Katholieke Universiteit Leuven
Faculteit Psychologie en Pedagogische Wetenschappen
Centrum voor Opleidingsdidactiek

Peer assessment as a tool for learning

Proefschrift aangeboden tot het verkrijgen van de graad
van Doctor in de Pedagogische Wetenschappen
door Sarah Gielen

O.l.v Prof. Dr. Filip Dochy en Prof. Dr. Patrick Onghena

2007

KATHOLIEKE UNIVERSITEIT LEUVEN
Faculteit Psychologie en Pedagogische Wetenschappen
Centrum voor Opleidingsdidactiek

PEER ASSESSMENT AS A TOOL FOR LEARNING

Proefschrift aangeboden tot
het verkrijgen van de graad van
Doctor in de Pedagogische Wetenschappen

door **Sarah Gielen**

o.l.v. Prof. Dr. Filip Dochy
Prof. Dr. Patrick Onghena

2007

SUMMARY

The dissertation includes three theoretical contributions and three empirical studies on peer assessment, a general introduction and final reflections including a discussion of the results, a discussion of the educational implications and a discussion of some methodological issues.

The first contribution delineates the role that peer assessment can play in raising the consequential validity of an assessment system. First, it clarifies the type of effects that assessment in general can have on learning, and formulates the design principles for increasing the consequential validity of an assessment system. Then, it is shown that peer assessment helps to meet the identified design principles that enhance consequential validity of an 'assessment system'. More specifically, this dissertation shows that peer assessment can make it more feasible to include challenging and authentic tasks in one's assessment system; it can help making the assessment demands more clear to the students; it can provide a supplement or a substitute for formative staff assessment; and finally, it can support the response to teacher feedback.

The second contribution goes beyond the impact of peer assessment on the consequential validity, and addresses the problem that the output of peer assessment is evaluated against a variety of quality criteria in the literature, resulting in a cluttered picture. The different conceptualisations of quality that appear in the literature are analysed. It is shown that discussions about the most appropriate quality criteria for the output of peer assessment should be brought back to the underlying differences in goals. The most obvious goal is its use as an assessment tool. The learning goal of peer assessment has also been well-established. Investigating the literature more closely yields three additional goals: installation of social control in the learning environment; preparation of students for self-monitoring and self-regulation in lifelong learning; and active participation of students in the classroom. Each of these goals results in different quality criteria. It is argued that only the criteria that are congruent with the goal that one is trying to achieve should be considered when evaluating the quality of peer assessment.

The third contribution starts from the observation that, together with the expansion of peer assessment research in the last decade also the diversity of peer assessment practices has increased exponentially. This diversity poses difficulties for practitioners as well as researchers. An inventory of peer assessment diversity is developed that may be of interest to practitioners, as a checklist of important decisions to take or an overview of possible alternatives to a specific practice, and to researchers as a guideline of which information to provide on the particularities of their peer assessment design.

The fourth contribution compares the impact of peer feedback and teacher feedback on student learning, addressing the question whether peer feedback can serve as a substitute for expert feedback. A pretest posttest control group design examines the long term learning effects of individual peer feedback and collective teacher feedback on writing assignments in secondary education (N=85). Moreover, it examines the added-value of two measures to support the response of the assessee to peer feedback: an a priori question form and an a posteriori reply form. The study showed no significant difference in students' progress on essay marks between the condition with plain substitutional peer feedback and the control condition with teacher feedback. However, both groups (plain peer feedback ànd teacher feedback) appeared to make significantly less progress then the groups in the 'extended' feedback conditions with the question or the reply form.

The fifth contribution examines a group of 68 first year students in secondary education who experienced formative peer assessment for three successive writing assignments. They were divided in two experimental conditions (similar to the 'extended' feedback conditions in the previous contribution) and a control condition with plain peer feedback. Students' progress in writing performance is examined against the constructiveness of the peer feedback they gave and received, and against the condition in which they participated. The effect of the constructiveness of feedback is studied from two directions: from the point of view of the receiver of the peer feedback ('assessment for learning') and from the point of view of the assessor who gave peer feedback ('assessing for learning'). The results of a repeated measures analysis show a significant positive effect of the composition of the received peer feedback on student performance. The constructiveness of feedback that students provided themselves was not found to improve their learning. Nevertheless, the overall level of constructiveness of the feedback was low. Possible barriers preventing students from providing good feedback, and solutions to these, are discussed in the paper. Finally, the study could not replicate the effect of condition that was found in the fourth contribution.

The sixth contribution compares strengths and weaknesses of peer feedback and staff feedback, from the student's perspective. The study is situated in a university course with 192 first year students in educational sciences. Generic, collective staff feedback on the draft versions of a series of cumulative assignments is complemented with a formative peer assessment system. Starting from a hypothetical forced choice, a further in-depth study addresses the perceived characteristics of both sources of feedback and their perceived contribution to a learning environment that attends the learner's needs. These perspectives are complemented with reasons reported by students to prefer one of both sources of feedback. Closed-ended questionnaire items are triangulated with qualitative data from open-ended questions. Results show that approximately half of the students were willing to trade in the credibility of staff feedback for the specificity of peer feedback if they have to choose. However, both sources of feedback showed to have their own strengths and weaknesses from the student's perspective. They were complementary and they even provided the conditions under which the complementary source became better.

