effects for Groups. For Content Domain, a significant effect is found regarding the variable 'overall vision on assessment', *F* (3,53) = 3.58, *MSE* = .561, *p* < .05. Post-hoc tests (Tukey) reveal that there was only one significant effect between two content domains, namely mathematics and philosophy (mean difference = .62, *p* < .05). The effect though is not caused by the treatment, and therefore less important.

For Time of Testing x Groups a significant interaction effect was found on the variable 'satisfaction classes', *F* (3,52) = 4.81, *MSE* = .173, *p* < .05. For the pre-test the experimental groups (*M* = 3.11; *SD* = .52) were more positive than the control groups (*M* = 2.99; *SD* = .54), while for the post-test the opposite pattern was shown (in order *M* = 3.76; *SD* = .55, *M* = 3.87; *SD* = .54).

For Time of Testing x 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.

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

Variable	Content domain	pre-test		post-test	
		M	*SD*	*M*	*SD*
Learning access level	Pedagogy	3.78	.71	3.76	.60
	Physics	3.13	.83	3.84	.76
	Mathematics	3.57	.66	3.25	.85
	Philosophy	3.76	.92	4.17	.60
Practical relevance	Pedagogy	3.68	.58	4.22	.51
	Physics	3.33	1.01	4.12	.62
	Mathematics	3.62	.61	3.45	.89
	Philosophy	3.61	.78	4.22	.64
Quality of the instruction	Pedagogy	2.84	.55	3.97	.36
	Physics	2.71	.74	3.93	.55
	Mathematics	2.99	.54	3.66	.67
	Philosophy	2.95	.72	4.34	.56
Teacher involvement	Pedagogy	3.44	.57	4.12	.46
	Physics	2.90	.80	3.96	.44
	Mathematics	3.49	.65	3.63	.84
	Philosophy	3.45	.70	4.48	.46
Relation instruction and assessment	Pedagogy	2.30	.65	4.04	.54
	Physics	1.98	.72	4.17	.57
	Mathematics	2.39	.63	3.44	1.02
	Philosophy	2.49	.59	4.29	.59
Obtrusiveness assessment	Pedagogy	2.38	.49	3.44	.59
	Physics	2.35	.56	3.41	.75
	Mathematics	2.55	.67	2.92	.69
	Philosophy	2.56	.57	3.69	.75

The means and Post-hoc analysis (Tukey) indicate that for all six variables the mathematics group shows a much lower increase or even a decrease, from the pre-test to the post-test than the three other groups.

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

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

Variable	Content domain	pre-test experimental		pre-test control		post-test experimental		post-test control	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Practical Relevance	Pedagogy	3.73	.43	3.62	.73	4.52	.41	3.88	.40
	Physics	3.42	.62	3.22	.62	3.89	.50	4.53	.65
	Mathematics	3.56	.37	3.39	.81	3.33	.69	3.56	1.04
	Philosophy	3.42	.92	3.81	.57	4.07	.64	4.44	.62
Obtrusiveness assessment	Pedagogy	2.28	.39	2.47	.57	3.63	.55	3.25	.60
	Physics	2.19	.56	2.53	.52	3.46	.76	3.31	.82
	Mathematics	2.68	.57	2.41	.77	2.55	.73	3.21	.53
	Philosophy	2.54	.65	2.59	.49	3.63	.73	3.76	.84

For ‘practical relevance’, the experimental mathematics group shows a decrease from the pre-test to the post-test while all other groups are more positive on the pre-test than the post-test. For ‘obtrusiveness assessment’, a similar pattern can be observed.

After the course, 16 students (8 from the experimental group, 8 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 indicate that 93,8% of the students rated the extent in 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. 75% of the students were satisfied with the followed 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 underpin the importance of peer assessment for their role as professional teachers. The majority of the students (83%) does not feel capable to assess a peer. 62,5% of the interviewed students stressed that it is still uncomfortable to assess a peer. 93,8% 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 realised 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: *"I think it is useful to pay attention to the development of assessment skills, because what you practice 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".*

Effects on teachers' perceptions

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.

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

Variables		PA task 1	PA task 1	PA task 1	PA task 1
invested time	*Mdn*	3.50	4.00	3.50	4.00
transparency of the task	*Mdn*	2.50	4.00	4.00	4.00
desired output	*Mdn*	3.50	4.00	4.00	5.00

All teachers indicate that the peer assessment tasks could be instructed 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 thirty 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 was 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 summarised. 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 differences between the control groups and the 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 following 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 peer assessment skills on students' performance. A difference between the performance quality of the students from the control and from 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 are found. The results showed a change of perception towards twelve 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 view on aspects of learning and assessment. They are more satisfied about the classes, the criteria and goals are clearer. The role of the teacher is 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 do 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 sometimes is hard to translate your 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 you have to prove that you understand the criteria before you may assess a peer. You need to have an objective perspective and give constructive criticism.

The factor content domain was of high influence, mainly caused by the domain mathematics. It was not possible to determine whether these findings are 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 instructing the peer assessment tasks did not occur. Because of the small number of teachers the interviews are not structurally elaborated. The answers teachers gave were illustrative for 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 operationalised. 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 (see also Boud, Cohen & Sampson, 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 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 provided far more paid time for planning: Japanese teachers for example spend about 40 percent of their paid time on professional development and collaboration compared with about 20 percent 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. And 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 it was 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 *N* 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 then 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.

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* (pp. 3-29). Boston: 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.

Boud, D., & Knights, S. (1994). Designing courses to promote reflective practice. *Research and Development in Higher Education, 16,* 229–234.

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

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* (pp. 131-141). 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). *Beroep in beweging. Beroesprofiel leraar primair onderwijs [Profession in action. Vocational training profile for the primary school teacher].* Utrecht: Forum Vitaal Leraarschap.

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 preservice 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* (pp. 97-114). Westport: Praeger.

Slavin, R. E. (1995). *Cooperative learning: theory, research and practice.* Boston: 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. (Center for Educational Technology and Expertise).

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* (pp. 19-32). Leuven-Apeldoorn: Garant.

Web-based Education Commission (2000). Helping isolated teachers make new connections. In *The power of the Internet for Learning*. Retrieved on March 13 from http://interact.hpcnet.org/webcommission/doc.htm

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

Williams, E. (1992). Student attitudes towards approaches to learning and assessment. *Assessment and Evaluation in Higher Education, 17,* 45–58.

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.

Chapter 5

Peer assessment training in redesigned teacher education: Effects on performance and perceptions[1]

Abstract

The implementation of peer assessment and the design of skill-based courses receive much attention in teacher education. This chapter reports the effects of a peer assessment training. 93 student teachers were randomly assigned to control groups and experimental groups. The experimental groups were trained in three important assessment skills, namely defining performance criteria, giving feedback and writing assessment reports. This was done through peer assessment tasks that were embedded in a redesigned skill-based course. Analyses of data derived from peer assessment reports written by the students showed that the experimental groups surpassed the control groups in the quality of the assessment skill. As a result of the training, students from the experimental groups also scored significantly higher grades on the content related end products of the course than students from the control groups. The results of the questionnaire showed that all students were significantly more satisfied with the redesigned course.

[1] Based on: Sluijsmans, D.M.A., Brand-Gruwel, Van Merriënboer, J.J.G. (2002). *Peer assessment training in teacher education.* Manuscript accepted for publication in Assessment and Evaluation in Higher Education.

Introduction

Peer assessment practices have been applied in many institutions for over 50 years. Although Kane and Lawler in their publication of 1978 address the failure of recognition of the use of peer assessment, many studies since then have proven the importance to involve students in the assessment process through peer assessment (e.g., Arter, 1996; Boud, Cohen & Sampson, 1999; Cutler & Price, 1995; Dochy & McDowell, 1997; Fallows & Chandramohan, 2001). There is an immense number of studies illustrating how peer assessment practices can be applied in curricula in both formative and summative ways (see Dochy, Segers, & Sluijsmans, 1999; Falchikov & Goldfinch, 2001; Sluijsmans, Dochy, & Moerkerke, 1999). Besides the analysis of peer-tutor correlations, which are reported in the majority of the studies, some of these studies also evaluate the perceptions and feelings of students regarding the process of peer assessment. These evaluations draw a diverse picture. On the one hand students regard peer assessment as useful for their own learning process. Orsmond, Merry and Reiling (1996) found that students enjoyed carrying out the peer assessment and considered that it was beneficial to their learning. Keaten and Richardson (1992) also affirmed that peer assessment fosters an appreciation for internal awards and interpersonal relationships in the classroom.

On the other hand, students found that criticising their friends was difficult. Hanrahan and Isaacs (2001) mention the hostility that students show towards 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). Especially the part of giving grades to each other, concerns students and is seen as 'risky and unfair' (Kwan & Leung, 1996). Besides these findings, students also doubt the objectivity of peer assessment and claim to have no training in such assessment practices (Cheng & Warren, 1997; Sluijsmans, Moerkerke, Dochy, & Van Merriënboer, 2001). This last point, the lack of instruction, is the central issue in this paper. Aim of the presented study is to examine how students can be involved in assessment and what the effects are of structural involvement on students performance and perceptions.

Critical evaluation of the performances of peers is especially important in the teacher-training context. Conducting a peer assessment is considered as a specific skill in the Dutch vocational training profile of primary school teachers (LPC, 1995). Unfortunately, there are no explicit training programs available that provide teachers in Teacher Training Colleges with guidelines how to train their student teachers in these assessment skills. The literature on peer assessment particularly focuses on the importance of negotiating about performance criteria (Falchikov, 1995; Mehrens, Popham, & Ryan, 1998; Orsmond, Merry, & Reiling, 1996; Orsmond, Merry, & Reiling, 2000), but that is only one a skill that is required for conducting reliable assessments. To make an overview of the important skills, literature review and expert interviews were conducted (Sluijsmans & Van Merriënboer, 2000). This resulted in a peer assessment model in which three main skills are taken into account. These skills 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 writing an assessment report, and (3) providing feedback for future learning: giving constructive feedback about the product of a peer.

A training program for peer assessment should be based on these skills. The design of such a training program results in a number of peer assessment tasks, which should be embedded in an existing course. After all, the course content determines the object of the peer assessment. Regarding the subskills in the model, it is a prerequisite that the teachers of the course agree on the course objectives and the assessment criteria, before they can guide students in their skills. If it turns out that an existing course contains course objectives that are ambiguous, it is necessary to discuss these course objectives with the teachers of the course. These discussions can lead to a thorough redesign to make it suitable for a training in assessment skills.

The peer assessment tasks can be regarded as the learning exercises in which the assessment skills are practiced. An example of a peer assessment task is the negotiation about assessment criteria for a specific product (e.g., Arter, 1996; Dochy & McDowell, 1997; Mehrens, Popham & Ryan, 1998; Stainer, 1997). This product is directly coupled to a certain course content related skill. In this view, the development of assessment goes hand in hand with the development of content related skills. The product, for which students define criteria, is object of peer assessment. A hypothesis is that understanding the criteria of a specific product in a certain course domain and analyzing the work of peers can lead to an improved awareness of the quality of one's own product (Falchikov, 1995; Freeman, 1995; Mehrens, Popham & Ryan, 1998).

The aim of this study is to determine the effect of a course-embedded peer assessment training on (1) students' assessment skill, (2) students' task performance in the domain of the course, and (3) students' perceptions regarding a redefined course. The students received the study points for the course only if they conducted the peer assessments seriously and handed in their assessment reports for the research purposes.

Method

Participants

The sample consisted of 93 second-year students of a Teacher Training College in the Netherlands (19 male, 74 female) with an average age of 20,7 years (*SD* = 1,6). The students were randomly distributed among four educational groups, of which two were experimental groups (*n* = 50) and two were control groups (*n* = 43). The control groups received training in defining criteria in a previous course (Sluijsmans, Brand-Gruwel, Van Merriënboer & Martens, submitted; see Chapter 4); this previous training did not yield the expected effects and makes the current experiment more conservative. Four teachers of the Teacher Training College participated in the study. Each teacher taught a certain content domain in the course. These domains were art, Dutch language, music and pedagogy.

Materials

Course. For the purpose of this study, a second-year course on creative learning was chosen. This course has been part of the curriculum for several years. Because the course objectives were not revised for years and teachers had developed multiple perspectives on what the content should be, the teachers that were jointly responsible for this course first redefined the course objective. It was decided that students were guided in one skill, namely the design of a creative lesson. At the end of the course, students had to make a videotape of a creative lesson that was designed and carried out by themselves.

The four teachers collaboratively decomposed the skill of designing a creative lesson. This resulted in a concept map with a number of constituent skills. For the domains art, Dutch language and music four one hour study tasks were defined, based on the constituent skills. In these tasks, students learned how each domain was related to creative learning and the design of creative lessons. The pedagogy teacher designed four one hour study tasks that integrated the tasks of the domains art, Dutch language and music. The whole course enclosed an introductory class, sixteen study tasks (four tasks per domain), and a concluding class in which the peer assessment was organized. The course was renamed into 'Designing Creative Lessons'.

Peer assessment training. The students were allocated to two control and two experimental groups. Four peer assessment tasks of one hour each were designed for the two experimental groups. These tasks were embedded in the study tasks of the pedagogy domain, and were closely related to the study tasks concerning designing creative lessons. The training focused on the three main constituent skills of the peer assessment model. The didactics used in the training were interactive discussion and elaboration.

In *Task 1*, students were introduced to the meaning of peer assessment and the product that they were going to peer assess at the end of the course. This product was a video of a creative learning lesson taught by two second-year students. After this introduction students watched a creative learning lesson on video, discussed and elaborated on the fragments in which creativity was applied. This resulted in a first rough draft of the criteria that are required for a creative lesson.

In *Task 2,* the skill '*defining criteria*' was addressed. Examples of valid and invalid criteria were presented. Students then further elaborated on the rough criteria for designing a creative lesson they formulated in the first task. This exercise resulted in a list of 15 criteria that are required for a creative lesson, which were accepted by the students and the teacher.

Discussing the purpose and guidelines for giving constructive feedback, was the central topic in *Task 3*. In the peer assessment model, this is the skill '*provide feedback for future learning*'. First, the teacher asked the students what their ideas were about feedback and criticism. After a short discussion, the teacher presented an expert-assessment report to the students. This was an assessment report on the video lesson that was analysed in Task 1, which was written by two experts on creative learning. Students discussed the good examples of constructive feedback. At the end of the task, students had to give each other feedback on some aspects of their own work. The output of this task was a list of criteria for constructive feedback.

In *Task 4*, the students were trained in the third main skill of the peer assessment model, namely '*judge the performance of a peer*'. In this final task, the three prior tasks were integrated. To confront the students with ways in which an assessment report can be written, they analysed the expert-assessment report and discussed the structure that was applied by the experts. They also discussed the language used in the assessment, for example the use of naive words, like 'nice'. Based on the findings, students defined a peer assessment form.

Peer assessment form. At the end of the course, all students had to assess the videotapes of the creative lessons of three peer groups on a peer assessment form. The peer assessment form defined by the students from the experimental groups in the fourth peer assessment task, consisted of four pages. The students from the control groups also used this form.

Rating form. To analyse the quality of the peer assessments that were written by the students, a rating form was developed. In this rating form seven variables were included. These variables were based on an expert-peer assessment that was used in the peer assessment training. One variable was related to the criteria ('use of criteria'), four variables were related to giving feedback ('positive comments', 'negative comments', 'constructive comments', and 'posed questions'), and another two variables were related to judging the performance of a peer by writing an assessment report ('naive word use', and 'structure'). Each student could score a maximum of 15 points on the first variable 'use of criteria', because the students defined 15 criteria in the first peer assessment task. The four variables for 'giving feedback' were measured by counting the number of comments and questions. The variable 'naive word use' was measured by counting words such as 'nice', 'good', 'excellent', and 'fine'. Four items in the rating form were included to measure the variable 'structure'. On these four items a maximum score of 13 points could be gained.

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

Examinations. To measure an effect of the peer assessment training on the task performance of students, the marks on the end products analysed. Four marks were given to each student: One for the collection of assignments students carried out during and between the study tasks, one for the group report about the process of the design process for a creative lesson, a mark for the individual report about the design of creative lessons, and one average mark for the creative products for the domains art, Dutch language and music. The pedagogy teachers marked the group report and the individual report. The other teachers marked the products for their own domain, based on criteria that were set beforehand. The score of each product could range from 0 to 100 for each product. The average of the four marks was calculated for each student. This score could also range from 0 to 100.

Student Questionnaire. Before and after the course, the students filled out a questionnaire about their perceptions on instruction and assessment. This questionnaire was developed in a prior study (Sluijsmans, Brand-Gruwel, Van Merriëmboer, & Martens, submitted). Ninety-two items are divided among sixteen variables. Six variables were related to instruction, five variables were related to vision on instruction and assessment and another four were related to the role of the student in assessment. The students had to score 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 on prior courses that were comparable to the course on 'Designing Creative Lessons'. These courses were not designed in a skill-based way. The post-test concerned students' perceptions after the redesigned course. Because of very low reliability coefficients, the items of three variables were removed from the questionnaire, 21 items in total. The clusters, variables, number of items, reliability coefficients and example items of the sixteen variables are presented in Table 1.

Table 1. Clusters, variables, number of items, reliability coefficients and example items of the thirteen variables of the student questionnaire

Variable	#	α	Example items
		Cluster: Instruction	
Satisfaction classes	5	.65	The study tasks evoked interesting discussions
Transparency classes	4	.68	The course objectives were comprehensible
Learning access level	4	.61	I felt that I could distinguish main issues from side issues
Practical relevance	3	.71	The study tasks are practically-oriented
Quality of the instruction	7	.88	The goals of the study tasks were instructed very clearly
		Cluster: Vision on instruction and assessment	
Relation instruction and assessment	4	.72	The study tasks and the assessment were interrelated
Fear for assessment	3	.58	I'm usually very nervous before taking an exam
Obtrusiveness assessments	5	.76	The questions on an exam have to be public to students before the exam is taken
Overall vision on assessment	2	.71	I support the way I am assessed
		Cluster: Role of student in assessment	
Involvement in assessment	8	.45	I think that students should be more involved in the development of assessment criteria
Group behaviour	5	.65	I don't like it when students don't make a individual contribution to a group product
Collaborative learning	3	.71	I prefer to elaborate on problems with my peers
Assessment skill	18	.87	I'm able to analyse a product of a peer

Design and procedure

The study was set up according to a pre-test-post-test control group design. Before the start of the course, all students filled out the questionnaire. During the course, all students worked in subgroups of five or six students on their design of a creative lesson, and the group report. They prepared their lesson that was taped on video and subject of the peer assessment. In between classes, each student worked individually on the individual report and the content domain related assignments.

During the course the students of the experimental groups performed the peer assessment tasks. Instead of these tasks the students of the control groups attended four extra hours in the pedagogic domain. During these hours, the control groups had the opportunity to elaborate on certain aspects of creative learning. At the end of the course, a peer assessment session was organized for each group (approximately 25 students), in which the video lessons of each subgroup were shown (four video lessons in each group).

The peers were instructed to write a qualitative peer assessment with regard to the content of the video lesson of each group. The experimental groups were free to use the output of the peer assessment tasks. For the peer assessment, the students from the control group had to use the regular course materials from the study tasks. Each student wrote three peer assessments, because in each group there were three other subgroups to assess. After the course, all students filled out the same questionnaire as in the pre-test.

Data-analyses

Three independent research assistants analysed the 279 peer assessment forms (93 students who completed 3 assessments). These research assistants had prior experience with the use of rating forms. One-way analyses of variance were applied to identify differences between the control groups and experimental groups on the seven variables of the rating form.

One-way analyses of variance were also applied to identify differences between the control and experimental groups on the product performances.

Means and standard deviations were calculated for the thirteen variables of the student questionnaire for all groups. The scores of each variable were analysed with a 2 (Groups) x 2 (Time of Testing) analysis of variance with repeated measures on the last factor.

Results

Effects on the assessment skill

Table 2 presents the means and standard deviations of the seven variables that were measured with the rating form for the experimental and control groups.

Table 2. Means and standard deviations of the experimental and control groups on the peer assessment forms at the post-test

Variable	experimental groups		control groups	
	M	*SD*	*M*	*SD*
use of criteria[a*]	9.62	2.34	6.68	2.14
positive comments	6.55	2.11	6.13	3.14
negative comments	1.47	1.06	1.32	0.88
constructive comments[*]	2.56	1.58	1.26	0.97
posed questions	0.14	0.21	0.13	0.24
naive word use[*]	1.42	0.78	1.92	0.87
structure[b*]	6.67	1.46	6.21	1.47

[*] $p < .01$.

[a] maximum score = 15

[b] maximum score = 13

The analyses reveal significant differences between the groups on four of the seven variables. The experimental groups applied more criteria (variable 'use of criteria', *F* (1,85) = 36.10, *MSE* = 5.07, *p* < .01) and gave more constructive feedback (variable 'constructive comments', *F* (1,85) = 20.13, *MSE* = 1.77, *p* < .01). They also used less naive words (variable 'naive word use', *F* (1,85) = 7.57, *MSE* = 0.68, *p* < .01). At last, the experimental groups scored better on the variable 'structure' (*F* (1,85) = 7.89, *MSE* = 1.98, *p* < .01).

Effects on the task performance in the course domain

All students handed in several products to ground their development in the skill to design creative lessons. Four marks were given to each student: One for the collection of assignments students carried out during and between the study tasks, one for the group report about the process of making the video lesson, a mark for the individual report about the design of creative lessons, and one average mark for the creative products for each domain. The average of the four marks was calculated. The score of the experimental groups was 72.77 (*SD* = 6.72) compared to an average score of the control groups of 68.42 (*SD* = 5.27). The difference between both groups was significant (*F* (1,83) = 10.59, *MSE* = 37.45, *p* < .01). The students from the experimental groups thus performed better on the skill of designing creative lessons than students from the control groups. Further analyses show that the significant difference between the two groups is caused by the significant effects in the performance of two of the four products, namely the collection of assignments and the group report. The students from the experimental groups scored an average of 72.34 (*SD* = 6.98) on the assignments against an average of 67.63 (*SD* = 6.75) of the students from the control groups, which is a significant effect (*F* (1,83) = 9.84, *MSE* = 47.34, *p* < .01). The average score on the group report gained by the experimental groups was 74.38 (*SD* = 11.09), compared to an average of 63.66 for the control groups (*SD* = 9.94). This difference was also significant (*F* (1,83) = 22.70, *MSE* = 111.87, *p* < .001).

Effects on perceptions of students

In Table 3 the means and standard deviations of the student questionnaire are given. The scores of each variable were analysed according to a 2 (Groups) x 2 (Time of Testing) analysis of variance with repeated measures on the last factor.

Table 3. 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
Cluster: Instruction					
Satisfaction classes	*M*	3.10	3.99	2.99	3.85
	SD	0.51	0.36	0.53	0.44
Transparency classes	*M*	3.15	3.99	3.28	3.88
	SD	0.63	0.34	0.54	0.58
Learning access level	*M*	3.45	3.92	3.49	3.79
	SD	0.81	0.41	0.80	0.51
Practical relevance	*M*	3.52	4.14	3.57	4.01
	SD	0.84	0.39	0.69	0.68
Quality of the instruction	*M*	2.81	3.63	2.91	3.61
	SD	0.75	0.55	0.50	0.67
Cluster: Vision on instruction and assessment					
Relation instruction and assessment	*M*	2.18	3.89	2.37	3.85
	SD	0.72	0.42	0.59	0.74
Fear for assessment	*M*	2.70	2.28	2.88	2.14
	SD	0.91	1.25	0.99	0.88
Obtrusiveness assessment	*M*	2.41	3.39	2.49	3.49
	SD	0.56	0.56	0.57	0.57
Overall vision on assessment	*M*	2.80	3.89	3.17	3.91
	SD	1.02	0.74	0.88	0.84
Cluster: Role of student in assessment					
Involvement in assessment	*M*	3.20	3.42	3.13	3.28
	SD	0.45	0.41	0.44	0.49
Group behaviour	*M*	4.17	4.27	3.93	4.04
	SD	0.45	0.51	0.54	0.60
Collaborative learning	*M*	3.80	3.92	3.98	3.87
	SD	0.60	0.59	0.44	0.74
Assessment skill	*M*	3.82	3.80	3.69	3.73
	SD	0.41	0.35	0.37	0.42

There were highly significant main effects for the factor Time of Testing. This is the case for 11 of the 13 variables. All students – experimental and control group students together - were more positive in the post-test than in the pre-test about the classes (*F* (1,68) = 65.82, *MSE* = 13.74, *p* < .001), and the transparency of the study tasks (*F* (1,68) = 60.42, *MSE* = 13.83, *p* < .001). Students found the redesigned course more attuned to their learning access level (*F* (1,70) = 74.38, *MSE* = 21.14, *p* < .001), and were more positive about the practical relevance of the course (*F* (1,70) = 38.19, *MSE* = 12.76, *p* < .001), and the quality of the instruction (*F* (1,68) = 37.90, *MSE* = 9.74, *p* < .001). They were less afraid about the assessment (*F* (1,70) = 72.59, *MSE* = 53.95, *p* < .001), and regarded the assessment in the redesigned course as more obtrusive (*F* (1,69) = 184.99, *MSE* = 37.36, *p* < .001). Students felt more involved in the assessment (*F* (1,69) = 17.57, *MSE* = 3.13, *p* < .01. The relation between instruction and assessment became more apparent for the students (*F* (1,70)

= 64.01, *MSE* = 29.04, *p* < .001), and students' overall vision on assessment changed positively (*F* (1,69) = 45.02, *MSE* = 32.99, *p* < .001).

On the factor Groups two significant main effects were found. One on the variable 'involvement in assessment' (*F* (1,69) = 4.66, *MSE* = 1.29, *p* < .05), and one on the variable 'group behaviour' (*F* (1,69) = 8.34, *MSE* = 3.09, *p* < .01). In general the experimental groups outscored the control groups.

The interaction Time of Testing x Groups yielded no significant effects. The peer assessment training had no effect on students' perceptions.

Conclusion and discussion

This study was carried out to investigate the effects of a peer assessment training embedded in a course on 'Designing Creative Lessons', on the development of the peer assessment skill. A second research question was if this training had effects on students' the task performance in the course domain. A third question was how students perceived the redesigned course.

Regarding the first question, the analyses of the qualitative peer assessment reports clearly revealed that the training had the expected positive effects on the development of the peer assessment skill. The experimental groups were more likely to use the criteria and to give more constructive comments than the student teachers from the control groups. The students who received training also scored higher on structure and used less naive words. In spite of these positive results, student teachers cannot be regarded as expert assessors after this training. The experimental groups only applied nine of the fifteen possible criteria. Also, two constructive comments per assessment report indicates just a small progress. Therefore longer training periods seem to be recommendable.

On the second research question, an even more important result was found: A positive effect of the peer assessment training on the content related performance. The student teachers from the experimental groups outperformed the students from the control groups. The overall difference was significant, caused by significant effects on two of the four end products. The students of the experimental groups scored higher on the collection of assignments and the group report. The result on the group report was of interest, because this product was most closely related to the peer assessment tasks that focused on the design of creative lessons. In the group report students had to justify this design. A more profound training may lead to an effect on the performance of all the end products.

The answer on the third research question revealed that the results of the questionnaire that measured students' self-perceptions regarding the redesigned course were positive. All student teachers indicate a positive change in their view on several aspects of assessment, instruction and the role of the teacher. Assessment and instruction were more aligned in the redesigned course. Test anxiety decreased and the student teachers felt more involved in the assessment procedure.

The result that there is no effect on the variable 'assessment skill' is surprising, because this is precisely the skill the students were trained in. An explanation for this odd effect may be caused by the design of the study. A problem in the design of the presented study is that the pre-test-post-test comparison results of the student questionnaire may be an inaccurate measurement of the impact of the redesigned course, because the students may have limited knowledge about assessing at the beginning of the course that prevents them from accurately self-assessing baseline opinions (Sprangers, 1989). By the end of the course, their new understanding of peer assessment and the skills involved may have an impact on the responses on their self-evaluation.

Reflecting on the results, some recommendations for future research can be given. First, the training period has to be extended considerably, to attain full mastery of the assessment skill. Second, we have to be aware of the finding that using more criteria not necessarily means that the experimental groups apply the criteria *adequately and correct*. Further analyses of the assessment reports by content experts is recommended. Third, it is important to investigate peer assessment in different subject matter domains to test the generalizibility of training. Peer assessment is a skill that can have a life beyond the task for which it is employed in this study (James, 2000). Peer assessment can be used in comparable tasks within other domains. The question is: Can students assess products in other subject matter domains after a training in one particular domain, or do they need additional training in order to be able to assess in other domains?

Fourth, a suggestion is to begin training students in their first year of teacher training. Observation during informal evaluation of the peer assessment procedure with students showed that students are conservative and conditioned in their attitude towards teachers and assessment. They still feel that the teacher is the expert and the only objective assessor. When students are trained to rely on their own judgment and that of their peers, they may develop a belief that a teacher is first of all a coach, who supports and adjusts the decisions that students make. The effects of ongoing involvement in assessment and instruction on the development of these beliefs need to be further examined.

A last issue that deserves further elaboration is the relationship between content knowledge and assessing (see Mehrens, Popham, & Ryan, 1998). Peer assessing is a complex skill that cannot be demonstrated outside a particular subject matter domain. It can be hypothesized that students who are novices in a certain domain are also less capable to assess. One student stressed this issue in his evaluation: *"I have problems with assessing a product of a peer that is of a higher level than my own product. Then I realize that I don't have enough domain knowledge to criticize it."* The other way around can also occur: The more domain expertise a student has, the more capable he or she is to evaluate the work of peers. An interesting question for further research would be to what extent domain expertise influences the development of assessment skills.

In summary, it can be concluded that students can well be trained in assessment skills and that such training positively affects the performance level of student teachers. A thorough redesign of courses, whereby student teachers become more involved in assessment and instruction, leads to more satisfaction among students. Although student questionnaires should not be considered as determinative, the results of this study at least tell that the benefits of peer assessment may only be realised after serious effort is made to incorporate it into the everyday teaching practices in a way which is positive, non-threatening and attractive to students. As Boud et al. (1999) state: *"It is unrealistic to discuss assessment in isolation from curriculum content and teaching strategies"* (p. 424).

References

Arter, J. (1996). Using assessment as a tool for learning. In R. Blum, & J. Arter (Eds.) *Student performance assessment in an era of restructuring* (pp. 1–6). Alexandria, VA: Association for Supervision and Curriculum Development.

Boud, D. (1989). The role of self-assessment in student grading. *Assessment and Evaluation in Higher Education,* 14, 20–30.

Boud, D., Cohen, R., & Sampson, J. (1999). Peer learning and assessment. *Assessment and Evaluation in Higher Education, 24,* 4, 413-426.

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

Cutler, H., & Price, J. (1995). The development of skills through peer assessment. In A. Edwards, & P. Knight (Eds.), *Assessing competence in higher education* (pp. 150–159). Birmingham, UK: Staff and Educational Development Series.

Dochy, F.J.R.C., & Mcdowell, L. (1997). Assessment as a tool for learning. *Studies in Educational Evaluation, 23*, 279–298.

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer-, and co-assessment in higher education: a review. *Studies in Higher Education, 24,* 3, 331-350.

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,* 3, 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,* 2, 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*, 1, 53-70.

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

Keaten, J.A., & Richardson, M.E. (1992, February). *A field investigation of peer assessment as part of the student group grading process*. Paper presented at the Western Speech Communication Association Convention, Albuquerque, NM.

Kwan, K., & Leung, R. (1996). Tutor versus peer group assessment of student performance in a simulation training exercise. *Assessment and Evaluation in Higher Education, 21*, 205–214.

LPC (1995). *Beroep in beweging. Beroepsprofiel leraar primair onderwijs [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 assessments. *Educational Measurement: Issues and Practice, 17,* 1, 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. (2000). The use of student derived marking criteria in peer and self-assessment. *Assessment and Evaluation in Higher Education, 25,* 23–38.

Rushton, C., Ramsey, P., & Rada, R. (1993). Peer assessment in a collaborative hypermedia environment: a case study. *Journal of Computer-Based Instruction, 20*, 3, 75.

Sluijsmans, D.M.A., Brand-Gruwel, S., Van Merriënboer, J.J.G., Martens, R. (2001, submitted). *Redesigning education for training peer assessment skills in teacher training*.

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

Sluijsmans, D.M.A., Moerkerke, G., Dochy, F., & Van Merriënboer, J.J.G. (2001). Peer assessment in problem based learning. *Studies in Educational Evaluation, 27*, 2, 153-173.

Sluijsmans, D.M.A., & Van Merriënboer, J. J.G. (2000). *A peer assessment model.* Heerlen: Open University of the Netherlands. (Center for Educational Technology and Expertise).

Sprangers, M. (1989). *Response shift and the retrospective pre-test*. SVO: The Hague.

Stainer, L. (1997) Peer assessment and group work as vehicles for student empowerment: a module evaluation. *Journal of Geography in Higher Education, 21*, 95-98.

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.

Chapter 6

The training of peer assessment skills to promote the development of reflection skills in teacher education[1]

Abstract

The main purpose of this study was to investigate the effects of a training in assessment skills. 110 student teachers were trained in peer assessment skills within three courses on mathematics. After each course, students wrote a reflection paper, which was assessed by a peer. Students were allowed to improve their first reflection paper after the second and third course, based on peer feedback. The teacher assessed the final paper two weeks after the third course. Based on analysis of the written assessments and grades of the reflection papers, it appeared that the training led to a progress in students' skill to assess and an increase of the quality of the reflection papers. The results of a questionnaire show that students' views on assessment changed positively.

[1] Sluijsmans, D.M.A., Brand-Gruwel, S., Van Merriënboer, J.J.G. & Bastiaens, T. (2002, submitted). The training of peer assessment skills to promote the development of self-assessment skills in teacher education.

Introduction

Effective assessment approaches, based on constructivist views, receive special interest 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 no longer has purely the function of crediting students with recognised certificates but is above all valuable for the monitoring of students' progress and for supporting them in improving their learning activities. The emphasis shifts to a representation of assessment as a *tool for learning* (Arter, 1997; Boud, 1990; Dochy & McDowell, 1997).

Assessment is an important issue in the current developments towards more problem based and competency based learning. 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 study the focus lies on the level of students, more specific on the students as assessors of their own work and that of peers.

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, whereby 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 the 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 that has to be answered is how this partnership should be developed in educational practice.

To get more insight in how self and peer assessment should be implemented in teacher training, a number of studies were analysed (Sluijsmans, Dochy, & Moerkerke, 1999). Results of the analysis led to three basic assumptions that could be supportive for structural implementation of self and peer assessment.

The first assumption is that conducting a self or peer assessment is a complex skill, in which students have to be guided from the moment they enroll in higher education. Before putting students into the role of assessor, it is a prerequisite that students understand which skills are involved while making a judgment 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 to assess 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 are elaborated (Sluijsmans, Dochy, & Moerkerke, 1999; Sluijsmans & Van Merriënboer, 2000). In this model, three main skills are taken into account. These skills are (1) defining assessment criteria; (2) providing 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, Dochy, & Moerkerke, 1999).

A second assumption is that the training of self and peer assessment might have positive effects on 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 as an isolated skill, 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, thus the assessment training will support students' development of their presentation skills.

The last assumption states that developed *reflection skills* are a prerequisite for conducting reliable self and peer assessments. To assess the work of themselves or a peer, students have to employ reflection skills to recognise strengths and weaknesses. In this view self and peer assessment are strongly related: analysing 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, 1995; Kremer-Hayon & Tillema, 1999; Newman, 1996; Richert, 1990). 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 the actual teaching performance (e.g., Anderson & Freiberg, 1995). Schön (1987) in this context distinguishes '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 on reflection of teaching (reflection during the act of planning 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 structure both 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, which help 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 chosen in which the assumptions mentioned above were applied. A training in assessment was embedded in three first-year courses on mathematics. The decision for selecting these specific courses was that the mathematics teacher has been using reflection papers for several years in these courses. 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 a routine behavior 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 criteria for a good reflection paper, how to give feedback regarding the reflection paper of a peer, and how to write an assessment report of the reflection paper of a peer.

Three existing first-year courses on mathematics were analysed to make them suitable for the assessment training. Two research questions were chosen to explore in depth.

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: the use of 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.

For the second research question it is explored what the students' perceptions are 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 change students' views positively towards mathematics and that students' 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 (*SD* = 2,7). The group of students was heterogeneous in educational background.