PREFACE
VOORWOORD

‘In learning, the tail wags the dog’, schreef mijn promotor ooit, hiermee verwijzend naar de belangrijke impact van assessment op het leerproces. Ik kan hem geen ongelijk geven. Ik heb in mijn doctoraatsproces zowel ‘pre-’, ‘true-’, als ‘post-assessment effecten’ ervaren, en ze hebben allemaal mee vorm gegeven aan mijn leerproces. De ‘pre-assessment effecten’ zijn verbonden met de hoge verwachtingen vanaf het begin, die me stimuleerden om het onderste uit de kan te halen. Filip en Steven, bedankt om in me te geloven en me de kans te bieden om in jullie onderzoekscentrum te tonen wat ik kon. Ik voelde me er thuis, en kreeg er alle vrijheid om me te ontwikkelen als onderzoeker, maar ook als onderwijskundige die een nauwe band met de onderwijspraktijk onderhield. Ook dank aan de wetenschappelijke commissie Psychologie en Pedagogiek van het Fonds voor Wetenschappelijk Onderzoek - Vlaanderen, die me in 2003 het vertrouwen schonk en me een beurs als ‘aspirant’ toekende om dit doctoraat voor te bereiden. Vervolgens vergeet ik ook niet de collega’s op het Centrum voor Opleidingsdidactiek, en de andere doctorandi aan het departement Pedagogische Wetenschappen, die me hielpen om de hoge verwachtingen te vertalen in kleinere tussenstappen en me toonden hoe ik ze kon bereiken. Katrien en Wouter, bedankt voor jullie goede raad, jullie waren inspirerende voorbeelden. Ook Goele, Stefan en Stijn, bedankt voor jullie aanmoedigingen en hulp waar nodig!

Het uiteindelijke schrijfproces, waarin ik rapporteerde over het uitgevoerde onderzoek, was een intensieve en leerrijke periode. Naast de ‘true-assessment leereffecten’ van het schrijven zelf, waren ook talrijke ‘post-assessment effecten’ verantwoordelijk voor de geboekte vooruitgang. In mijn onderzoek heb ik het over het belang van feedback, en dat is evenzeer op mijzelf van toepassing. Filip, bedankt voor je feedback en advies, ik was blij dat ik ten volle op jou mocht rekenen tijdens de laatste maanden. En dat geldt zeker ook voor Patrick. Je noemde jezelf ooit ‘co-promotor van het laatste uur’ omdat je niet vanaf het begin bij mijn onderzoeksproject betrokken was, maar dat ‘laatste uur’ was dan ook cruciaal. Jouw komst bracht een nieuwe dynamiek in mijn doctoraat waarvoor ik je dankbaar ben.

Verder dank ik alle andere co-auteurs van de verschillende manuscripten - Sabine, Katrien, Stijn, Stefan, Liesje, Elien, Steven, Wouter, en ook Wilfried - voor het grondig nalezen, corrigeren en becommentariëren van mijn eerdere versies. Bovendien hoort hier ook een dankjewel aan het 'Peer Assessment Collaboration Team' voor hun stimulerende feedback tijdens de eindfase. Ik heb effectief aan den lijve mogen ondervinden dat het krijgen van 'peer feedback' naast 'staff feedback' een onschatbare meerwaarde biedt.

Daarnaast wil ik zeker ook nog mijn appreciatie uitdrukken aan alle 'betrokkenen' bij mijn onderzoek voor hun bijdrage tot dit doctoraat:

Bedankt Elien, Liesje en Anneleen, om mijn klankbord te zijn tijdens jullie thesisperiode en om mee gestalte te geven aan een van de onderzoeksprojecten.

Bedankt Steven, Herman, Hans, Veerle, Roel en Katrin, en alle studenten van 1e bachelor Pedagogische Wetenschappen aan de KULeuven, academiejaar 2005-2006.

Bedankt Jo en Willem, en de deelnemende eerstejaarsleerlingen van het schooljaar 2004-2005 van het St.-Pieterscollege te Leuven.

Bedankt Kathleen, Lieve, José, en alle studenten van het derde jaar van de opleiding leraar lager onderwijs in de Katholieke Hogeschool Mechelen in de academiejaren 2004-2005 en 2005-2006.

En tenslotte, bedankt Kristin en de deelnemende studenten van 1e bachelor Toegepaste Economische Wetenschappen aan de KULeuven, academiejaar 2004-2005.