Materials

Courses. 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). Within the three courses students were confronted with basic skills that are required for teaching mathematics to pupils. Students wrote a personal reflection paper after each course.

Peer assessment training. 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 didactics 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 of 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 course. 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 was developed to analyze the quality of the peer assessments that were written by the students. Prior studies were used for the definition of this rating form (Sluijsmans, Brand-Gruwel, Van Merriënboer, submitted; Sluijsmans, Brand-Gruwel, Van Merriënboer, & Martens, submitted). The rating form is derived from an expert assessment report, which represents the expert peer assessor. In this rating form seven variables are included consisting of 36 dichotomous items. These variables are: the use of criteria (1), positive comments (2), negative comments (3), constructive comments (4), posing questions (5), structure (6), and naive word use (7).

Each student could score a maximum of 19 on the first variable 'use of criteria', because the students defined 19 criteria in the first part of the peer assessment training that focused on the definition of criteria for a reflection paper. The variables concerning 'giving feedback' were measured by counting the number of positive, negative and constructive comments and the number of posed questions the student used in the peer assessment. Two variables were related to 'writing an assessment report'. Six items in the rating form were included to measure the variable 'structure'. One item for example checks if the student starts his assessment report with a positive comment. The variable 'naive word use' was measured by counting the words such as nice, good, excellent, and fine.

Finally, three dichotomous items were included to obtain some additional information about the effects of the training. One item was only measured in 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 moments 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 with the rating form. For each variable the interrater-reliabilities were calculated. These reliabilities were acceptable for all variables (Cohen's Kappa > .95).

Examinations. To measure an 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 analysed. The score could range from 0 to 100. It was decided that a score of 75 was given when the student used the criteria. Extra 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 – in the intake - and after the third course – in the outtake - the students filled out a questionnaire about their perceptions on mathematics, instruction and assessment. 116 items were divided among twelve variables. Four variables were related to expectations regarding mathematics, four were related to expectations regarding assessment of mathematics and another four were related to the assessment skill. The students had to answer the items on a five point Likert-scale, varying 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. Because of low corrected item-total correlations, 11 items were removed from the questionnaire. The clusters, variables, number of items, reliability coefficients and example of items are presented in Table 1.

Table 1. Clusters, variables, number of items, reliability coefficients and example items of the twelve variables of the student questionnaire

Variable	#	α	Description	Example items
			Cluster: Expectations regarding mathematics	
Role of the mathematics teacher	7	.65	Perception of the activities a mathematics teacher should perform	The mathematics teacher has to stimulate creative thinking
Self esteem of the student	16	.74	Self-perception on learning mathematics	I'm capable to work independently on assignments
Vision on mathematics and didactics	16	.68	Vision about mathematics and didactics	Pupils learn mathematics by working in groups
Mathematics skills	7	.85	Self-perception on their mathematics skills	I'm capable to help peers with their mathematical problems
			Cluster: Expectations regarding assessment of mathematics	
Fear for assessment	6	.86	Students' fear to take exams	I'm usually very nervous before taking an exam
Obtrusiveness assessment	5	.60	Perception of the extent in which mathematics exams are public to them	The questions on an exam have to be public to students before the exam
Predictability of assessment	4	.77	Perception of the extent in which mathematics exams are predictable	I know before I take the test if I will pass or fail
Involvement in assessment	7	.82	Perception of the extent in which they want to be involved in assessment	I think that students should have more participation in the development of assessment criteria
			Cluster: Assessment skill	
Group behaviour	4	.68	Perception of peers' group behaviour	I don't like it when students don't make a individual contribution in groups
Collaborative learning	8	.84	Perception of collaborative learning	I prefer to elaborate on problems with my peers
Peer assessment	7	.74	Perception on peer assessment	Peer assessment is useful
Assessment skill	18	.84	Self-perception on their assessment skills	I'm able to analyse a product of a peer

Teacher Interview, Field Notes and Students' Reflection Papers. Additional qualitative data was 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 analyse reflections of the mathematics teacher 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 set up according to a within subject repeated measures design. Students participated for a period of seven months in the experiment. 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 questionnaire, writing a reflection paper about prior experiences in mathematics, and assessing an anonymous reflection paper. This anonymous reflection paper was previously marked as an 'unsatisfactory' one by the mathematics teacher. After the intake, all students attended three successive courses on mathematics. During each course, the students worked on products that were content related. Besides that, the students had to write a reflection paper after the first course, which could be improved after the second and third course to submit the final version of the paper two weeks after the last feedback session. All students received training in the assessment skill during the courses. 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 this 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, work field experiences, personal expectations, and strengths/weaknesses.

In the second training, 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 their 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 acted 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. After 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 the 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 in 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 the test anxiety and to lengthen the period in which the peer assessment skills were trained, students received no grades of the mathematics teacher for their reflection paper after each course. The role of the teacher was limited to a coaching and chairing one in the feedback sessions. The reflection paper that was written based on the given peer feedback 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 same reflection paper that was presented in the intake session.

Data-analyses

Two research assistants analysed 550 peer assessment forms gathered at five testing moments, 220 forms from the intake and outtake and 330 gathered after 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 analysed separately from the peer assessment forms written after the three courses. The reason was that the conditions in which the students wrote these peer assessments were not similar to the conditions of the other three testing 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 in their own pace at home.

A non-parametric test for two related samples (Wilcoxon) 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 analysed 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 test 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 analysed to see if any overall effects occurred. These reports were analysed 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) was applied. This analysis revealed that the training had a positive effect on six of the seven variables. In the intake for example students only used less than one criterion (M = .57), while this number significantly increases in the outtake (M = 5.0). Because there was no significant effect on the variable 'naïve word use', this variable was excluded from further analyses.

Table 2. Means, standard deviations and z-values of assessment forms after intake and outtake

Variable	Intake		Outtake		
	M	*SD*	*M*	*SD*	*z*
Defining criteria					
use of criteria[a]	0.57	.70	5.00	4.32	-8.15*
Giving feedback					
positive comments	0.86	.35	1.41	1.36	-3.77*
negative comments	1.17	.73	2.73	2.22	-6.37*
constructive comments	0.29	.46	1.96	1.88	-7.25*
posed questions	0.12	.33	1.69	0.81	-4.26*
Writing an assessment report					
structure[b]	.075	.45	2.46	1.21	-8.77*
naive word use	1.62	.76	2.17	1.86	-.17

* *p* < .001.
[a]maximum score = 19
[b]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 ANOVA's reveal which increase or decrease of the means presented in Table 2 was significant. Beforehand, it was tested if the assumptions 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.

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

Variable	Course I		Course II		Course III				
	M	*SD*	*M*	*SD*	*M*	*SD*	*MSE*	*df*	*F*
Defining criteria									
use of criteria[a]	9.25	4.60	6.35	4.68	10.45	5.27	19.31	(2, 210)	24.97*
Giving feedback									
positive comments	2.75	2.94	3.31	2.67	6.08	3.72	7.60	(2, 210)	45.64*
negative comments	4.28	4.03	2.50	2.73	4.04	2.73	8.26	(2, 208)	13.23*
constructive comments	2.05	1.01	2.28	1.81	2.96	2.72	4.41	(2, 210)	24.79*
posed questions	1.71	.69	1.92	.88	3.56	1.96	5.07	(2, 210)	10.56*
Writing an assessment report									
structure[b]	2.07	.97	2.55	.90	4.19	1.44	1.18	(2, 210)	109.5*

* *p* < .001.
[a]maximum score = 19
[b]maximum score = 6

'Defining Criteria'. On the variable 'use of criteria' a significant decrease of the number of criteria used can be noticed from the training in the first course (*M* = 9.25), to the second course about giving 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)). In Figure 1a it is illustrated how students develop in the skill 'use of criteria'.

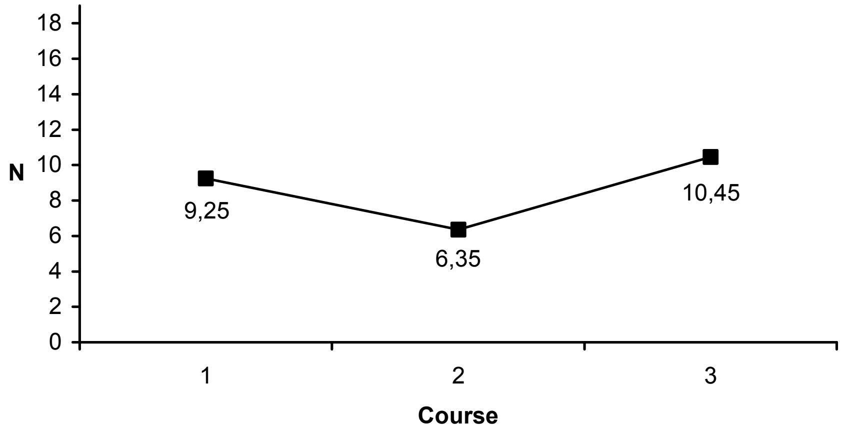


Figure 1a. Development in 'Defining of Criteria'

'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 found from the first course (*M* = 4.28) to the training in the second course about giving 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 writing an assessment report, *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 the training 'giving feedback' (*M* = 2.28), *F* (1,105) = 7.47, *MSE* = 9.56, *p* < .01), and 'writing an assessment report' (*M* = 2.96), *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). The passage of the four variables related to giving feedback are drawn in Figure 1b.

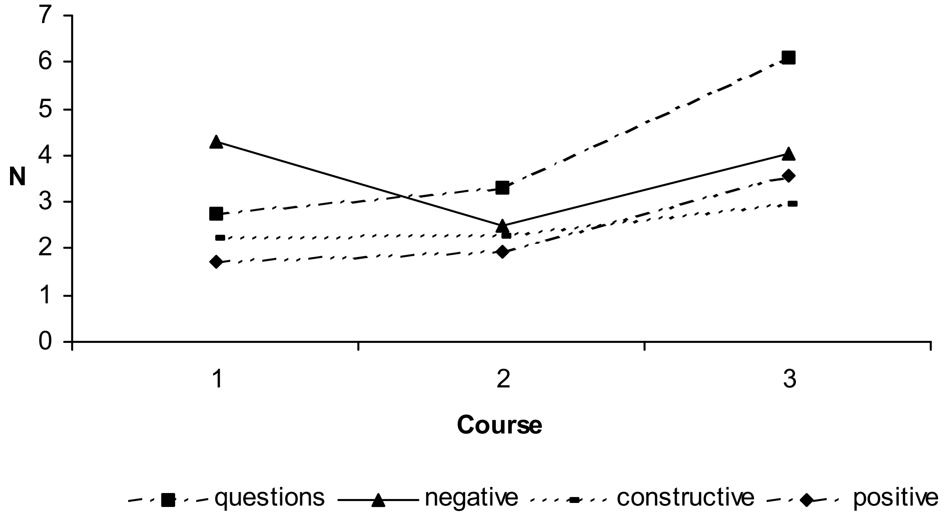


Figure 1b. Development in 'Giving Feedback'

'Writing an assessment report'. 'Structure' increases significantly from the first course (*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). From Figure 1c it can be derived how students develop in their skill of writing an assessment report.

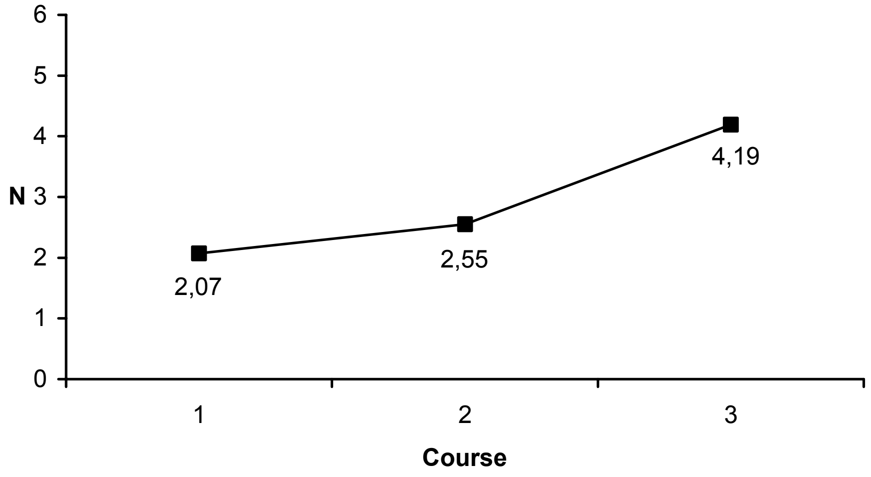


Figure 1c. Development in 'Writing a Structured Assessment Report'

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 about defining criteria (first course), and that this percentage decreased 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 confirmative 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 the 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.

Research question II: Change in perceptions

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

Table 4. Means and standard deviations of the students questionnaire results at the intake and outtake on a five-point Likert-scale (N=110)

		Intake	Outtake	*t*
Cluster: Expectations regarding mathematics in teacher education				
Role of the mathematics teacher	*M*	3.32	3.59	-4.76***
	SD	0.46	0.47	
Self esteem of the student	*M*	2.30	2.41	-1.87*
	SD	0.48	0.49	
Vision on mathematics and didactics	*M*	2.35	3.14	-13.38***
	SD	0.40	0.48	
Mathematics skills	*M*	3.33	3.59	-2.58***
	SD	0.84	0.74	
Cluster: Expectations regarding assessment of mathematics in teacher education				
Fear for assessment	*M*	2.98	2.58	3.80***
	SD	0.88	0.84	
Obtrusiveness assessment	*M*	3.35	3.56	-1.70*
	SD	0.67	1.21	
Predictivity of assessment	*M*	2.73	3.04	-2.77***
	SD	0.85	0.90	
Involvement in assessment	*M*	3.14	3.26	-1.29
	SD	0.74	0.68	
Cluster: Assessment skill				
Group behaviour	*M*	4.22	4.37	-2.42**
	SD	0.52	0.48	
Collaborative learning	*M*	3.40	3.25	1.67*
	SD	0.72	0.67	
Peer assessment	*M*	4.23	4.27	-.77
	SD	0.42	0.42	
Assessment skill	*M*	3.62	3.87	-4.61***
	SD	0.43	0.41	

$^{*}p < .05.$ $^{**}p < .01.$ $^{***}p < .001$

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 vision on mathematics and didactics, a higher self-esteem and they feel more skilled in mathematics. Their fear of assessment decreased significantly, they evaluate the assessment as more open and predictive. Student's perceptions towards the group behaviour changed positively. The students indicated that they are more capable in assessing 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 assessment skills of students and their skills on writing reflection papers as an effect of assessment training that was integrated in three mathematics courses. It was also measured how students' views on several aspects on mathematics and assessment changed in 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. In the following the results found will be summarised 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 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 counts for 'structure', of which only a score of 2.46 was reached of the six possible points. 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.

An effect on the variable 'naïve word use' failed to occur. A first explanation could be that if students write longer assessments, it is logical that more of those words will be used. Second, it may relate to the limited vocabulary of first-year students, who tend to use words as 'nice' and 'good' regularly. An third explanation could be that the use of naïve words is not necessarily an indication for poor assessment skills, if the other important skills are developed on a satisfactory level. In other words, if a student uses the criteria adequately, makes constructive comments, and structures his assessment report, it is less important how often he uses naïve words. For these reasons, it is recommendable to reconsider this variable as a superfluous one in determining the level of assessment skill.

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 that 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 training part about defining criteria. However, this decreased after the second course. An explanation could be that the criteria 'faded away' for some period, because the attention was focused on the second part of the assessment training about giving feedback. In conclusion, students reached a higher level on all skills after the third course compared to the first course. Besides a growth in their 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 after each course to conduct the peer assessment. They also felt that their level of self-esteem to solve mathematical tasks increased over the courses. This is also supported by the quantitative data. An important finding is that their test anxiety decreased. Students who developed a negative attitude towards mathematics changed in their perceptions by using reflection and peer assessment. Students' attitude towards mathematics often differ from their attitude towards other domains, like for example pedagogy or history. First year student teachers are highly influenced in their perception on mathematics by their experiences in secondary education. While some of these students enjoy mathematics, a lot of students have less positive experiences (Ashcraft, Kirk, & Hopko, 1998; Fennema & Sherman, 1978; Jacobs, Watson, & Sutton, 1996). 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 choose a profession a 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 mathematical anxiety and increasing mathematical skills.

Another explanation for the decrease of 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). In the approach of formative assessment as a learning and feedback tool, test-driven behaviour is reduced. Dalziel (1998) adds to this that it is important that students are informed about the assessment that will be used in the courses and that they are involved in the decision making process about instructional issues. In other words, assessment is broader defined than just a paper and pencil test.

During the feedback sessions, that took place three times, students reported their written assessment reports orally to the peer. In the first feedback session, where the students had to report their assessment report to the rest of the group for the first time, student felt very insecure to report their findings to the assessed peer. Some students were very nervous and showed physical signs like sweating, stuttering en red cheeks. 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, moreover because they did not receive grades for their reports. The teacher explains 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 beamed 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 a peer assessment report or did not make a constructive contribution in the group, this was noticed and criticised by the other peers. As far as this I concerned, it was nice to observe that the role that I used to have now shifted to the students".*

Although the majority of the findings supported the hypotheses, some limitations of this study are in order. 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 a higher quality of the skill to write a reflection paper and the skill to assess. Two prior studies however showed that a short training period leads to only small differences in development between trained groups and control groups (Sluijsmans, Brand-Gruwel, Van Merriënboer, & Martens, submitted; Sluijsmans, Brand-Gruwel, & Van Merriënboer, submitted). Because of the length of the study, it is possible that changes in perceptions are caused by other external factors. The students attended other courses in the same period and developed experience in teaching at elementary schools.

A second limitation lies in the approach that was chosen. Peer assessment in this study was most of all considered 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 less quality than his own paper. The opposite also occurred: a student who had little knowledge about the meaning of the criteria of a reflection paper, was less capable in giving constructive feedback. Regarding this issue, the mathematic teacher's view was that a student may not benefit from feedback they receive *from* peers who have insufficient knowledge to give constructive comments, but that the student will learn by giving feedback *to* peers. The question in how far writing reliable assessments requires knowledge about the criteria of a product, is an interesting issue that deserves further elaboration in future research.

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

Supporting students to developing assessment skills can play an important role in the ongoing evolution of teacher education towards skill-based curricula. Results from the presented study reveal that students can be trained in their assessment skills and that this positively influences the quality of the reflection papers. The peer assessment tasks and the feedback sessions encouraged the students to explore concerns involved in giving and getting feedback on writing reflection papers. Giving helpful feedback to others may also develop students' interpersonal skills and assists 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 from and learn from others – in the classroom, on the job, or in other areas of their lives.

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 vague and differ per 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; Reilly-Freese, 1999; Hayon-Kremer, & Tillema, 1999; Korthagen, 2000), 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 supports in making these perspectives more explicit.

References

Anderson, J.B., & Freiberg, H.J. (1995). Using self-assessment as a reflective tool to enhance the student teaching experience. *Teacher Education Quarterly, 22,* 77–91.

Arter, J. (1996). Using assessment as a tool for learning. In R. Blum, & J. Arter (Eds.) *Student performance assessment in an era of restructuring* (pp. 1–6). Alexandria, VA: Association for Supervision and Curriculum Development.

Ashcraft, M.H., Kirk, E.P., & Hopko, D. (1998). On the cognitive consequences of mathematics anxiety. In C. Donlan (Ed.), *The development of mathematical skills*. Hove, England: Taylor & Francis.

Bangert, A.W. (1995). *Peer assessment: an instructional strategy for effectively implementing performance-based assessments*. University of South Dakota.

Boud, D. (1990). Assessment and the promotion of academic values. *Studies in Higher Education, 15*, 101–111.

Dalziel, J. (1998). Using marks to assess student performance: some problems and alternatives. *Assessment and Evaluation in Higher Education, 23,* 351-366.

Dochy, F., & McDowell, L. (1997). Assessment as a tool for learning. *Studies in Educational Evaluation, 23*, 279–298.

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.

Fennema, E., & Sherman, J. (1978). Sex related differences in mathematics achievement and related factors: A further study. *Journal for Research in Mathematics Education, 9,* 189-203.

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.

Kelly, W.P., & Tomhave, W.K. (1985). The study of math anxiety/math avoidance in preservice elementary teachers. *Arithmetic Teacher, 32,* 51-53.

Korthagen, F.A.J. (1985). Reflective teaching and pre-service education in the Netherlands. *Journal of Teacher Education, 36,* 11-15.

Korthagen, F.A.J. (2001). *Linking practice and theory: the pedagogy of realistic teacher* education. Mahwah, NJ: Lawrence Erlbaum Associates.

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* (pp. 131-141). 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.

Loughran, J.J. (1996). *Developing reflective practice*. UK: Falmer Press.

Madaus, G.F. (1988). The influence of testing on the curriculum. In L.N. Tanner and K.J. Rehage (Eds.), *Critical issues in curriculum* (pp. 83-121). Chicago: The University of Chicago Press.

Mehrens, W.A., Popham, W.J., & Ryan, J.M. (1998). How to prepare students for performance assessments. *Educational Measurement: Issues and Practice, 17,* 1, 18-22.

Newman, S.J. (1996). Reflection and teacher education. *Journal of Education of Teaching, 22*, 297-310.

Patton, M.Q. (1990). *Qualitative evaluation and research methods*. Newbury Park, CA: Sage Publications.

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

Richert, A.E. (1999). Teaching teachers to reflect: a consideration of programme structure. *Journal of Curriculum Studies, 22,* 509-527.

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

Sluijsmans, D.M.A., Brand-Gruwel, S., Van Merriënboer, J.J.G., Martens, R. (2001, submitted). *Redesigning education for training peer assessment skills in teacher training*.

Sluijsmans, D.M.A., Brand-Gruwel, S., Van Merriënboer, J.J.G. (2001, submitted). *Peer assessment training in teacher education.*

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

Sluijsmans, D.M.A., 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.M.A., & Van Merriënboer, J.J.G. (2000). *A peer assessment model.* Heerlen: Open University of the Netherlands. (Center for Educational Technology and Expertise).

Stefani, L. (1998). Assessment in partnership with learners. *Assessment and Evaluation in Higher Education, 23,* 339-350.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Sullivan, P. (1989). Factors inhibiting change: A case study of a beginning primary teacher. *Research in Mathematics Education in Australia,* August, 1-9.

Trujillo, K. M., & Hadfield, O. D. (1999). Tracing the roots of mathematics anxiety through in-depth interviews with preservice elementary teachers. *College Student Journal, 33,* 219-232.

Watson, J. (1987). The attitudes of preservice primary teachers toward mathematics: Some observations. *Research in Mathematics Education in Australia,* August, 48-56.

Chapter 7

General discussion

The curricula in higher education are subject to reform, based on ongoing changing demands in the labour market. Alternative ways in instruction and assessment receive growing interest. Within these alternative approaches, it is generally agreed that students are considered as active participants in performance assessments, who are able to formulate learning goals in collaboration with their peers. Introducing self and peer assessment activities may contribute to the development of reflective and responsible practitioners, on condition that these activities are embedded in educational practice. As a consequence, existing course material needs to be substantially revised, whereby the theory of constructive alignment, student involvement and instructional design is taken into account.

Five studies were conducted to gain more insight in the issue of student involvement in assessment in relation to performance and perceptions. Each experimental study was conducted within a teacher training context, in which the skill to assess peers' work is considered to be important. By means of a literature study, a pilot study and three experimental studies, it was possible to elaborate on the integrated theoretical framework outlined in Chapter 1.

In this final chapter, the main results of each study are summarised. The results are reflected upon by presenting four main considerations in the design of peer assessment. Limitations of the studies are provided and suggestions for further research are given. The final section of this dissertation provides some 'rules of thumb' for implementing (peer) assessment in teacher education and higher education in general.

Summary of the main results

The literature review enclosed an analysis of self-, peer- and co-assessment practices in higher education. The development of evaluative and critical abilities and possibilities for student collaboration seemed to be the main advantages of student involvement. However, training in the skill to self assess or to peer assess has to be provided for at least beginning students. This training would have an optimal impact on learning if it is embedded in carefully (re)designed learning environments (Frederiksen, 1994; Pilot, 1997; Topping, 1998).

Chapter2 reported a pilot study in which the reliability of peer ratings was investigated. Overall conclusion was that the peer assessment method used could be improved. Peer assessment the way it was conducted in the studies did not prevent rating errors like friendship marking, resulting in over-marking; collusive marking, resulting in a lack of differentiation within groups; decibel marking, where individuals who dominate groups get the highest marks; and parasite marking, where students fail to contribute but benefit from group marks (Pond, Ul-Haq, & Wade, 1995). Recommendations were directed at more training, supported by Arter (1996) and Perkins (1986), and a more qualitative approach in peer assessment.

Findings from the first part of the dissertation that consisted of a literature study and a pilot study, led to the design of an integrated framework. This framework contained the guidelines for the design of the three following studies. For clarification of the main results of these studies, the framework is repeated in Figure 1.

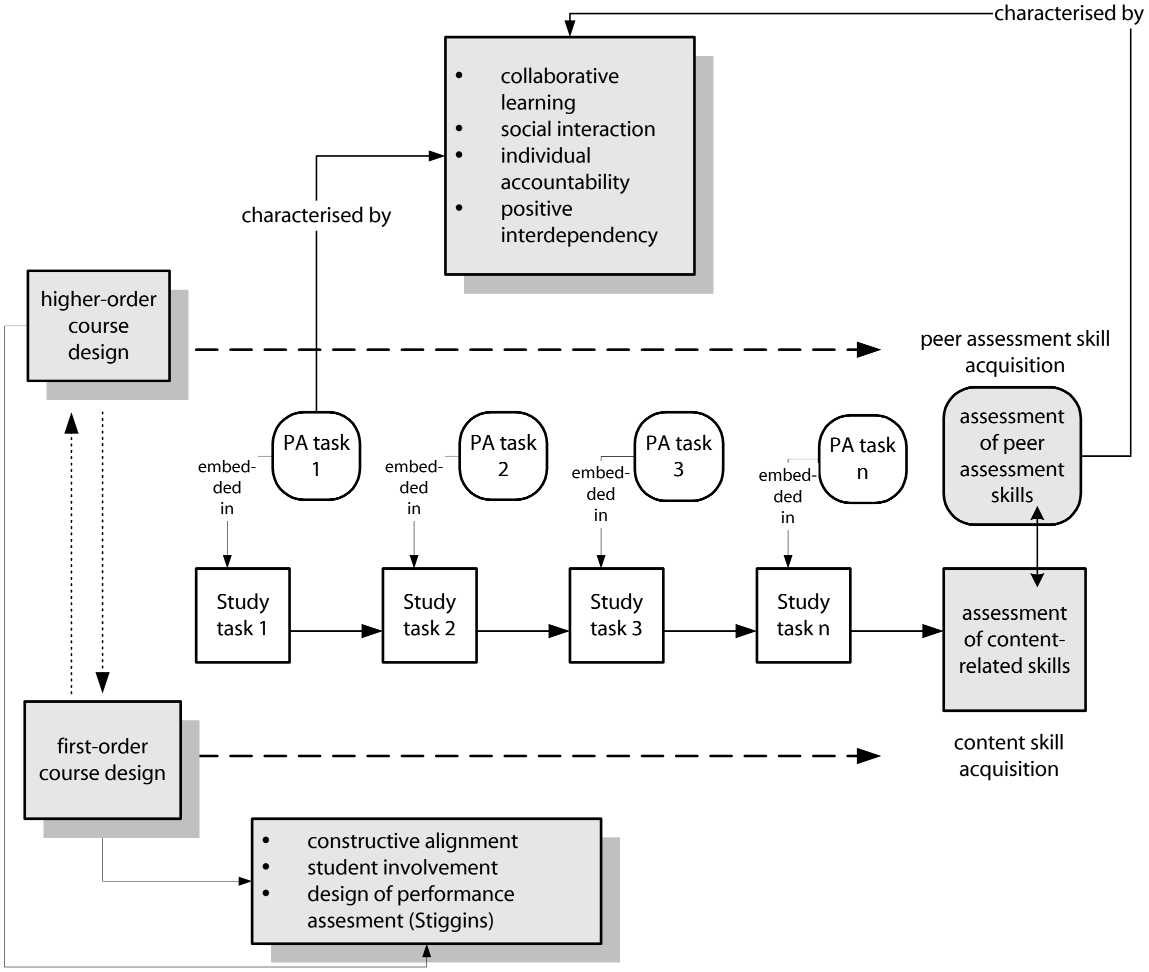

characterised by
collaborative learning
social interaction
individual accountability
positive interdependency
characterised by
higher-order course design
peer assessment skill acquisition
PA task 1
PA task 2
PA task 3
PA task n
embed-ded in
assessment of peer assessment skills
Study task 1
Study task 2
Study task 3
Study task n
assessment of content-related skills
first-order course design
content skill acquisition
constructive alignment
student involvement
design of performance assesment (Stiggins)

Figure 1
Student involvement and course design for powerful learning environments – an integrated framework

The objective of the studies reported in Chapters 4, 5 and 6 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. It was also explored what the effects of the redesign of the course in a more skill-based way were on several variables, based on student's perceptions. Teacher perceptions were taken into account regarding the process of the peer assessment training and the redesign of the course.

Positive findings were found regarding the first question whether training in peer assessment by means of peer assessment tasks, leads to the development of the skill to assess the work of peers. This is illustrated by the 'higher-order course design path' in the integrated framework presented in Figure 1. It can be concluded that each of the three studies led to an improvement of the skill to assess, and that this progress was the least in the study reported in Chapter 4, and increased in the following studies in which was chosen to train the whole skill over a longer period of time. This means that it is possible to train students in their assessment skills.

The effect of training peer assessment skills on students' content-related performance is illustrated by the 'first-order course design path' in the integrated framework presented in Figure 1. At first, a difference between the performance quality of the students from the control and from the experimental group failed to occur (Chapter 4). Explanations were sought in the small progress in the peer assessment skill and the short training period. Moreover, it was hypothesized that the redesign of the course had an effect on the learning result of all students. In the next study, in which the whole peer assessment skill was trained, a positive effect of the peer assessment training on the actual learning results was found (Chapter 5). The student teachers from the experimental groups outperformed the students from the control groups. This same result was found in the last study reported in Chapter 6, where the whole group of students wrote better reflection reports after the training than before the training.

The intervention of a peer assessment training had several implications for students. Overall results of the student questionnaire in each study were that the whole group was more positive about the instruction and the integration of assessment and instruction after they participated in 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 view on aspects of learning and assessment. They are more satisfied about the classes and the criteria and goals are clearer. The role of the teacher is also evaluated in a more positive way. The student teachers indicated that they are more capable in assessing than before the redesigned course. Test anxiety decreased and the student teachers felt more involved in the assessment procedure. In the interviews though students also indicated that they did not feel like an expert-assessor after the training (Chapter 4). Comments from the students in their reflection papers (Chapter 6) implied that they felt less intimidated after each course to conduct the peer assessment. They also felt that their level of self-esteem to solve mathematical tasks increased over the courses. This is also supported by the quantitative data. An important finding is that their test anxiety decreased. Students who had developed a negative attitude towards mathematics changed in their perceptions by using reflection and peer assessment (Chapter 6).

A result that was found in the study reported in Chapter 4 was that the content domain influenced students' perceptions. This finding was mainly caused by the domain mathematics. Explanations for this effect could be sought in the specific character of the mathematics domain.

Teacher experiences were investigated in the study reported in Chapter 4 and in the longitudinal study (Chapter 6). In both studies, it appeared that the teachers had no major problems in instructing the peer assessment tasks. They indicated that implementing the peer assessment training led to a rethinking of the existing course and stimulated them to view the content from a different perspective. The need for revision of courses that lasted for several years though did also lead to some resistance. Some teachers doubted the value of the peer assessment and were sometimes reluctant to give up some part of their content expertise on behalf of the 'higher order' skills. The role of the teachers and these feelings of resistance are further discussed with regard to future research. First, four main decisions in the design of peer assessment are given.

Four main considerations in the design of peer assessment

Four considerations were central to the reported work. The first one concerns the decision for assessing products or processes of peers. The next two considerations were if students had to assess peers in a qualitative or quantitative way and if this was done by means of a written or an oral peer assessment report. A last consideration was if the peer assessments had to be conducted in a non-anonymous setting. Each consideration is clarified in the following.

Product versus process peer assessment

When students assess their peers, the object of assessment is a certain product or a process evaluation. In Chapter 2, a study was reported in which students assess one another on aspects as active contribution to the group discussions, quality of these contributions, preparedness to be involved in tasks, and actual contribution to the teamwork. This process-oriented approach is very suitable when the free-rider effect – students who do not participate fully in problem based groups – occurs (William, 1999). Peer assessment offers students the opportunity to express their dissatisfaction by rating free-riders low on the scale. The possibility for this kind of implementation of peer assessment however is not pursued in the three following studies presented in Chapters 4, 5, and 6 Peer assessment should not become a tool for sanctions. It may be useful to reduce free-rider effects, but in this case the peer assessment does not function as a learning tool. The majority of the studies presented in the literature study (Chapter 2) are product oriented, where students assess their peers on products as presentations, video's and essays (e.g., Conway, Kember, Sivan, & Wu, 1993; Falchikov, 1986; Falchikov, 1995; Fry, 1990) in a quantitative way. The difference between quantitative and qualitative peer assessment is described in the following.

Quantitative versus qualitative peer assessments

Peer assessment is mostly applied as a quantitative method where student and teacher marks are compared (Falchikov & Goldfinch, 2000). The studies reported in Chapter 2 were directed at a quantitative approach in which students rated one another on four criteria. Nominations, rankings and ratings however, have been found to create quite strong adverse reactions (Hanrahan & Isaacs, 2001; Kwan & Leung, 1996; Rushton, Ramsey, & Rada, 1993). This was also a finding of the studies reported in Chapter 2. Because students expressed their need for more detailed instruction beforehand, it was decided to choose a qualitative approach in the subsequent studies, in which the focus was on peer assessment as a learning activity. In that way, peer assessment did not aim at scoring peers on a number of criteria, like in many peer assessment studies (see Boud & Falchikov, 1989; Falchikov & Goldfinch, 2001), but on the quality of peer assessments of individual student teachers. According to this perspective, conducting a peer assessment is regarded as a skill on its own which subsequently can be assessed by the teacher.

Written peer assessments versus oral peer assessment

To find an answer to the question whether students can be trained in their peer assessment skills, students wrote assessment reports that were analysed. These reports varied from one page to sometimes eight pages. Only in the last study (Chapter 6), students amplified their written comments orally in feedback sessions. On the one hand, based on additional comments in the questionnaires, students indicated that they found it difficult to express their feedback in writing. Observations from the feedback sessions on the other hand revealed that when students have to report their feedback face to face in a group, this also entails insecure feelings. In this perspective, it is interesting to discuss if students have to give feedback orally, in writing, or both. In teacher training, both writing and presentation skills are important. If students are very competent in their writing skills, it is possible that they pass exams, because they are capable of 'writing what the teacher wants to read'. This for example is the case when students have to write essays or reflection papers. This however does not make a student a good teacher. The opposite is also possible: students that have strong verbal skills, can easily convince others or dominate others, for example in problem based learning groups. What is said however, does not have to be appropriate in the context.

In an oral performance, the advantage is that the students are present and interacting with their peers so that the interpretation of the performance is necessarily a joint product of the students and the peers. This was particularly stimulated in the peer assessment tasks, in which students had to discuss criteria and feedback rules. Based on the experiences of the last study, in which the differences between written and oral peer assessment became clearly visible, the preference is given to a combination of the two modes of communication. In this case, students can practice both writing and communication skills.

Anonymous versus non-anonymous peer assessment

A discussion that reared its head during the design of the experiments, was if the peer assessment should or should not take place in an anonymous setting. In the study reported in Chapter 2, students rated their peers in a quantitative way on four criteria. It was not brought into the open how each individual rated the peers, and the study therefore was anonymous. Reflecting on the results of this study, it was decided for a semi-anonymous peer assessment, which implied that students received the written assessment reports of their peers after the peer assessment that was conducted in presence of the assessees. Students were assessed as a dyad (Chapter 4) or as a group (Chapter 5) by all of their peers. In the mathematics study (Chapter 6), a non-anonymous approach was taken. Students were individually assessed by one other peer. Every student received a 'who assesses who-scheme'. An additional activity compared to the prior studies was that students also had to report their peer assessment orally in a feedback session.