Niet al deze betrokkenen zullen 'hun verhaal' rechtstreeks herkennen in dit proefschrift, want ik heb meer projecten opgezet en meer data verzameld dan nodig was. Elk van deze verhalen is evenwel een onmisbare bouwsteen geweest in mijn groeiende inzicht in peer assessment als leermiddel.

Tot slot nog een welgemeend woord van dank aan alle vrienden en familieleden die me door dik en dun gesteund hebben tijdens de afgelopen jaren!

Sarah
10677 meter hoog boven Gaspésie (Canada)
15 april 2007

Dit proefschrift werd mede mogelijk gemaakt dankzij een beurs als Aspirant bij het Fonds voor Wetenschappelijk Onderzoek – Vlaanderen.

This doctoral dissertation was funded by a grant as a Research Assistant of the Research Foundation – Flanders.

TABLE OF CONTENTS

CHAPTER 1

GENERAL INTRODUCTION

GENERAL INTRODUCTION

The dissertation includes three theoretical contributions and three empirical studies on peer assessment, a general introduction and final reflections including a discussion of the results, a discussion of the educational implications and a discussion of some methodological issues with regard to the research. The six main chapters can be read on their own. They take the format of article manuscripts, each with their own abstract, introduction, methodology, results and discussion of the results. Each of these manuscripts is submitted to an international peer reviewed journal. The second chapter is an adaptation from a published chapter in an international edited book.

All chapters deal with a different conceptual area or a different empirical question within the domain of peer assessment, and all empirical studies are based on different datasets. However, some overlap in the introductory sections of the chapters could not be avoided, since each chapter is a stand-alone manuscript. The common theme in all manuscripts is the effectiveness of peer assessment as a tool for learning. Figure 1 provides a representation of the topics of the six main chapters, and their relationships, which are further elaborated below.

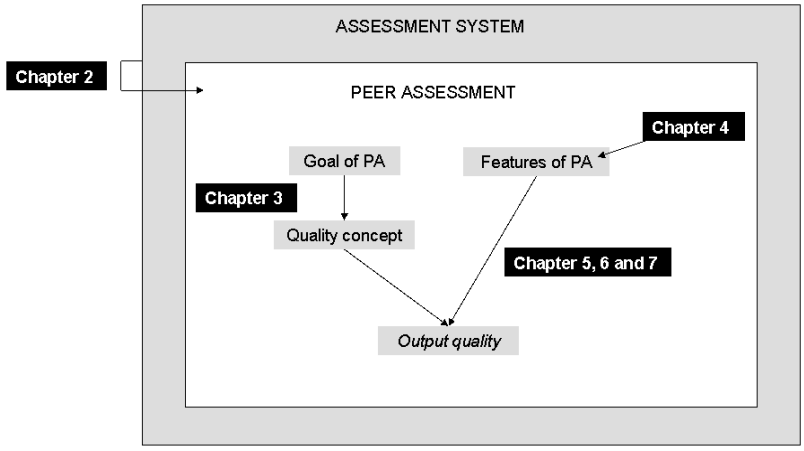


Figure 1. Overview of the main chapters of this dissertation.

Chapter 2, 3 and 4 are theoretical in nature and are based on a review of the available literature on peer assessment. However, all three chapters take a step further than merely providing a summary of the previous literature. Chapter 2, 3 and 4 contribute to a new framework for the study of peer assessment. The second and the third chapter address the place of peer assessment in the learning environment.

Chapter two, entitled "The impact of Peer Assessment on the Consequential Validity of Assessment", examines the role of peer assessment in the larger assessment system of a learning environment, and discusses its impact concerning the effects of this assessment system on the learning processes of students. Peer assessment is shown to be able to enhance the consequential validity of the larger assessment system. By introducing peer assessment, the contribution of an assessment system to a powerful learning environment can be strengthened.

Chapter three, entitled "Goals of Peer Assessment and their Associated Quality Concepts", focuses on peer assessment in itself, but goes beyond its impact on the learning processes, by analysing all goals that peer assessment can serve in an educational setting. Distinguishing the goals for which peer assessment is applied, appears useful to define and demarcate the appropriate quality conceptualisations and quality criteria to evaluate the effectiveness of peer assessment for a certain use. This analysis helps to clarify which quality criteria are appropriate to be investigated in the subsequent empirical studies that address the potential of peer assessment as a tool for learning.

The fourth chapter, entitled "An Inventory of Peer Assessment Diversity", develops a tool to analyse the features of a peer assessment practice, thereby providing a framework to capture and categorise the diversity of peer assessment in education. The aim of this 'inventory of peer assessment diversity' is to support practitioners and researchers in the field, by providing a checklist of important decisions to take when designing a peer assessment application, by providing an overview of possible alternatives when revising a specific application, and by providing a guideline of variables to describe when reporting on a peer assessment study. Moreover, it may inspire scholars to systematically study the effectiveness of some values of the variables in the inventory, or to combine previous research results in a way that takes account of the differences between studies with regard to the variables in the inventory. Examples of how this inventory can guide a

researcher in the description of a peer assessment practice under study can be found in the three subsequent manuscripts reporting on empirical studies of peer assessment.