It was a deliberate choice to shift from an anonymous to a non-anonymous peer assessment. Teachers who work together, have to learn from each other (Verloop & Wubbels, 2000). Student teachers have to get used to an open discussion about criteria and to giving constructive feedback face to face. Opponents argue that knowing the peer and subjective opinions about the peer affect the objectivity of peer assessments. Students' vary in their opinions about anonymity. While some students indicated that they prefer assessment not to be anonymous (e.g., Pâquet & Des Marchais, 1998), students in the studies presented in this dissertation found it often difficult to write a peer assessment to a student who is aware that he is assessed by that student (Chapter 6).

Regarding the results of studies that compare face-to-face situations with more anonymous settings (e.g., Jonassen & Kwon, 2001; Marttunen, 1992), it can be further discussed whether peer assessment should be anonymous or not. This decision may be dependent on the goals aimed at. For the studies is this dissertation, it is a difficult point, because both social factors and the training of argumentative skills are of importance. The development of social skills prefer a non-anonymous setting, while research advocates an anonymous setting for the development of argumentative skills. In conclusion, a combination of the two seems desirable.

Limitations and future research

The studies in this dissertation, the design, certain decisions that were taken and the results found, uncover the limitations of the studies and evoke even more questions than there were before. Therefore, it is interesting to discuss the subjects that need further attention.

Reconsidering the peer assessment model

For the design of the peer assessment tasks, the peer assessment model was directive. Looking back upon the results, some adjustments of the model can be suggested.

First, the model as it is formulated for the studies does not clearly distinguish between assessment of process and assessment of product, quantitative or qualitative peer assessment, and if the peer assessment is anonymous or not. Second, the model moves directly into peer assessment. Experts in the field hold the opinion that some self assessment is necessary before students can engage in peer assessment. The study reported in Chapter 6 showed that self and peer assessment are interrelated. It is interesting to analyse which skills are conditional for peer assessment. Besides self assessment skills, one could think of social and collaborative skills. Third, it may be unclear to what extent this is a generic model of skills needed for peer assessment, or whether it is just meant to analyse and develop the skills needed in a particular context. At the broadest level the specified skills are always needed, but when it gets to more detailed levels, this is specific to the way in which peer assessment will be conducted. The difference between skills and activities is hard to make. At last, the model does not distinguish performance levels. Sometimes it is quite easy to understand what a criterion means in general terms. For example, when students are peer assessing an oral presentation a criterion might be 'maintains a good relation with the audience, using eye contact, movement and gestures'. Most students can understand what this means, but what is a 'good enough' level of performance on this criterion so that the student can pass? In the studies reported in Chapter 4, 5, and 6, a rating form was developed to measure students' assessment skills, but it needs further research to determine what level of proficiency for example a first-year student-teacher has to reach.

Knowledge about assessing versus knowledge about content

An interesting issue is the relationship between content knowledge and assessing (see Mehrens, Popham, & Ryan, 1998). Peer assessing is a complex skill that cannot be demonstrated outside a particular subject matter domain. It can be hypothesized that students who are novices in a particular domain are also less capable to assess in this domain. The skill to assess may probably vary considerably according to context. There are some situations where it would be quite easy to peer assess – to define criteria and judge performance – because the performance is very straightforward, it is easy to know what is right and wrong. However, in other cases much more judgement and domain knowledge is needed on the part of the (peer) assessor to do the assessing.

The peer assessment should be worthwhile for each 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 less quality than his own paper (see Chapter 6). The opposite also occurred: a student who had little knowledge about the meaning of the criteria of a reflection paper, was less capable in giving constructive feedback. An interesting question would be to what extent domain expertise influences the development of assessment skills and in how far writing reliable assessments requires knowledge about the criteria of a product.

The design of the peer assessment tasks

Acquisition of peer assessment skills and content related skills were two goals in the studies. It became apparent that the training had to be much more systematic and of longer duration than we were able to organise in the available context and time span. A relevant question for future research is how the design of courses and the design of assessment training is most conducive to skill acquisition. This dissertation focused only on short-term effects of the training in peer assessment. It is conceivable that peer assessment training and more critical reflection about assessment might have a long-term effect for students.

There is considerable controversy in the literature on the optimal training strategy (Lane, 1987; Lee & Magill, 1985; Van Merriënboer, 1997). Researchers disagree on whether it is better to train in isolation or in context, whether to practice the whole task or its components, whether uninterrupted consistent repetitions of the same task or separated variations of a task are more beneficial, and whether a more structured guidance approach to learning or an unstructured discovery approach is better. A mixture of training and practice strategies and schedules is probably required, whereby an integration of training higher order skills and content related skills is preferred (Lin, 2001).

Meanwhile, a research project had been started at the Open University, that focuses on the design of self and peer activities that enhance student learning (Kirschner & Prins, 2002). The design of assessment support for students is an interesting aspect in this research project. Two examples of assessment support that will be elaborated are worked-out examples and process worksheets.

Peer assessment and collaborative learning

The issue of collaborative learning that was brought to notice several times, is also an important factor in the integrated framework. Collaborative learning is inherently more complex than competitive or individualistic learning because students have to engage simultaneously in teamwork. Social skills for effective cooperative work do not magically appear when tasks are employed in which students are expected to collaborate. Instead, social skills must be taught to students just as purposefully and precisely as academic skills. Since collaboration and conflict are inherently related (see Johnson & Johnson, 1995), the procedures and skills for managing conflicts constructively are especially important for the long-term success of learning groups. It is important to focus more on social skills in future research, since they may have predictive value for the development of peer assessment skills.

Assessing the assessment skill

An important instrument in the studies was the *rating form* that was used for the analysis of the written assessments reports and was derived from the peer assessment model. Certain decisions were made regarding the variables included in the rating form in the study reported in Chapter 4. It was experienced that the development of a reliable and valid instrument to measure students' assessment skills is quite complex. A limitation of the rating form is that it only measures the use of the appropriate criteria, and the extent to which students made positive, negative or constructive comments. This however does not necessarily mean that the students apply the criteria adequately and correctly. In-depth analyses of students' written assessment reports by content experts are recommended with regard to the limitations of the rating form. It needs further analysis and research to develop a reliable instrument for analysing assessment skills. The rating form as it is, has to be elaborated and tested. To have more insight if students actually develop skills like for example 'describe a personal report on course objectives' (see peer assessment model in Chapter 1), these skills have to be evaluated separately, in other words, assessment has to be more analytic and process oriented, instead of holistic and product oriented, as was done in the studies. Evaluation after each peer assessment task for example will give more information about which skills are easier to train than other skills and provide points for improvement for new training programs.

Role of teachers in teacher training colleges

In this dissertation, every study was predominantly focused on training students in assessment skills. The teachers though, who instructed the peer assessment tasks, were also confronted with the complexity of designing skill-based courses, performance assessments and peer assessment. They also had to collaborate and were to a large extent interdependent, because courses had to be redesigned from a skill-based perspective instead of a content-based perspective. In some way, teachers were forced to be critical about one another and about study tasks, in order to reach the desired goals of the project. As yet, it appeared to be very important to invest time in training teachers in new instructional design approaches, alternative assessment and student involvement in assessment. From the 'practice as you preach' – philosophy, teachers have to be receptive for self-reflection and change. The teachers in Teacher Training Colleges also have to become reflective practitioners (Schön, 1987). This however is an ongoing issue in the area of educational reform. The implementation process of certain innovations does not pass off without any resistance. The change in role-definition is sometimes hard to accept for teachers who have many years of teaching experience. Other teachers appeared to be more positive about their role in self and peer assessment activities and performance assessment.

Students' perceptions

In the last three studies, much effort was put in the investigation of students' perceptions. Students' perceptions were measured by means of questionnaires. These instruments can be regarded as self assessments, in which the students reflect on their opinion about instruction and assessments, but also on their role in collaborative learning groups and on their level of assessment skills. The results showed that students may overestimate themselves in a questionnaire, which was for example manifested in a high score on the peer assessment skills-items in the questionnaire and a low score on their written assessment reports.

A problem in the design of the presented study is that the pre-test-post-test comparison results of the student questionnaire may be an inaccurate measurement of the impact of the redesigned course. Because a pre-test was used at the beginning of the redesigned course, students had no way to correct an answer at the end of the course if they made an inaccurate evaluation in the baseline opinion (Sprangers & Hoogstraaten, 1989). By the end of the course, their new understanding of peer assessment and the skills involved may have an impact on the responses on their self-evaluation.

A 'post-then-pre-test design' would be a solution for this problem (Howard & Dailey, 1979). The problem is handled by not giving a pre-test at the beginning of the redesigned course. At the end of the course, students are asked two questions. The first question asks about behaviour as a result of the course. This is the post-test question. Then the student is asked to report what the behaviour had been before the course. This second question is really the pre-test question, but it is only asked after the course when the student has sufficient knowledge to answer the question validly. The retrospective pre-test at the end of the course would be more accurate because it is answered in the same frame of reference as the post-test. The problem of what has been called the 'response-shift bias' will be minimised (Sprangers & Hoogstraaten, 1989).

Practical implications

Research on assessment in teacher education and higher education in general is complex by nature. The success of sound assessment practices lies on the one hand in a close relationship between learning, instruction, and assessment; on the other hand in qualified assessors. This dissertation attempted to make a contribution on both aspects. The design of the peer assessments and performance assessments, the redesign of the courses, and the experiences of teachers and students, yielded many issues regarding the practical implications. These issues can be looked upon as guidelines or questions one could consider before implementation of peer assessment. In conclusion, the most important guidelines for practice and their justifications are outlined in Table 1.

Table 1. Guidelines and justification for implementing (peer) assessment

Guideline	Justification
Students have to be trained in (peer) assessment skills	Students learn to understand their role in the assessment process. Involvement in learning, including assessment, is vital to effective learning, and the teacher is the key person to help students to develop this learner autonomy (Dickinson, 1998).
Determine performance criteria beforehand	The importance of criteria is a factor which should not be underestimated. Criteria can be defined as guidelines, rules, characteristics, or dimensions that are used to judge the quality of student performance. Criteria are the basis of evaluating student progress; they identify the critical aspects of a performance or a product that describes in specific terms what is involved in meeting the learning outcomes (Boud, 1995).
Think backwards starting from the desired performance	The definition of a course objective in terms of a performance simplifies the discussion which content and knowledge is important. In current practices, it is still common to define first the course content before the assessment task, which can lead to an imbalance between instruction and assessment.
Stimulate collaborative learning	In the design of both peer assessment tasks and study tasks it is important to integrate collaborative activities, for example reflective interactions in the definition of performance criteria. Collaborative learning activities provide a platform for exchanging ideas and perceptions. Students are stimulated to share knowledge about aspects of instruction and assessment. The teacher takes the role of a coach, who mediates the discussions between students, without giving the right answers.
Create interdependency	In peer assessment students have a responsibility towards their peers. If a student for example does not write an assessment or does not make a constructive contribution in the group, this will be noticed and criticised by the other peers. In this example, the free-rider effect is locked, because each student has an individual responsibility.
Start training students in their first year of teacher training	Observation during informal evaluation of the peer assessment procedure with students showed that students are conservative in their attitude towards teachers and assessment. They still feel that the teacher is the expert and the only objective assessor. When students are trained to rely on their own judgment and that of their peers, they may develop a belief that a teacher is first of all a coach, who supports and adjusts the decisions that students make.

table 1 continued

Prevent illegitimate use of peer assessment	Peer assessments have to be based on valid reasons, for example as a learning tool. Peer assessment is often implemented as a tool that decreases the load of teachers. Yet, it is more desirable to see this application of peer assessment as a welcome side effect, besides the application as a learning tool. In this way, it works both ways.
Provide training for teacher educators	A structural change in course design and assessment implies a serious effort of teachers. Teachers often are not educated in instructional design or assessment. In-service training in topics such as performance assessment, peer assessment, and instructional design increases the successful implementation of new assessment approaches.
Create favourable conditions for implementation	The implementation of new assessment approaches requires both a top-down and bottom-up approach. Top-down in creating the conditions, such as time and training, bottom-up to create 'good examples' that can be deployed by colleagues.
Aim at using peer assessment as a tool for summative assessment	Peer assessment is most of all considered as a learning tool that supports students in their assessment skills. After adequate training, it is possible to use peer assessment for summative purposes on the understanding that students are capable to assess a peer. Teachers have to assess if students acquired these skills.

Assessment drives the learning process and overrides practically every other aspect of curriculum design (Longhurst & Norton, 1997). Changing assessment practices towards more performance based approaches, will inevitably lead to a revision of instruction. Instruction, assessment, and learning and teaching strategies have to be completely aligned. Educators must develop appropriate assessments that have no single right answer and in which students' argumentation is key in defending their solution. In this dissertation, it was chosen to design performance assessments in which students were actively involved through peer assessment activities. Performance assessments suit the new visions on instruction and assessment well. Performance assessments are increasingly based on ill-structured problems, on which no single correct answer can be given. With regard to the complex nature of the meaning of competence came a sense of the insufficiency of the traditional multiple choice test as a measure for complex skills. Teachers therefore have to realise that some tests they apply in their current practice are one-time measures, that rely on a single correct response to each item. Because they usually require brief responses, which are often machine-scored, students construct their responses in only the most minimal way and often by only plugging in a piece of knowledge. This is no problem, if this is the goal of the test. The split between instruction and assessment arises when the test does not measure what it actually purports to measure (*validity*), or if the assessment scores are not consistent and precise (*reliability*).

By involving students in the design of instruction and assessment, they become aware of how and on what knowledge and skills they are assessed. Peer assessment can be conceived as an evaluative device, but in the approach chosen in this dissertation it is moreover a learning activity. The student is introduced as an important collaborator with the teacher in the creation of tasks as well as in developing guidelines for scoring and interpretation. Until today, many tests are kept under lock and key so students do not have knowledge about them ahead of time. By doing this, students will study in a particular way in the hope that this will improve their test performance, but there is virtually no way that students can 'learn by doing' in the way that they learn while engaging in a performance based assessment in which they were involved as one of the assessors (Frederiksen, 1984).

An indirect goal of this dissertation was to encourage teachers to think about the performance assessment at the beginning of a course design process. The involvement of students in these processes implies an extra investment. Although the studies in this dissertation focused mainly on the training of student teachers, it became increasingly apparent that much effort has to be put into the professional development of teacher educators. Meanwhile, initiatives are conducted to define a vocational profile for teacher educators (Koster & Korthagen, 2001). The competencies of teacher educators are operationalised (Plake, Impara, Fager, 1993). Designing rich, authentic performance assessment is one of these competencies that deserves special attention. After all, assessment is the tail that wags the dog.

References

Arter, J. (1996). Using assessment as a tool for learning. In R. Blum, & J. Arter (Eds.) *Student performance assessment in an era of restructuring* (pp. 1–6). Alexandria, VA: Association for Supervision and Curriculum Development.

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

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

Conway, R., Kember, D., Sivan, A., & Wu, M. (1993). Peer assessment of an individual's contribution to a group project. *Assessment and Evaluation in Higher Education, 18,* 45–56.

Dickinson, L. (1988). Collaborative assessment: an interim account. In H. Holec (Ed.), *Autonomy and self-directed learning: present fields of application* (pp. 121–128). Strasbourg, France: Council of Europe.

Falchikov, N. (1986). Product comparisons and process benefits of collaborative peer group and self assessments. *Assessment and Evaluation in Higher Education, 11*, 146–166.

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.

Frederiksen, N. (1984). The real test bias: influences of testing on teaching and learning. *American Psychologist, 3*, 193-202.

Fry, S.A. (1990). Implementation and evaluation of peer marking in higher education. *Assessment and evaluation in higher education, 15,* 177–189.

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

Howard, G.S., & Dailey, P.R. (1979). Response-shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology, 66*, 144-150.

Johnson, D. W., & Johnson, R. T. (1995). *Teaching students to be peacemakers* (3rd ed.). Edina, MN: Interaction Book Company.

Jonassen, D.H., & Kwon, H.I. (2001). Communication patterns in computer mediated versus face-to-face group problem solving. *Educational Technology Research and Development, 49*, 35-51.

Kirschner, P., & Prins, F. (2002). *Electronic peer assessment during learning by design*. Heerlen: Open University of the Netherlands (Center for Educational Technology and Expertise).

Koster, B., & Korthagen, F. (2001). Training teacher educators for the realistic approach. In F. Korthagen (Ed), *Linking practice and theory: the pedagogy of realistic teacher education*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kwan, K., & Leung, R. (1996). Tutor versus peer group assessment of student performance in a simulation training exercise. *Assessment and Evaluation in Higher Education, 21*, 205–214.

Lane, N.E. (1987). *Skill acquisition rates and patterns: Issues and training implications.* New York: Springer-Verlag.

Lee, T.D., & Magill, R.A. (1985). Can forgetting facilitate skill aquisition. In D.Goodman, R.B. Wilberg, & I.M. Franks (Eds), *Differing perspectives in motor learning, memory, and control* (pp. 3-21). North-Holland: Elsevier Science Publishers B.V.

Lin, X. (2001). Designing metacognitive activities. *Educational Technology Research & Development, 49*, 2, 23-40.

Longhurst, N., & Norton, L.S. (1997). Self-assessment in coursework essays. *Studies in Educational Evaluation, 23*, 319–330.

Marttunen, M. (1992). Commenting on written arguments as a part of argumentation skills – a comparison between students engaged in traditional versus on-line study. *Scandinavian Journal of Educational Research, 36*, 289-302.

Mehrens, W.A., Popham, W.J., & Ryan, J.M. (1998). How to prepare students for performance assessments. *Educational Measurement: Issues and Practice, 17*, 1, 18-22.

Perkins, D. N. (1986). *Knowledge as design*. Hillsdale, NJ: Erlbaum.

Pilot, A. (1997, December). *Curricula voor een op informatie gerichte maatschappij*. [Curricula in information society]. Paper presented at the Consortium COO HBO and SUNCOO Conference, Utrecht, The Netherlands.

Pâquet, M.R., & Des Marchais, J.E. (1998). Students' acceptance of peer assessment. *Education for Health, 11*, 25-36.

Plake, B.S., Impara, J.C., & Fager, J.J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement, Issues and Practice, 12*, 10-12.