In Chapter 5, 6 and 7, being empirical in nature, the focus lies on 'peer assessment as a tool for learning', and the impact of specific features of a peer assessment practice on its 'effectiveness' or 'output quality'. Based on the analysis in Chapter 3, the quality concept that is associated with the goal of 'peer assessment as a tool for learning' is defined as 'the effects of peer assessment on the learning environment and the learning outcomes of students'. This quality concept is the central issue of investigation in the empirical studies of the dissertation. The general concept is operationalised in several dependent variables within the different studies, as is shown in Figure 2.

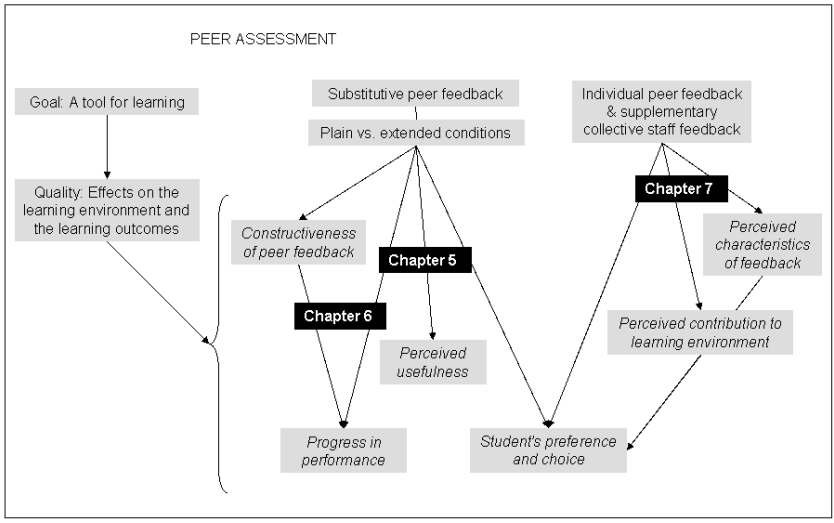


Figure 2. Overview of the research questions in the empirical studies of the dissertation.

Chapter 5, entitled "Peer Feedback as a Substitute for Teacher Feedback", and Chapter 6, entitled "The Effects of Constructiveness of Peer Feedback on Performance", study a substitutive peer feedback situation (i.e. formative peer assessment of draft artefacts that replaces formative teacher

assessment) and compare conditions with plain peer feedback and extended conditions with extra features. In Chapter 5 also a control condition with teacher feedback is present. The output measure in these studies is progress in performance, and in Chapter 5 also perceived usefulness of feedback and student's preference and choice for peer feedback. At an intermediate level, Chapter 6 studies the constructiveness of the provided peer feedback as an output as well as an explanatory variable.

The study in Chapter 7, entitled "A Complementary Role for Peer Feedback and Staff Feedback in Powerful Learning Environments", is situated in a setting where individual peer feedback is supplemented with collective, generic staff feedback. In this study, the first output measure is again the student's preference and choice, this time between peer and staff feedback. At the intermediate level, this study looks for explanations within the perceived characteristics of both sources of feedback and their perceived contribution to the learning environment.

The three empirical studies in this dissertation are complementary in their methodological approach, since Chapter 5 and 6 adopt a quasi-experimental approach and Chapter 7 adopts an in-depth case-study approach.

Chapter 8 closes this dissertation with some final reflections. It contains a summary and discussion of the results, in which the individual studies are exceeded and results are combined and compared. The contribution of this dissertation to the educational theory is demonstrated and suggestions for further research are added. Furthermore, this chapter discusses the educational implications of this dissertation. Finally, a reflection on some methodological issues is appended, in which strengths and weaknesses of the different research designs of the preceding chapters are discussed, and in which some additional information is provided with regard to the research process.

CHAPTER 2

THE IMPACT OF PEER ASSESSMENT ON THE CONSEQUENTIAL VALIDITY OF ASSESSMENT

Adapted from Gielen, Dochy, and Dierick, 2003

THE IMPACT OF PEER ASSESSMENT ON THE CONSEQUENTIAL VALIDITY OF ASSESSMENT

Introduction

In conjunction with the constructivist paradigm on learning and teaching (De Corte, 1996; Pellegrino, Chudowsky, & Glaser, 2001; Tynjälä, 1997), new ideas about assessment have arisen. These are referred to as the 'assessment culture' (Birenbaum, 1996; Segers, Dochy, & Cascallar, 2003). The role of assessment and evaluation in education has been crucial, probably since the earliest approaches to formal education. Currently this role is being broadened in educational theory, however. Whereas in the past we have seen assessment primarily as a means to measure the achievement of goals and thus for certification and selection, there is now a belief that the potential usages of assessing are much wider and impinge on all stages of the learning process. The new assessment culture strongly emphasizes the integration of instruction and assessment. This fundamental change in our views of assessment is represented by the notion of 'assessment as a tool for learning' (Dochy & McDowell, 1997).