Pond, K., Ul-Haq, R., & Wade, W. (1995). Peer review: a precursor to peer assessment. *Innovations in Education and Training International, 32*, 314–323.

Rushton, C., Ramsey, P., & Rada, R. (1993). Peer assessment in a collaborative hypermedia environment: a case study. *Journal of Computer-Based Instruction, 20*, 75–80.

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

Sprangers M., & Hoogstraaten, J. (1989). Pretesting effects in retrospective pretest-posttest designs. *Journal of Applied Psychology, 74*, 265-272.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*, 3, 249-276.

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* (pp. 19-32). Leuven-Apeldoorn: Garant.

William, J. B. (1999). On the free-rider problem in cooperative learning. *Journal of Education for Business, 74*, 271-74.

Summary

Teacher education is a fascinating domain for educational research. The teacher training colleges face an enormous job in changing their educational practices, because of its complex double task to educate student teachers who have to educate pupils in elementary schools in the future. Two developments that have been brought to notice in the last few years are the involvement of students in assessment and the design of more competency-based curricula. Both issues are strongly interrelated, since it is assumed that instruction, learning and assessment have to be completely aligned for optimal learning results (Biggs, 1996, 1999, 2001). Within the theme diagnosis of the research program 'Instructional Design for Competency-based Learning in Post-secondary Higher Education' (Van Merriënboer, 1998), the research project 'Student involvement in assessment' was started that focused on self and peer assessment as learning tools in higher education.

Based on a theoretical framework on constructive alignment, instructional design and student involvement, the main aim of this dissertation was to investigate the issue of student involvement in teacher education, more specifically the effect of an embedded training in peer assessment skills on students' performance in their peer assessment skills and content based skills. The thesis contains two parts. The first part consists of a literature study and a pilot study. The second part consists of three experimental studies, which are conducted within the context of teacher education. The following five research questions were addressed in the first part:

1. How are self-, peer- and co-assessment applied in higher education?
2. What are the effects of the use of these forms of assessment on the quality of the learning environment?
3. Are peer ratings in problem based learning groups reliable?
4. Do students have idiosyncratic (i.e., personal) strategies in peer assessment?
5. What are students' experiences with peer assessment and problem based learning?

A literature review was conducted to investigate the first two research questions. Results on *research question 1* were diverse. Instruments for self assessment vary from Likert- scales, ability listings and written tests to portfolios, audiotape assessments or electronic interactive systems (e.g., Anderson & Freiberg, 1995; Gentle, 1994; Longhurst & Norton, 1997). Peer assessment practices are mainly focused on peer ranking (each group member rank all of the others from best to worst 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). Self assessment and peer assessment are mostly used for formative purposes, but summative assessment is possible when these forms are combined with co-assessment (e.g., Hall, 1995). The answers on *research question 2* revealed that self-, peer and co-assessment to some extent do improve different aspects of the quality of learning of students. It increased students' confidence in the ability to perform (e.g., Cutler & Price, 1995), as well as their awareness of the quality of their own work (e.g., Anderson & Freiberg, 1995; Gentle, 1994). Several studies found that students reflected more on their own behaviour and/or performance (Anderson & Freiberg, 1995; Gentle, 1994; Longhurst & Norton, 1997; Sobral, 1997). Other studies reported increased quality of learning (e.g., Cutler & Price, 1995; Freeman, 1995; Hassmèn, 1997; Horgan, et al., 1997), and independency and responsibility of students (e.g., Anderson & Freiberg, 1995). At last, it was concluded that students have both positive and negative experiences with self and peer assessment (Cheng & Warren, 1997; Conway et al., 1993; Cutler & Price, 1995; Peters, 1996; Warkentin, et al., 1995; Williams, 1992). Only a few studies examined directly the effects of self and peer assessment on end-of-course individual exams (e.g., Warkentin, et al., 1995). In both studies, no effects were found.

Chapter 2 reported a pilot study in which *research questions 3,4 and 5* are examined. The setting was problem based, which means that students worked collaboratively in groups on tasks. At the end of the course period, students assessed one another on a list of four criteria concerning their participation in the group. Two studies within the teacher training context were conducted; one within a university setting with two problem based working groups, another within a higher vocational education context with four problem based working groups. Findings of the *research question 3* yielded a two-fold answer: the peer assessment leads in one group to acceptable reliability results and in the other group to unacceptable results. This was also the case in the second study, where only the results in the third and fourth group led to acceptable results. This finding is less positive than results from other studies, where student-tutor marks are more highly correlated (Freeman, 1995; Longhurst & Norton, 1997), but in concordance with the studies of Kwan and Leung (1996) and Rushton, Ramsey and Rada (1993). It appeared that in the first study just one of the students was using an idiosyncratic rating strategy and only one outlier was found in the second study *(research question 4)*. The results from the evaluation questionnaire that examined *research question 5* show that the students in general felt uncomfortable in assessing their peers. Students in the second study found it unacceptable to give negative scores without having the opportunity to give informative feedback. All students were positive about implementing peer assessment, but the students in the first study predicted many more implications for the institution. Half of all students see possibilities for peer assessment in other courses with problem based learning.

Regarding the literature review and the results of the pilot study, the concepts of constructive alignment, student involvement, and course design were integrated in a framework that underlies particularly the design of the studies reported in Chapters 4, 5, and 6. In this framework, peer assessment is interpreted as a complex skill in which students have to be trained. A peer assessment model was designed (Sluijsmans & Van Merriënboer, 1997), and a training in assessment skills was developed.

The integrated framework contains two important 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, Popham, & Ryan, 1998). In this view, the assessment skill is not trained as an isolated skill, but is directly linked to course content. On that line of argument, student teachers will always be guided in at least two skills: the skill to assess work of peers and a content-related skill, which contains the object 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). Second, training students in skills in general 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, Kessels, & Meijers, 2000). This means that skills are trained in the context of different content domains.

Five research questions were explored in depth in the second part of this dissertation, in which three studies are reported (Chapter 4, 5, and 6):

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

In the first two studies 93 second-year student teachers participated, in the third study 110 first-year student teachers. Positive findings were found regarding the effects of a training in peer assessment by means of peer assessment tasks, on the students' skill to assess the work of peers *(research question 1)*. The results found vary between each study. In the study reported in Chapter 4, it was found 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 effect though was very small, and effects on other variables of the rating form failed to occur, probably due to the training that only focused on training in the skill 'defining criteria'. In the subsequent study, reported in Chapter 5, the analyses of the qualitative peer assessment reports revealed that the experimental groups were more likely to use the criteria and to give more constructive comments than the student teachers from the control groups. The students who received training also scored higher on structure and used less naive words. In spite of the positive results reported in Chapters 4 and 5, it was concluded that student teachers could not be regarded as expert assessors after a peer assessment training in one course. A longer training period appeared to be recommendable. Therefore, a study was conducted according to a longitudinal within-subjects design in order to investigate the development of assessment skills as an effect of assessment training over a period of almost one year instead of in only one course period of six weeks (Chapter 6). The training in this longitudinal study was integrated in three successive mathematics courses. Analysis of the peer assessments from the intake and outtake data revealed significant effects. This progress was significant for most variables. All students used the criteria more adequately, gave more constructive feedback, and wrote more structured assessment reports after the training period of ten months. Students also adopted a more critical attitude in the outtake than in the intake.

The next research question focused on the effect of training peer assessment skills on students' content-related performance (*research question 2*). At first, a difference between the performance quality of the students from the control and from the experimental group failed to occur (Chapter 4). Explanations were sought in the small progress in the peer assessment skill and the short training period. Moreover, it is possible that the redesign of the course had an effect on the learning result of all students. In the next study, in which the whole peer assessment skill was trained, a positive effect of the peer assessment training on the actual learning results was found (Chapter 5). The student teachers from the experimental groups outperformed the students from the control groups. This same result was found in the last study reported in Chapter 6, where the whole group of students wrote better reflection reports after the training than before the training.

The intervention of a peer assessment training had several implications for students. Not only did they have to work on peer assessment tasks, they also were confronted with a course designed in an alternative way, in which collaborative activities that required an active contribution played an important role. Besides these changes, they were assessed on their peer assessment skills and content-related performance. To identify students' perceptions towards these developments, a questionnaire was developed that was taken before and after each course in which peer assessment training was embedded (*research question 3*). The variables included were related to instruction, to vision on instruction and assessment, and to the role of the student in assessment. The questionnaire applied in the study reported in Chapter 4 was slightly revised in the subsequent two studies, but the variables stayed for the most part the same.

Overall results of each study were that 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, which was designed from a skill-based perspective and consisted of tasks that fostered collaborative learning and interaction, led to an active participation of student teachers and the teachers. It can be concluded that the student teachers changed positively in their view on aspects of learning and assessment. They are more satisfied about the classes and the criteria and goals are clearer. The role of the teacher is also evaluated in a more positive way. The student teachers indicated that they are more capable in assessing than before the redesign of the course.

A result that was found in the study reported in Chapter 4 was that the content domain influenced students' perceptions (*research question 4*). This finding was mainly caused by the domain mathematics. Explanations for this effect could be sought in the specific character of the mathematics domain.

To embed a training in assessment skills in existing courses, it was necessary to reflect on these existing courses. The redesign process and the implementation phase demanded a lot of effort of the teachers that were involved in the courses. Teacher experiences were investigated in the study reported in Chapter 4 and in the longitudinal study (Chapter 6) (*research question 5*). In both studies, it appeared that the teachers had no major problems in instructing the peer assessment tasks. They indicated that implementing the peer assessment training led to a rethinking of the existing course and stimulated them to view the content from a different perspective. The need for revision of courses that lasted for several years though did also lead to some resistance. Some teachers doubted the value of the peer assessment and were sometimes reluctant to give up some part of their content expertise on behalf of the 'higher order' skills.

Implications and future research

Acquisition of peer assessment skills and content related skills were two goals in this dissertation. Although positive effects were found, it became apparent that the training had to be much more systematic and of longer duration than was feasible to organise in the available context and time span. This dissertation focused only on short-term effects of the training in peer assessment. It is conceivable that peer assessment training and more critical reflection about assessment might have a long-term effect for students. Limitations of the materials were found in the use of the student questionnaire that measured students' perceptions and the rating form that measured the quality of students' assessment skills. A problem with the student questionnaire was that the pre-test-post-test comparison results of the student questionnaire may be an inaccurate measurement of the impact of the redesigned course. Because a pre-test was used at the beginning of the redesigned course, students had no way to correct an answer at the end of the course if they made an inaccurate evaluation in the baseline opinion (Sprangers & Hoogstraaten, 1989). By the end of the course, their new understanding of peer assessment and the skills involved may have an impact on the responses on their self-evaluation. A restriction of the rating form is that it only measures the use of the appropriate criteria, and the extent to which students made positive, negative or constructive comments. This however does not necessarily mean that the students apply the criteria adequately and correctly. In-depth analyses of students' written assessment reports by content experts are recommended with regard to the limitations of the rating form. It needs further analysis and research to develop a reliable assessment instrument for analysing assessment skills.

A relevant question for future research is how the design of courses and the design of assessment training is most conducive to skill acquisition. A reconsideration of the peer assessment model and the collaborative activities appeared to be desirable. It is also interesting to elaborate further on the relationship between peer assessment skill acquisition and content skill acquisition, and to what extent domain expertise influences the development of assessment skills.

Changing assessment practices and views on learning and the role of students in this, is a considerable challenge in teacher education and higher education in general. The success of sound assessment practices lies on the one hand in a close relationship between learning, instruction, and assessment, on the other hand in qualified (student) assessors. This dissertation attempted to make a contribution on both aspects. Important guidelines for practice are that students need to be guided in their skill-development, that a clear definition of performance criteria is crucial for effective assessments, that collaborative activities need to be stimulated, and that teacher educators receive training in instructional design and alternative assessment approaches. From the 'practice as you preach' - philosophy, an important condition for successful initiatives on the student level is that teachers are receptive for self-reflection and change.

References

Anderson, J.B., & Freiberg, H.J. (1995). Using self assessment as a reflective tool to enhance the student teaching experience. *Teacher Education Quarterly, 22*, 77–91.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*, 347-364.

Biggs, J.B. (1999). *What the student does: Teaching for quality learning at university*. Buckingham: Open University Press.

Biggs, J.B. (2001). The reflective institution: Assuring and enhancing the quality of teaching and learning. *Higher Education, 14*, 221-238.

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

Conway, R., Kember, D., Sivan, A., & Wu, M. (1993). Peer assessment of an individual's contribution to a group project. *Assessment and Evaluation in Higher Education, 18,* 45–56.

Cutler, H., & Price, J. (1995). The development of skills through peer assessment. In A. Edwards, & P. Knight (Eds.), *Assessing competence in higher education* (pp. 150–159). Birmingham, UK: Staff and Educational Development Series.

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

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

Gentle, C.R. (1994). Thesys: an expert system for assessing undergraduate projects. In M. Thomas, T. Sechrest, & N. Estes (Eds.), *Deciding our future: technological imperatives for education* (pp. 1158–1160). Austin, TX: The University of Texas.

Hall, K. (1995, February). *Co-assessment: participation of students with staff in the assessment process: a report of work in progress*. Invited paper presented at the Second European Electronic Conference on Assessment and Evaluation. EARLI-AE list European Academic & Research Network (EARN) (EARLI_AE on Listserv@nic.surfnet.nl).

Hassmén, P., Sams, M.R., & Hunt, D.P. (1996). Self assessment responding and testing methods: effects on performers and observers. *Perceptual and Motor Skills, 83,* 1091–1104.

Horgan, D.D., Bol, L., & Hacker, D. (1997, August). *An examination of the relationships among self, peer, and instructor assessments.* Paper presented at the meeting of the European Association for Research on Learning and Instruction, Athens, Greece, 1997.

Kane, J.S., & Lawler III, E.E. (1978). Methods of peer assessment. *Psychological Bulletin, 85*, 555–586.

Longhurst, N., & Norton, L.S. (1997). Self-assessment in coursework essays. *Studies in Educational Evaluation, 23*, 319–330.

Mehrens, W.A., Popham, W.J., & Ryan, J.M. (1998). How to prepare students for performance assessments. *Educational Measurement: Issues and Practice, 17*, 1, 18-22.

Peters, M. (1996). Student attitudes to alternative forms of assessment and to openness. *Open Learning*, 11(3), 48–50.

Rushton, C., Ramsey, P., & Rada, R. (1993). Peer assessment in a collaborative hypermedia environment: a case study. *Journal of Computer-Based Instruction, 20*, 75–80.

Sluijsmans, D., & Van Merriënboer, J.J.G. (2000). *A peer assessment model.* Heerlen: Open University of the Netherlands. (Center for Educational Technology and Expertise).

Sobral, D.T. (1997). Improving learning skills: a self-help group approach. *Higher Education*, *33,* 39–50.

Sprangers M., & Hoogstraaten, J. (1989). Pretesting effects in retrospective pretest-posttest designs. *Journal of Applied Psychology, 74*, 265-272.

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. (1998). *Instructional design for competency-based learning in post-secondary higher education*. Heerlen: Open University of the Netherlands. (Center for Educational Technology and Expertise).

Warkentin, R.W., Griffin, M.M., Quinn, G.P., & Griffin, B.W. (1995, April). *An exploration of the effects of cooperative assessment on student knowledge structure*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Williams, E. (1992). Student attitudes towards approaches to learning and assessment. *Assessment and Evaluation in Higher Education, 17*, 45–58.

Dutch summary

Het domein van de lerarenopleidingen is een fascinerende context voor onderwijskundig onderzoek. Lerarenopleidingen hebben te maken met een complexe taak met betrekking tot het veranderen van hun onderwijskundige praktijk. Twee ontwikkelingen die de laatste jaren regelmatig in de belangstelling staan zijn de toenemende betrokkenheid van studenten bij hun eigen toetsing en het ontwerp van meer competentiegericht onderwijs. Beide ontwikkelingen zijn nauw verbonden met elkaar, aangezien instructie, leren en toetsing volledig op elkaar moeten zijn afgestemd, om optimale leerresultaten te bereiken (Biggs, 1996, 1999, 2000). Binnen het thema 'Diagnosis' van het onderzoeksprogramma 'Instructional Design for Competency-based Learning in Postsecondary Higher Education' (Van Merriënboer, 1998) is het promotie project 'Student-betrokkenheid in assessment' gesitueerd dat gericht is op self en peer assessment als leermiddelen in het hoger onderwijs.

Gebaseerd op een theoretisch kader rondom 'constructive alignment', instructie-ontwerp en studentbetrokkenheid, is in dit proefschrift het thema van studentbetrokkenheid in toetsing binnen de context van de lerarenopleiding voor primair onderwijs aan een nadere analyse onderworpen. Meer specifiek zijn de effecten van een training in beoordelingsvaardigheden op de ontwikkeling van de beoordelingsvaardigheid van studenten en hun vaardigheden binnen een bepaald inhoudsdomein onderzocht. Het proefschrift bestaat uit twee delen. Het eerste deel omvat een literatuurstudie en een pilot studie. Het tweede deel rapporteert drie experimentele studies die zijn uitgevoerd binnen de context van een lerarenopleiding. De volgende vijf vragen stonden centraal in het eerste deel:

1. Hoe wordt studentbetrokkenheid in toetsing in hoger onderwijs bewerkstelligd door vormen als self-, peer- en and co-assessment?
2. Wat zijn de effecten van de invoering van self- peer- en co-assessment op de kwaliteit van de leeromgeving?
3. Zijn peer beoordelingen in probleem gestuurd onderwijs betrouwbaar?
4. Hebben studenten persoonlijke strategieën in peer assessment?
5. Wat zijn de ervaringen van studenten met betrekking tot peer assessment in probleem gestuurd onderwijs?