It is not only teaching that has an influence on learning: assessment – both 'summative' and 'formative' – also almost inevitably influences learning. Assessment is therefore now recognised as being a part of the learning environment. It is expected to take its responsibility in supporting and steering the active learner: to become 'assessment *for* learning' in addition to 'assessment *of* learning' (Black & Wiliam, 1998).

The present chapter will focus on how peer assessment can be a tool to assist teachers who want to attain the 'assessment for learning' goal. We will first explain how assessment in general influences student learning, and relate this to the concept of consequential validity. Secondly, we will seek an answer to the question "How should assessment be designed to be a tool for learning?". A framework of design principles for assessment to support

learning will be introduced. This framework provides an overview of how the consequential validity of an assessment can be increased. Thirdly, we will investigate what the role of peer assessment may be in a learning environment that aims at assessment for learning, and what its impact is on the consequential validity of the existing assessment systems.

Learning & Assessment

Research into student learning has provided considerable evidence to suggest that student behaviour and student learning are influenced by assessment to a large extent (Black et al., 1998; Boud, 1990; Ramsden, 1992; Scouller, 1998; Thomas & Bain, 1984). This influence of assessment can occur on different levels and depends on the function of the assessment (summative versus formative). The question that is raised is *how* assessment can influence learning. We first need a map of possible 'influencing channels' before we can try to manipulate these channels: that is, the characteristics of the assessment that determine the extent or the nature of that influence on learning. The effects of assessment on learning can be categorized into three groups, as represented in Figure 1 and discussed below.

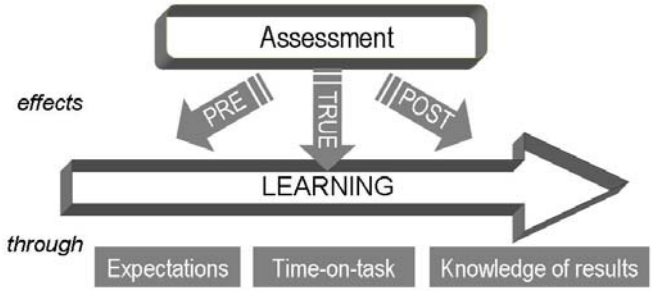


Figure 1. Effects of assessment on learning.

Post-Assessment Effects

The most well-known influence of assessment is due to the activity of looking back after the completion of the assessment task (referred to as "post-assessment effects"). Post-assessment effects deal with how

judgements about the quality of student performances shape, and hopefully improve, the students' competence by short-circuiting the randomness and inefficiency of trial-and-error learning (see also Sadler, 1989). Feedback is the most important trigger for these post-assessment effects to occur (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Hattie & Timperley, 2007; Kluger & DeNisi, 1996). Teachers give students information about the quality of their performance and support students in reflecting on the learning outcomes and the learning processes they are based on (Martens & Dochy, 1997). Feedback can also be obtained from fellow students and third persons, or it can even be generated by students themselves: self-reflection can lead to a kind of self-feedback or internal feedback, on condition that the assessment demands are clear (Butler & Winne, 1995). When students have the necessary metacognitive knowledge and skills to reflect on their achievements, weaknesses, and learning processes themselves, teacher feedback can be reduced.

Pre-Assessment Effects

A second kind of influence, however, is less obvious, but significant. This influence works pro-actively, since students tend to adjust their attention and learning behaviour to what they expect to be assessed. These effects can be described as pre-assessment effects, since such effects occur before assessment takes place, but are also called 'systemic validity' by Frederiksen and Collins (1989), 'backwash effects' by Biggs (1996), and finally the 'feedforward-function' by Starren (1998). The study of pre-assessment effects is related to the 'discovery' of the subjective learning environment, referring to students' perceptions of the learning environment, which may differ from the 'objective' learning environment (Sambell & McDowell, 1998). This 'discovery' is grounded in the increased attention given to student perceptions. All students have perceptions of the learning environment: some perceive it in approximately the same way that the teacher intended, others perceive it quite differently (Dart et al., 2000; Entwistle, 1991). Some students actively try to seek cues about what counts for the teacher (referred to as "cue-seekers" by Miller & Parlett, 1974), others just pick up what they come across ("cue-conscious") and still others just make up their own perceptions of the assessment, without noticing any cues ("cue-deaf"). In the subjective learning environment, the perceptions of the summative assessment requirements control student learning activities and

study motives to a large extent. “In learning, the tail wags the dog” (Dochy & McDowell, 1997, p. 291). Various authors now believe that assessment (including summative assessment) can no longer be regarded merely as an ‘independent observer’ that judges the worth of the teaching and learning process, by measuring the progress and achievements of students (Birenbaum & Dochy, 1996; Black et al., 1998; McDowell, 1995; Pellegrino et al., 2001; Segers et al., 2003; Wolf, Bixby, Glenn, & Gardner, 1991). Assessment is an accessory to this teaching and learning process, whether intentionally or not.

To make use of the pre-assessment effects (the expectations of students) in order to support learning, assessment should exude an “incentive power”. It should stimulate students to learn in a deep and thorough way and it should direct students to the desired learning goals. An important difference between the pre- and post-assessment effects is that the latter are often intentional whereas the first are, in most cases, more a kind of side-effect. Both, however, are important effects that need attention from teachers and instructional designers.

True-Assessment Effects

Nevo (1995) and Struyf, Vandenberghe, and Lens (2001) point to a third kind of learning effect from assessment. Students also learn during assessment itself, because they often need to reorganize their acquired knowledge, use it in different ways to tackle new problems, and to think about relationships between ideas that they had not discovered previously during their studies. Challenging assessment tasks provide an extension of the ‘time-on-task’ (Gibbs & Simpson, 2004). When assessment stimulates learners towards thinking processes of a higher cognitive level, it is possible that assessment itself becomes a rich learning experience for students (Struyf et al., 2001). We call this the true-assessment effect. In this true-assessment effect, the assessment task functions as a learning task, in the same way as an appropriate learning task would do without any assessment features. What could make a difference, however, is that students with an achievement motivation instead of an intrinsic learning motivation would probably engage less seriously with just a plain learning task, compared to an assessment task that ‘pays off’ (Gibbs, 1999).

Consequential Validity and its Extension with Regard to a Powerful Learning Environment

Not only in the educational literature, but also in the psychometrical theory, the answer to the question "What is a good assessment like?" has been broadened recently. The traditional consideration is whether the assessment is able to measure the 'real' construct under study, without any bias, to make a good judgement, is not longer perceived as sufficient (Kane, 2001; Linn, 1997; Moss, 1992). Among others, Messick (1989) has proposed to use a unified concept of 'construct validity', in which several aspects of validity are brought together. One of the extensions is to consider the values and consequences that are inherent to the assessment within its validity inquiry. "The consequential aspect [of construct validity] appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice" (Messick, 1995, p. 745). The latest version of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) also expanded the quality criteria for assessment by paying attention to the appropriateness of the use of a test or assessment.

This concern for the consequences of assessment has now been widely accepted as a valuable addition to the traditional quality indicators, and is often referred to as 'consequential validity' nowadays (e.g., Boud, 1995; Sambell, McDowell, & Brown, 1997). Messick (1995) himself, however, did not use the term 'consequential validity' to emphasise the unified nature of validity, in which the investigation of its consequences is just one part. On the other hand, in contrast to for instance Popham (1997) who argued that the concern about consequences should not be part of the 'validity investigation' in the strict sense of the concept, Messick preferred to include this inquiry within the concept of validity. Values and consequences are "intrinsic to the meaning and outcomes of the testing and have always been", and by making this explicit these aspects are "exposed to open examination and debate as an integral part of the validation process" (Messick, 1995, p. 748).

The expanding role of assessment in educational theory, described earlier, sheds another light on the meaning of 'consequences of assessment'. Within an *educational context,* researchers have taken a step further than only examining the values inherent to test interpretation and the consequences associated with the use of test scores. With the notion of 'assessment as a tool for learning', and as a result of the identification of pre- and true-assessment effects that take place *before* test scores are available, consequential validity is extended with the consequences of using a specific assessment instrument in itself. This goes beyond the consequences of the actual interpretation and use of its scores. The 'traditional' consequences referred to by psychometricians seem to fit with the 'post-assessment effects' of assessment. We argue, however, that in an educational framework a broader set of consequences for teaching and learning should be taken into account.

Moreover, because of the additional role of assessment as a tool for learning in educational settings, having 'neutral' consequences will not be sufficient either. 'Consequences' should be translated in 'positive pre-, true-, and post-assessment effects' and these positive effects should be capitalised. Only then can assessment be a powerful tool for learning. Quality criteria for assessment that are concerned with this *educational* responsibility of assessment are referred to by some authors as '*edu*metric criteria' (e.g., Dierick & Dochy, 2001). The definition of 'consequential validity' in this paper might thus be described as an 'edumetric view on consequential validity', and comes down to the evaluation of the strengths and weaknesses of an assessment in its role as a tool for learning.

The inquiry related to consequential validity investigates whether the actual consequences of assessment are also the expected consequences. In Table 1 an overview is given of questions that can be used as guidelines to collect supporting evidence for, and to examine possible threats to, the consequential validity of an educational assessment from an edumetric point of view. This information can be brought to the surface by methods such as presenting statements of expected (and unexpected) consequences of assessment to the student population, by holding semi-structured key group interviews, by recording student time logging (i.e., logging the time dedicated to assessment), or by administering self-review checklists.