Om een antwoord te vinden op de eerste twee onderzoeksvragen is een literatuurstudie uitgevoerd. De resultaten van de eerste onderzoeksvraag waren uiteenlopend. Instrumenten voor self assessment varieerden van Likert-schalen, vaardighedenlijsten, geschreven testen tot portfolio's, audiotape assessments of elektronische interactieve systemen (Anderson & Freiberg, 1995; Gentle, 1994; Longhurst & Norton, 1997). Peer assessment toepassingen zijn voornamelijk gericht op peer ranking (ieder groepslid plaatst alle andere groepsleden in een lijst van 'beste' tot 'slechtste' op één of meer criteria), peer nomination (ieder groepslid geeft aan welke student hij het beste beoordeelt op een bepaald criterium) of peer rating (ieder groepslid scoort elk ander groepslid op een set van karakteristieken op basis van één of meerdere beoordelingschalen) (Kane & Lawler, 1978). Self assessment en peer assessment worden overwegend gebruikt voor formatieve doeleinden. Summatief gebruik is mogelijk wanneer self en/of peer assessment wordt toegepast in combinatie met een vorm van co-assessment (Hall, 1995). De bevindingen met betrekking tot de tweede onderzoeksvraag lieten zien dat self-, peer en co-assessment in bepaalde mate de kwaliteit van verschillende aspecten van het leerproces van studenten verhoogt. Het blijkt dat studenten meer vertrouwen kregen in hun eigen handelen (Cutler & Price, 1995) en dat ze meer inzicht kregen in de kwaliteit van hun eigen werk (Anderson & Freiberg, 1995; Gentle, 1994). Uit een aantal studies is gebleken dat studenten meer reflecteerden op hun eigen studiegedrag (Anderson & Freiberg, 1995; Gentle, 1994; Longhurst & Norton, 1997; Sobral, 1997). Andere studies rapporteren verbeterde leerprestaties (Cutler & Price, 1995; Freeman, 1995; Hassmèn, 1997) en een verhoogd onafhankelijkheids- en verantwoordelijkheidsgevoel bij studenten (Anderson & Freiberg, 1995). Tot slot is gebleken dat studenten zowel positieve als negatieve ervaringen hebben met self en peer assessment (Cheng & Warren, 1997; Conway et al., 1993; Cutler & Price, 1995; Peters, 1996; Warkentin, et al., 1995; Williams, 1992). Slechts enkele studies hebben de directe relatie tussen self / peer assessment en prestaties van summatieve toetsen onderzocht (e.g., Warkentin, et al., 1995). In beide studies bleven effecten uit.

In hoofdstuk twee wordt verslag gedaan van een pilot studie waarin de derde, vierde en vijfde onderzoeksvraag centraal stonden. Deze studie vond plaats binnen het probleemgestuurde onderwijs. In dit type onderwijs werken studenten binnen een groep gedurende een aantal weken samen aan probleemtaken volgens een bepaalde stappenmethode. Aan het einde van de blokperiode hebben de studenten elkaar kwantitatief beoordeeld aan de hand van een lijst met vier criteria die gericht waren op de participatie in de groep. Twee deelstudies zijn uitgevoerd: één binnen de setting van een universitaire lerarenopleiding met twee groepen studenten, de ander binnen een PABO met vier groepen studenten. *Onderzoeksvraag 3* leverde in de eerste deelstudie een tweeledig antwoord op: de peer assessment leverde bij de ene groep acceptabele overeenstemming op tussen peer en tutor scores, in de andere groep niet. Dit was ook het geval in de tweede deelstudie, waar slechts bij twee van de vier groepen acceptabele scores werden gevonden. Deze resultaten zijn minder positief dan andere studies waarbij student-tutor scores hoger zijn gecorreleerd (Freeman, 1995; Longhurst & Norton, 1997), maar in overeenstemming met de studies van Kwan en Leung (1996) en Rushton, Ramsey and Rada (1993). Het bleek dat in de eerste deelstudie slechts één student een idiosyncratische strategie hanteerde en dat slechts één uitbijter te vinden was in de tweede deelstudie (*onderzoeksvraag 4*). De uitkomsten van de evaluatie-vragenlijst ten behoeve van *onderzoeksvraag 5* lieten zien dat de studenten zich in het algemeen ongemakkelijk voelen bij het geven van een beoordeling van hun medestudenten. Studenten in de tweede deelstudie gaven aan het onacceptabel te vinden om negatieve scores te geven aan hun medestudenten, zonder daarbij de mogelijkheid te hebben om feedback te geven. Alle studenten stonden positief tegenover de invoering van peer assessment, maar de studenten uit de eerste deelstudie voorspelden meer implicaties voor de onderwijsinstelling. De helft van alle studenten zagen mogelijkheden voor peer assessment in andere blokken binnen de context van het probleemgestuurd onderwijs.

Op basis van de bevindingen van de literatuurstudie en de pilot studie zijn de concepten van ‘constructive alignment’, studentbetrokkenheid en instructie-ontwerp geïntegreerd in een kader dat als uitgangspunt is genomen voor de studies die in de hoofdstukken 4, 5 en 6 zijn gerapporteerd. In dit kader wordt peer assessment beschouwd als een complexe vaardigheid waarin studenten moeten worden getraind. Een peer assessment model is opgesteld (Sluijsmans & Van Merriënboer, 2000) en een training in assessmentvaardigheden is ontworpen.

Het geïntegreerde kader gaat uit van twee assumpties. De eerste assumptie is dat een training in assessment vaardigheden positieve effecten kan hebben op de prestaties met betrekking tot inhoudsgerelateerde vaardigheden, mits de training is ingebed in het modulemateriaal dat ontworpen is volgens een op vaardigheden gebaseerde benadering (Mehrens, Popham, & Ryan, 1998). In dit opzicht wordt de beoordelingsvaardigheid niet getraind als een geïsoleerde vaardigheid, maar direct gerelateerd aan de inhoud van de module. In navolging hierop worden studenten altijd in ten minste twee vaardigheden getraind: de beoordelingsvaardigheid en een inhoudsgebonden vaardigheid, waarop studenten summatief worden beoordeeld. Dit leidt tot de hypothese dat als studenten worden ondersteund in hun beoordelingsvaardigheden, dit ook zou moeten leiden tot een beter begrip van inhoudsgerelateerde criteria. Het wordt verondersteld dat het begrijpen van de criteria van een bepaald product en het observeren van het werk van medestudenten leidt tot een verbeterde kijk op de kwaliteit van het eigen werk (Falchikov, 1995; Freeman, 1995). De tweede assumptie is dat het trainen van studenten in hun beoordelingsvaardigheid consequenties heeft voor het ontwerp van het onderwijs. Competentiegericht onderwijs betekent dat de inhoud niet langer wordt gedefinieerd vanuit het perspectief van de vakinhouden, maar vanuit het perspectief van de te ontwikkelen vaardigheden (Tillema, Kessels, & Meijers, 2000). Dit impliceert dat vaardigheden worden getraind in de context van verschillende vakinhouden.

Vijf onderzoeksvragen zijn uitvoerig onderzocht in het tweede deel van dit proefschrift, waarin verslag wordt gedaan van drie studies (hoofdstuk 4, 5, en 6):

1. Leidt training in peer assessment vaardigheden tot een hoger beoordelingsvermogen?
2. Leidt training in peer assessment vaardigheden tot een verbetering van het leerresultaat met betrekking tot de module-inhoud?
3. Wat zijn de percepties van studenten met betrekking tot herontworpen modules?
4. Beïnvloedt vakinhoud deze percepties?
5. Wat zijn de percepties van docenten met betrekking tot herontworpen modules en de peer assessment training?

In de eerste twee studies participeerden 93 studenten, in de derde studie 110 studenten. Positieve resultaten werden gevonden met betrekking tot de effecten van de training in peer assessment op de vaardigheid van het beoordelen (*onderzoeksvraag 1*). De resultaten verschillen echter per studie. In de studie die wordt gerapporteerd in hoofdstuk 4 bleek dat de experimentele groepen beter in staat waren in het toepassen van criteria dan de controlegroepen. Het verschil tussen beide groepen was echter erg klein en effecten op de andere variabelen werden niet gevonden. Dit laatste waarschijnlijk omdat de training alleen was gericht op het leren definiëren en toepassen van beoordelingscriteria. In de daaropvolgende studie die wordt gerapporteerd in hoofdstuk 5, werd na analyse van de peer assessment rapporten gevonden dat de studenten van de experimentele groepen niet alleen meer criteria toepasten dan de controle groepen, maar dat zij ook meer constructieve feedback gaven. De studenten uit de experimentele conditie schreven ook gestructureerdere rapporten en gebruikten minder naïeve woorden. Ondanks de significante resultaten van de twee studies, is geconcludeerd dat studenten nog geen expert beoordelaars zijn na een training in één module. Een langere trainingsperiode blijkt nodig.

In een volgende longitudinale studie (hoofdstuk 6) zijn studenten getraind in hun beoordelingsvaardigheden gedurende een periode van bijna een jaar in plaats van in één module. De training is ingebed in drie opeenvolgende wiskundemodules. Bij de vergelijking van de data van de voormeting en nameting werden significante verschillen gevonden. Alle studenten gebruikten de vastgestelde criteria in hun assessment rapporten en schreven gestructureerdere assessment rapporten na een periode van tien maanden. Studenten durfden ook een kritische houding aan te nemen.

Met betrekking tot de *tweede onderzoeksvraag* is onderzocht of een training in beoordelingsvaardigheden leidt tot een verbetering van het modulegebonden leerresultaat. Dit bleek niet het geval te zijn in de eerste studie (hoofdstuk 4). Verklaringen hiervoor zijn de korte periode van training en het kleine verschil in beoordelingsvaardigheid tussen de experimentele en controlegroepen. Het is ook mogelijk dat de herontworpen module een effect heeft gehad op het leerresultaat van alle studenten. In de daaropvolgende studie, waarin de studenten zijn getraind in een aantal beoordelingsvaardigheden, werd wel een effect gevonden op het inhoudsgebonden leerresultaat (hoofdstuk 5). De studenten van de experimentele groepen scoorden hogere cijfers op het eindproduct dan de controlegroepen. Ditzelfde resultaat werd gevonden in de derde studie, waarbij de hele groep studenten na de modules betere reflectiepapers schreven dan bij de voormeting (hoofdstuk 6).

Het volgen van de peer assessment training hield voor de studenten een behoorlijke verandering in. Naast het volgen van een aantal peer assessment taken, werden zij ook geconfronteerd met een module die op een andere manier was georganiseerd dan zij gewend waren. In deze herziene module was een actieve rol voor de studenten weggelegd. Naast deze veranderingen werden zij beoordeeld op de kwaliteit van hun peer assessment rapportages en hun eindproducten die waren gekoppeld aan de module-inhoud. Om de percepties van studenten te achterhalen, ontwikkelden de onderzoekers een vragenlijst die werd afgenomen vooraf aan de herziene module en na het volgen van de module (*onderzoeksvraag 3*). De variabelen omvatten vragen rondom instructie, visie op instructie en assessment en visie op hun eigen rol in assessment.

De vragenlijst zoals die is gebruikt in de studie van hoofdstuk 4 is enigszins aangepast voor de daaropvolgende studies, maar dezelfde variabelen zijn gehandhaafd.

De algemene bevindingen lieten zien dat de studenten de instructie en de integratie van instructie en assessment van de herziene module positief waardeerden. De hernieuwde module, die ontworpen was vanuit een vaardighedenperspectief en taken omvatte die gericht waren op samenwerkend leren en interactie, leidde tot een actieve participatie van de studenten en de docenten. Er mag worden geconcludeerd dat de studenten positief zijn veranderd in hun visie op aspecten van leren en assessment. Ze zijn meer tevreden over de studietaken waarvan de doelen en criteria helder zijn. De rol van de docent is tevens positiever beoordeeld. De studenten gaven aan dat ze zich meer in staat voelen een beoordeling van een medestudent te geven dan voor de module en de training.

Een resultaat dat werd gevonden in de eerste studie (hoofdstuk 4) was dat vakinhoud van invloed is op de percepties van studenten (*onderzoeksvraag 4*). Deze bevinding werd voornamelijk veroorzaakt dor het wiskundedomein. Verklaringen voor dit verschijnsel worden gezocht in het specifieke karakter van dit domein.

Om een training in beoordelingsvaardigheden te kunnen integreren in bestaande modules, was het noodzakelijk te reflecteren op deze bestaande modules. Het proces van herontwerp en de invoeringsfase vereisten veel inspanning van de betrokken docenten. De ervaringen van docenten zijn onderzocht in de eerste en derde studie (hoofdstuk 4 en 6). In beide studies bleken de docenten geen problemen te ondervinden bij de instructie van de peer assessment taken. Ze gaven aan dat de invoering van de peer assessment training een herbezinning op hun eigen onderwijs impliceerde. De noodzaak tot herziening van modules die al jarenlang een plaats binnen het opleidingscurriculum hebben, leidde ook tot enige weerstand. Enkele docenten twijfelden aan de waarde van peer assessment en stonden soms aarzelend tegenover het feit dat het aanbieden van bepaalde vakinhoud ter discussie kwam te staan ten behoeve van de ontwikkeling van de 'hogere orde' vaardigheden.

Implicaties en toekomstig onderzoek

Het verwerven van beoordelingsvaardigheden en inhoudsgebonden vaardigheden als gevolg van een interventie waren twee doelen van dit proefschrift. Alhoewel positieve effecten zijn gevonden, is ook duidelijk gebleken dat een dergelijke training systematischer moet worden uitgebreid dan nu mogelijk was binnen de beschikbare randvoorwaarden. Dit proefschrift richtte zich alleen op korte termijn effecten van de training. Het is denkbaar dat peer assessment training en kritische reflectie lange termijn effecten hebben. De vragenlijst die studentpercepties onderzocht en het beoordelingsformulier voor de analyse van de peer assessments blijken een aantal beperkingen te hebben. Het probleem van de vragenlijst ligt in het feit dat de vergelijking van de voor- en nameting resultaten een inaccurate meting kan zijn voor de percepties op de herziene module. Aangezien een voormeting heeft plaatsgevonden voor het begin van de herziene module, hebben studenten geen gelegenheid hun mening bij te stellen aan het einde van de module (Sprangers en Hoogstraten, 1989). Hun nieuwe visie aan het eind van de module op peer assessment en de vaardigheden kan een impact hebben op de gegeven antwoorden. Een beperking van het beoordelingsformulier voor het meten van de kwaliteit van de peer assessments is dat het instrument alleen meet of de criteria worden gebruikt en de mate waarin studenten positieve, negatieve of constructieve commentaren geven. Er kunnen echter geen uitspraken worden gedaan of de studenten de criteria op de juiste manier gebruiken. Diepte-analyses van de geschreven assessment rapporten door inhoudsdeskundigen zijn noodzakelijk om meer inzicht te verkrijgen in bepaalde kwaliteitsaspecten. Verder onderzoek is nodig om tot een betrouwbaar en valide instrument te komen voor het meten van beoordelingsvaardigheden.

Een belangrijke vraag voor toekomstig onderzoek is hoe het ontwerp van modules en het ontwerp van een assessment training zodanig kan worden vormgegeven dat het daadwerkelijk tot het verwerven van vaardigheden leidt. Een herziening van het peer assessment model en de gerelateerde samenwerkingsactiviteiten is wenselijk. Het is tevens interessant om verder de relatie tussen beoordelingsvaardigheden en inhoudsgebonden vaardigheden te onderzoeken en in hoeverre vakinhouden deze relatie beïnvloeden.

Het veranderen van assessments binnen de opleiding en de rol van studenten hierin, is een uitdaging in lerarenopleidingen en het hoger onderwijs in het algemeen. Het succes van betrouwbare assessments ligt aan de ene kant in een nauwe relatie tussen leren, instructie en toetsing, aan de andere kant in getrainde assessoren. Dit proefschrift heeft gepoogd een bijdrage te leveren aan beide aspecten. Belangrijke richtlijnen voor de praktijk zijn dat studenten begeleid moeten worden in hun ontwikkeling in vaardigheden, dat performance criteria helder moeten worden geformuleerd, dat samenwerkingsactiviteiten worden ingebouwd en dat docenten in lerarenopleidingen training ontvangen in instructieontwerp en alternatieve assessment benaderingen. Vanuit de 'practice as you preach'- filosofie is een cruciale randvoorwaarde dat docenten ontvankelijk zijn voor reflectie en verandering.

References

Anderson, J.B., & Freiberg, H.J. (1995). Using self assessment as a reflective tool to enhance the student teaching experience. *Teacher Education Quarterly, 22*, 77–91.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*, 347-364.

Biggs, J.B. (1999). *What the student does: Teaching for quality learning at university*. Buckingham: Open University Press.

Biggs, J.B. (2001). The reflective institution: Assuring and enhancing the quality of teaching and learning. *Higher Education, 14*, 221-238.

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

Conway, R., Kember, D., Sivan, A., & Wu, M. (1993). Peer assessment of an individual's contribution to a group project. *Assessment and Evaluation in Higher Education, 18,* 45–56.

Cutler, H., & Price, J. (1995). The development of skills through peer assessment. In A. Edwards, & P. Knight (Eds.), *Assessing competence in higher education* (pp. 150–159). Birmingham, UK: Staff and Educational Development Series.

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

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

Gentle, C.R. (1994). Thesys: an expert system for assessing undergraduate projects. In M. Thomas, T. Sechrest, & N. Estes (Eds.), *Deciding our future: technological imperatives for education* (pp. 1158–1160). Austin, TX: The University of Texas.

Hall, K. (1995, February). *Co-assessment: participation of students with staff in the assessment process: a report of work in progress*. Invited paper presented at the Second European Electronic Conference on Assessment and Evaluation. EARLI-AE list European Academic & Research Network (EARN) (EARLI_AE on Listserv@nic.surfnet.nl).

Hassmén, P., Sams, M.R., & Hunt, D.P. (1996). Self assessment responding and testing methods: effects on performers and observers. *Perceptual and Motor Skills, 83,* 1091–1104.

Horgan, D.D., Bol, L., & Hacker, D. (1997, August). *An examination of the relationships among self, peer, and instructor assessments.* Paper presented at the meeting of the European Association for Research on Learning and Instruction, Athens, Greece, 1997.

Kane, J.S., & Lawler III, E.E. (1978). Methods of peer assessment. *Psychological Bulletin, 85*, 555–586.

Longhurst, N., & Norton, L.S. (1997). Self-assessment in coursework essays. *Studies in Educational Evaluation, 23*, 319–330.

Mehrens, W.A., Popham, W.J., & Ryan, J.M. (1998). How to prepare students for performance assessments. *Educational Measurement: Issues and Practice, 17*, 1, 18-22.

Peters, M. (1996). Student attitudes to alternative forms of assessment and to openness. *Open Learning*, 11(3), 48–50.

Rushton, C., Ramsey, P., & Rada, R. (1993). Peer assessment in a collaborative hypermedia environment: a case study. *Journal of Computer-Based Instruction, 20*, 75–80.

Sluijsmans, D., & Van Merriënboer, J.J.G. (2000). *A peer assessment model.* Heerlen: Open University of the Netherlands. (Center for Educational Technology and Expertise).

Sobral, D.T. (1997). Improving learning skills: a self-help group approach. *Higher Education*, *33,* 39–50.

Sprangers M., & Hoogstraaten, J. (1989). Pretesting effects in retrospective pretest-posttest designs. *Journal of Applied Psychology, 74*, 265-272.

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. (1998). *Instructional design for competency-based learning in post-secondary higher education*. Heerlen: Open University of the Netherlands. (Center for Educational Technology and Expertise).

Warkentin, R.W., Griffin, M.M., Quinn, G.P., & Griffin, B.W. (1995, April). *An exploration of the effects of cooperative assessment on student knowledge structure*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Williams, E. (1992). Student attitudes towards approaches to learning and assessment. *Assessment and Evaluation in Higher Education, 17, 45–58.*

Appendices

Appendix 1. Description of the constituent peer assessment skills

First level	*Description*
Define assessment criteria	The student actively participates in a group discussion to reach a common understanding about the assessment criteria for the product to be assessed
Judge the performance of a peer	The student assesses individually a product of a peer by first analysing the product and then formulating the discrepancies between the product and the criteria. The formulated discrepancies are written down in a peer assessment report
Provide (anonymous) feedback for future learning	The student writes a feedback report that provides feedback for future courses. This feedback: • confirms that the peer's understanding of what the product required was correct; • helps the student to add information to his own knowledge when they experience an information gap; • helps the peer to replace the erroneous information with more accurate information.

Second level	*Description*
Develop 'personal' course objectives on the basis of given course objectives and group discussion	The student presents his personal interpretations of the course objectives and argumentates his view in a group session
Describe a personal report on course objectives	The student individually writes a report that reflects his interpretation of the course objectives
Couple course objectives to study tasks	In collaboration with his peers, the student relates the defined course objectives to the different tasks he has to carry out to reach the course objectives and formulates which part of the task contributes to which course objective
Develop measurable criteria for each study task	In collaboration with his peers, the student lists the criteria that were decided for the task; these criteria are the result of the task analysis
Analyse the performance of a peer	The student individually applies the assessment criteria to the product of the peer after reading the product and marks the evidence for the presence of the criteria
Formulate discrepancies in a peer assessment report	The student writes an assessment report on the quality of the product which reflects evidence for reaching the desired criteria at a certain level
Formulate points for improvement	The student writes individually a number of points for improvement based on the assessment criteria and the group discussions in which the assessment criteria were decided
Reflect on points of improvement for the peer	Based on the assessed product, the student individually presents and argumentates points for improvement to the peer

Third level	*Description*
Analyse given course objectives	The student interprets given course objectives based on prior knowledge and personal values
Summarise results of the group discussion	The student takes an active role in the group discussion and writes a report which represents the outcomes of the discussions
Analyse the study task	The student discusses the study task with the peers and formulates common criteria that the student must meet to carry out the task in a proper way

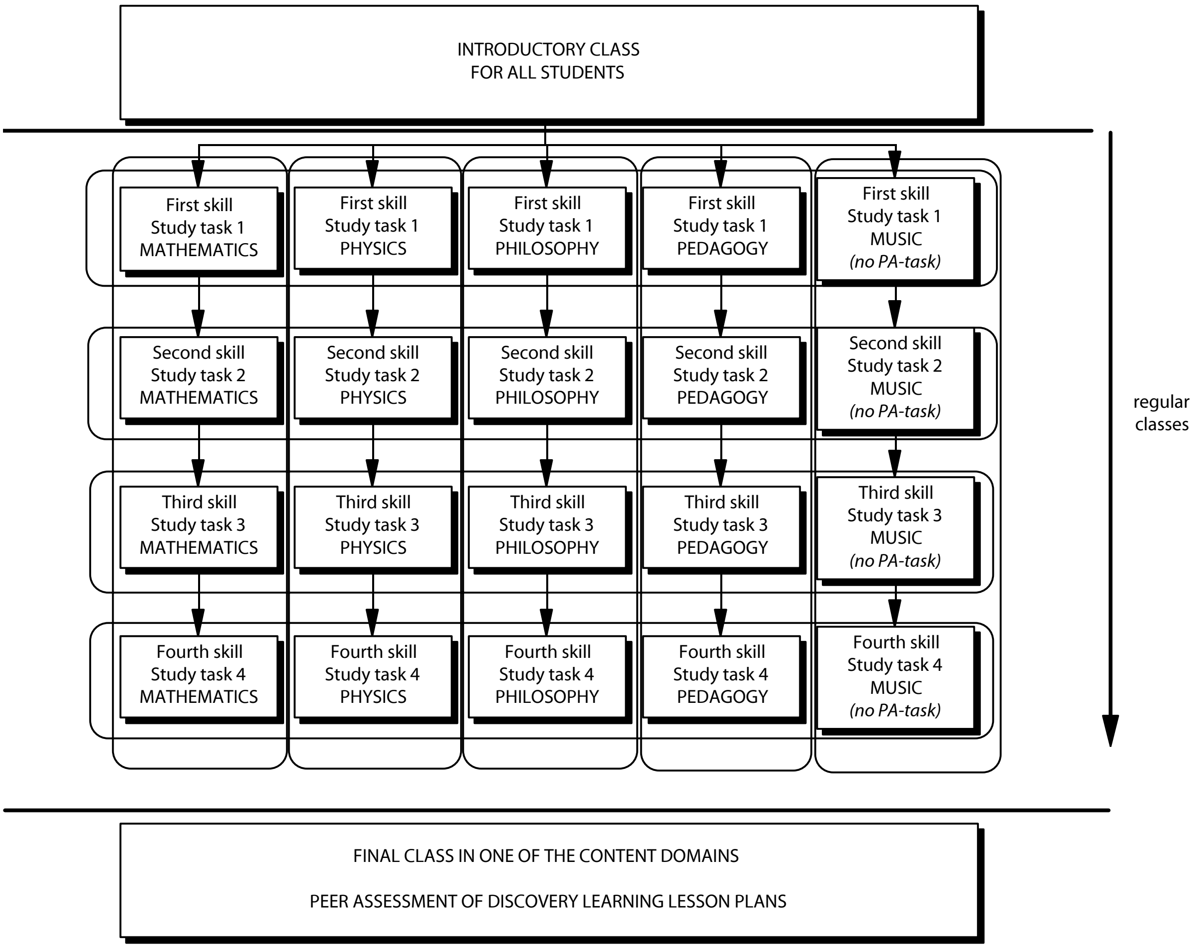
Appendix 2. Example of a part of a scheme in which studies are recorded according to several variables

	Author(s)/ publication year	*Title*	*Aim of the study*	*Subject; total n; type of education*	*Advantage*	*Disadvantage*	*Treatment/ Instrument*	*Criteria*	*Value of statistics reported*	*Conclusions*
1	Falchikov, N. (1995)	Peer feedback marking: Developing Peer Assessment	capitalize the benefits of peer assessment in terms of improving the learning process, sharpening critical abilities and increasing student autonomy	13 students of human developmental psychology (12 female, 1 male), mean age 21	PA is generally a useful, reliable and valid exercise	student over-marking	individual exercise	presentation: -structure -knowledge of topic -amount of information -delivery	no statistics reported	-PFM regarded as more fairer, more informative and challenging and harder than traditional marking, but less accurate, beneficial and enjoyable -PFM is making the learner think more, learn more and more critical
2	Conway, R., Kember, D., Sivan, A., & Wu, M. (1993)	Peer assessment of an Individual's Contribution to a Group Project	it examines ways in which students may be awarded individual marks, reflecting personal effort, for group projects	students in an optometry course at the Hong Kong Polytecnhic n=52	Not clearly mentioned	Not clearly mentioned	-a system which modified the group mark by a two-part weighting factor -a simpler scheme with a one-part multiplicative weighting factor	presentation: -enjoyment -preparation for future career evaluation of PA 12 students interviewed	PA of the group's presentation (25%) lecturer assessment of the group's presentation (25%) lecturer assessment of the group's report (50%) six statements on a Likert scale	-PA is a good method, fair enough -students should play a part in the assessment, but in a lesser portion

3	McDowell. L. (1995)	The Impact of Innovative Assessment on Student Learning	this article discusses the impact of innovative assessments forms on student learning	case studies at the University of Northumbria (England)	-develop-ment of evaluative and critical abilities -opportu-nities for skills development -integration of knowledge -collabora-tion, motivation, satisfaction, reward for effort -clarity of expectations	-organisational issues: cheating and fairness -the limitations of realism -erosion of motivation -time constraints -mismatch between learning and marks	unconventional exams, oral presentations, group projects, and peer assessment	No explicit criteria reported	no statistics reported	many positive effects on learning have been identified and students strongly support new assessment methods; problems include the difficulty in accommodating individual learner needs and acknowledging progress
4	Orsmond, P., Merry, S. & Reiling, K. (1996)	The importance of Marking Criteria in the Use of Peer Assessment	report of a study which implemented and evaluated a method of PA as a formative and summative assessment procedure	British university first-year undergraduate biology students' (n=39)	PA encourages students to take a more active role in cooperative learning, it develops students' own learning	-	in pairs students had to produced a scientific poster, a poster marking form was used to mark the posters	see appendix 1 en 2 (poster assignment and evaluation questionnaire)	correlations between tutor and peer assessment	a comparison between the tutor and the student peer mark may be misleading as a guide to the validity of peer assessment when individual criteria were analysed, agreement of peers and teacher ranged from 31-62%

5	Warken-tin, R. et al. (1995)	An Exploration of the Effects of Cooperative Assessment on Student Knowledge Structure	cooperative assessment was investigated in a classroom setting, examining achievement outcomes as measured by a multiple choice post-test of course content, a post-test of knowledge structure representation, and student perceptions of the cooperative assessment procedure	eighty-three undergraduate psychology students participated in this non equivalent control group study design	co-assessment offers opportunities for the negotiation of understanding it increases student motivation and the quantity of communication between teacher and students during the test	group instability, noisy classrooms, "stifling" of academically stronger students	pre-test: 30 MC-items	one figure illustrates the discussion, and two appendixes provide supplemental information	analysis of covariance	there were no significant differences between the groups on the post-test and that the hypothesis was not supported. there were also no differences between groups on similarity or coherence measures of student knowledge structure. Student reactions to the cooperative assessment procedure were positive
6	Pond, K., Ul-Haq, R. & Wade, W. (1995)	Peer Review: A Precursor to Peer Assessment	describes a two-year study focusing on the enhancement of student learning in tutorials through peer review and peer assessment	first-year undergraduates at Loughborough University of Technology (England)	SA and PA may promote higher levels of thinking	results may be unreliable or students may feel resentful	-	-	-	the findings of the two-year study generally support the available literature regarding reliability of peer generated marks and underlines the importance of "ownership" and understanding of review criteria

Appendix 3. Organisation of the redesigned course ‘Designing Discovery Learning Lesson Plans’



Curriculum Vitae

Dominique Sluijsmans was born on the 5th of November 1973, in Kerkrade, The Netherlands. After completing a pre-university secondary education at the 'Katholiek Gymnasium Rolduc' in 1993, she started her graduation in Educational Sciences at the Catholic University of Nijmegen. She graduated in 1997, after a work placement of a year at the KPC-Group in Den Bosch. Since September 1997 she has been working as a research associate at the Educational Technology Expertise Centre (OTEC) at the Open University of the Netherlands. From September 1998 till February 2002 she worked on her thesis on 'Student involvement in assessment'. This project is situated in a large research program 'Instructional Design for Competency-based Learning in Post-secondary Higher Education', within the theme diagnosis (for information see http://www.ou.nl/otecresearch). Besides her work as a Ph.D.-student, she also advised teachers in higher education in the design of various forms of assessment. Since March 2002 she has been working as an educational technologist at OTEC.

Publications

Dochy, F., Segers, M., & Sluijsmans, D. (1999).The use of self-, peer-, and co-assessment in higher education: a review. *Studies in Higher Education, 24,* 3, 331-350.

Kirschner, P., Valcke, M., & Sluijsmans, D. (1999). Design and development of third generation distance learning materials: from an industrial second generation approach towards realizing third generation distance education. In J. Van den Akker, R.M. Branch, K. Gustavson, N. Nieveen, & Tj. Plomp (Eds.), *Design Approaches and Tools in Education and Training* (pp.81-93). ICO: Kluwer Academic Publishers.

Segers, M., Pletincx, J., Dochy, F., & Sluijsmans. D. (1999). *Using new assessment forms in problem based learning: some relations with program evaluation*. Paper presented at the American Educational Research Association, April 19-23, Montreal, Canada.

Sloep, P., Slot, W., Sluijsmans, D. & De Haan, D. (1999). Competentiegericht leren in een Virtueel Bedrijf. [Competency-based learning in a Virtual Company]. In K. Schlusmans, S. Everwijn, G. Kinkorst, C. Nagtegaal, & R. Slotman (Eds.), *Computergerichte leeromgevingen [Competency-based learning environments]* (pp. 193-208). Lemma: Utrecht.

Slot, W., Jansen, D., De Haan, D., & Sluijsmans, D. (1999). Het Virtueel Bedrijf: een leeromgeving voor de competentiegroei van een ICT-er. [The Virtual Company: a learning environment to develop competencies]. *Tijdschrift voor Informatica Onderwijs, 8*, 3, 101-105.

Sluijsmans, D. (in druk). Peer assessment als complexe vaardigheid. *Tijdschrift voor Medisch Onderwijs*. [Peer assessment as complex skill.]

Sluijsmans, D., Boon, J. & De Haan, D. (1998). *De alpha-run van het virtueel bedrijf: een evaluatie [The virtual company: a pilot study]*. Heerlen: Open University of the Netherlands. (Center for Educational Technology and Expertise).

Sluijsmans, D., & Brand-Gruwel, S. (2000). *Elkaar leren beoordelen in competentiegericht onderwijs. [Evaluating peers in competency-based learning]*. Paper gepresenteerd op de Onderwijs Research Dagen, May 24-26, Leiden.

Sluijsmans, D., & Brand-Gruwel, S. (2000). *Integratie van onderwijs en instructie in PABO's. [Integration of instruction and assessment in teacher education*]. Paper gepresenteerd op het VELON-congres, 10-11 April, Veldhoven.

Sluijsmans, D., & Brand-Gruwel, S. (2001). Integratie van toetsing en onderwijs: de docent als toetsontwerper. [Integration of instruction and assessment: the teacher as test designer]. In A. Wald en J. van de Linden (Eds) *Leren in perspectief* (143-156). Garant: Leuven/Apeldoorn.

Sluijsmans, D., & Brand-Gruwel, S. (2001). *Het trainen van beoordelingsvaardigheden bij PABO-studenten*. [Training student teachers in assessment skills]. Paper gepresenteerd op de Onderwijs Research Dagen, 25-27 juni, Amsterdam.

Sluijsmans, D., Brand-Gruwel, S., Van Merriënboer, J. (2001, submitted). Redesigning education for training peer assessment skills in teacher education.

Sluijsmans, D., Brand-Gruwel, Martens, R., Van Merriënboer, J. (2002). Peer assessment training in teacher education. Manuscript accepted for publication in Assessment and Evaluation in Higher Education.

Sluijsmans, D., Brand-Gruwel, S., Bastiaens, T, R., Van Merriënboer, J. (2002, submitted). The training of peer assessment skills to promote the development of self-assessment skills in teacher education.

Sluijsmans, D., Brand-Gruwel, S., Mittendorff, M., & Van der Veen, A. (2001). Het trainen van beoordelingsvaardigheden door peer assessment. [Training assessment skills through peer assessment]. In: J.C.A. van Loon, *Maatwerk in uitvoering* (pp. 157-164). Tilburg: Fontys Hogescholen.

Sluijsmans, D. & Dochy, F. (1998). Alternatieve toetsmethoden in studentgericht onderwijs. [Alternative assessment in student centred education]. *Tijdschrift voor Hoger Onderwijs, 16*, 4, 298-314.

Sluijsmans, D., & Dochy, F. (1999). *The use of self-, peer- and co-assessment in higher education: a review of the literature*. Paper presented at the biennial EARLI conference, Göteborg, Sweden, August 24-28.

Sluijsmans, D., Dochy, F., & Moerkerke, G. (1998). *The use of self- peer- and co-assessment in higher education*. Paper gepresenteerd op de Onderwijs Research Dagen, 13-15 May, Enschede.

Sluijsmans, D., Dochy, F., & Moerkerke, G. (1998). *The use of self- peer- and co-assessment in higher education*. Heerlen: Open University of the Netherlands. (Center for Educational Technology and Expertise).

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., & Imandt, B. (2001). *Leren beoordelen binnen het wiskunde-onderwijs op de PABO* [Assessment training within mathematics in teacher education] . Paper gepresenteerd op het VELON-congres, 27 March, Veldhoven.

Sluijsmans, D., Martens, R., & Verheijen, H. (2000). Peer assessment en onderwijsontwerp.[Peer assessment and instructional design]. *Onderwijsinnovatie, 2*, 17-24.

Sluijsmans, D., & Moerkerke, G. (1999). *Peer assessment in probleemgestuurde onderwijsomgevingen: ervaringen en kwaliteiten van studenten* [Peer assessment in problem based learning environments: experiences and qualities of students]. Paper gepresenteerd op de Onderwijs Research Dagen, 20-21 May, Nijmegen.

Sluijsmans, D., & Van Merriënboer, J.J.G. (2000). *A peer assessment model.* Heerlen: Open University of the Netherlands. (Center for Educational Technology and Expertise).

Sluijsmans, D., & Moerkerke, G. (1999). Student involvement in performance assessment: a research project. *European Journal of Open and Distance Learning.* http://kurs.nks.no/eurodl/eurodlen/index.html.

Sluijsmans, D., Moerkerke, G., Dochy, F., & Van Merriënboer, J. (2001). Peer assessment in problem based learning. *Studies in Educational Evaluation, 27*, 2, 153-173.

Sluijsmans, D., & Dierick, S. (2002). Studentbetrokkenheid in assessment door peer assessment. [Student involvement in assessment through peer assessment]. In F. Dochy, L. Heylen, & H. van de Mosselaer (Eds.). *Assessment in onderwijs*. Utrecht: Lemma.