Table 1

A framework for collecting supporting evidence for, and examining threats to, the consequential validity of an assessment

CONSEQUENCES OF ASSESSMENT USE
Procedure Searching for evidence: What does the assessment claim to do? Investigating if the actual consequences are also the expected consequences. What are the effects on the system of using the assessment, other than what the assessment claims? Are these effects beneficial or detrimental to learning?
Review questions How do students prepare themselves for education? What kind of learning strategy is used by students? What instructional strategies are used to prepare students for the assessment? Which kind of knowledge is measured? Does assessment stimulate the development of various skills? Does assessment stimulate students to apply their knowledge in realistic situations? Are long term effects perceived? Is breath and depth in learning actively rewarded, instead of merely by chance? Is independence stimulated by making expectations and criteria explicit? Is relevant feedback provided for progress?

Designing Assessment as a Tool for Learning

In the literature, several conditions under which assessment may support learning have been identified (Gibbs et al., 2004). These conditions can be translated into three design principles for assessment as a tool for learning. Paying attention to these three principles increases the consequential validity of an assessment. Firstly, a teacher is advised to create challenging and authentic tasks that match ambitious learning goals. Secondly, a teacher should make the assessment demands transparent to students. Finally, a formative function has to be integrated into the assessment system. These three principles are important to create a learning benefit from assessment through its pre-, true-, and post-assessment effects.

Create Challenging and Authentic Tasks that Match Ambitious Learning Goals

Deep level learning and collaborative learning can be provoked by making the assessment tasks authentic and challenging (Birenbaum et al., 1996; Gipps, 1994). The influence works through the pre-assessment effect (assuming students hold appropriate perceptions of the demands of the

assessment); through the post-assessment effect (assuming there is proper feedback); as well as through the true-assessment effect. Facing a challenging assessment is a stimulator to look up additional information, question the content more critically, discuss it with peers (collaborative learning), and structure it more personally (Gibbs et al., 2004). Facing an authentic assessment lets students focus on their ability to use their knowledge in a creative way to solve problems in ill-structured domains, and increases students' motivation to engage in the task (Wolf et al., 1991). When tasks are perceived as interesting and relevant, students will engage more effort in them, which in its turn is beneficial for learning (Dochy & Moerkerke, 1997). Even when students 'practiced' the assessed competencies before the assessment by means of real learning tasks, the assessment task may still realize a true-assessment effect since it will not allow a mere reproduction of what is learned. It thus extends the time-on-task in a meaningful way.

Make Assessment Demands Transparent to Students

A second design principle for enhancing the consequential validity of an assessment from an edumetric point of view is the transparency of the assessment process. Different authors point out that the transparency of the used assessment criteria has a positive influence on students' learning processes (Dochy, Segers, & Sluijsmans, 1999; Rust, Price, & O'Donovan, 2003). Indeed, "meeting criteria improves learning": if students know exactly which criteria will be used when assessing a performance, their performance will improve because they know which goals have to be attained and they can internalize these goals.

This also has an effect on student motivation. When learning goals become internalized by students, and when feedback emphasizes the relationship between performances and the achievement of these goals, effort becomes linked to goals instead of grades, being a more autonomous source of motivation (Deci & Ryan, 1985).

As has been indicated previously, making the assessment expectations clear to students by making the judgment criteria transparent, or even by involving students in the assessment, also has a supportive role in the anticipated effect of creating challenging and authentic tasks. In order to realize the pre-assessment effect of the first design principle, students have to have a correct interpretation of the assessment demands. If they do not

understand that deep level learning is required, they will not act upon that (Entwistle, 2000).

Furthermore, the positive effect of transparency on the quality of the learning behaviour is not only attributable to high expectations becoming clear, but also to the reduction in uncertainty about what is important. Uncertainty creates fear, and fear brings about the use of a surface approach (McDowell, 1995).

Finally, the transparency of assessment is not merely a way to clarify the demands of an assessment, so that the right learning will take place: it is also a way to stimulate the students' metacognitive skills. Having access to the assessment criteria supports the self-evaluation and self-regulation of students (Butler et al., 1995; Nicol & Macfarlane-Dick, 2006). It also contributes to the more effective use of feedback, since it is clear which goals need to be attained.

Integrate a Formative Function into the Assessment System

It is not only the challenging and authentic nature of the tasks and the transparency of the demands which foster a deep approach to learning. Students should also be guided through the learning process. Assessment, in addition to teaching, can play an important role in this (Black et al., 1998). The integration of assessment into the learning process ensures that students are encouraged to study in a more profound way during the course. They are encouraged to study at a stage when there is not yet great pressure on their time, which makes it possible to study in a more profound and personal way instead of 'quickly learning by heart' (Askham, 1997; Dochy et al., 1997; Sambell et al., 1997; Thomson & Falchikov, 1998). Additionally, the integration of assessment into the learning process has the advantage that students, through external and internal regulation, can get confirming or corrective input concerning deep learning behaviour (the formative function of assessment) (Martens et al., 1997). External regulation refers to the assistance provided by the teacher giving explicit feedback about their learning process and results. Internal regulation of the learning process is stimulated when students, based on the received feedback, gain insight themselves into their own levels of competence and how they can improve their learning behaviour (Askham, 1997).

Moreover, feedback can also have a positive influence on the intrinsic motivation of students. The key factor to obtain these positive

effects of feedback seems to be whether students perceive the primary goal of the assessment to be controlling their behaviour or providing informative and helpful feedback on their progress in learning (Deci et al., 1985; Keller, 1983; Ryan, Connell, & Deci, 1985). Birenbaum and Dochy (1996) emphasise that powerful learning environments are characterized by a good balance between discovery learning and personal exploration on one hand, and systematic instruction and guidance on the other, always taking into account the individual differences in abilities, needs and motivation between students. By giving descriptive feedback - not just a grade - and organizing different types of follow-up activities, a teacher creates a powerful learning environment.

A final crucial aspect of the positive influence of feedback is the way it is presented to students. Crooks (1988) identifies the following conditions for feedback in order for it to be effective. "First of all, feedback is most effective if it focuses on students' attention to their progress in mastering educational tasks" (p. 468). Therefore, it is necessary that an absolute, or self-referenced, norm is used (Meltzer & Reid, 1994; Wolf et al., 1991), so that students can compare actual and reference levels of performance and use the feedback information to alter the gap. This is also an essential condition to offer students with a normative concept of ability a possibility to realize constructive learning behaviour, since this context does not generate competitive feelings between them (which make them use defensive learning strategies). Moreover, Crooks (1988, p. 469) continues, "feedback should be given while it is still clearly relevant. This usually implies that it should be provided soon after a task is completed and that the student should then be given opportunities to demonstrate learning from feedback. Thirdly, feedback should be specific and related to its needs".

In short, formative assessment will have a positive influence on the intrinsic motivation of students, accelerating and sustaining the required (or desired) constructive learning processes. In order to do this, it should be embedded in a powerful learning environment and should take into account some crucial conditions for its feedback to be effective.

Increasing the Consequential Validity of Assessment through Peer Assessment

Definition of Peer Assessment

In practice there are several possible strategies to realize the aforementioned design principles for a consequentially valid educational assessment system. This study will examine one of them: the introduction of peer assessment into the assessment system.

Peer assessment in itself is not an assessment method like essay writing, portfolio assessment, the 'overall test', performance assessment, short answer test, or multiple choice test. Peer assessment can, in fact, be combined with all these assessment methods since the only fixed feature is that peers take the role of the assessor. To define the essence of peer assessment, a limited definition such as Topping's is actually sufficient: "Peer assessment is defined as an arrangement in which individuals consider the amount, level, value, worth, quality or success of the products or outcomes of learning of peers of similar status" (Topping, 1998, p. 250).

All the other characteristics of peer assessment are values of a list of variables, which can vary from case to case. For instance, it may include previous discussion or agreement over criteria, or use a teacher-defined list of criteria; it may involve feedback of a qualitative nature or, on the other hand, may involve students in marking. The assessment may be formative or summative; it may be supplementary to, or a substitute for, staff feedback; and it may be peer-to-peer or in groups (Dochy et al., 1999). In their inventory of peer assessment diversity, Gielen, Dochy, and Onghena (2007), list 20 variables to systematically describe and distinguish the multitude of divergent peer assessment practices.

The feature of involving peers in assessment, which is common to all peer assessment designs can, however, have a considerable impact on the consequential validity of the 'parent' assessment method, such as a portfolio, a test, or a performance assessment. This impact of peer assessment on student learning will be analysed in the remainder of the paper, which is organised according to what peer assessment may do to support the design principles described above.

Make Feasible Challenging and Authentic Tasks that Match Ambitious Learning Goals

Firstly, peer assessment makes it feasible to enrich assessment with more open-ended or complex assignments which are directed at deeper understanding, complex skills and attitudes. This type of assessment task is more challenging and authentic, but teachers may hesitate to use it frequently because it largely requires observation of behaviour or careful reading of extended reports, instead of correction of short answers. Limited resources are a constraint on teachers in addressing higher order learning goals in their assessments, and as a consequence they risk being left out of assessment (Wolf et al., 1991). By introducing peer assessment, assessments of these learning goals can take place more often since the observation or reading burden can be shared among multiple assessors. Moreover, the increased validity of the assessment of these open-ended assignments through multi-source assessments including peers as assessors (Conway & Huffcutt, 1997; Johnson, Olson, & Courtney, 1996) might also pave the way for a greater use of this type of assessment task. Finally, some information is not even accessible for teacher assessment, so peer assessment might be the only source of evaluative information apart from self-assessment. Social skills and engagement in group work are such an example. These aspects of performance cannot otherwise be evaluated, unless the teacher was to be present during all group meetings (Kane & Lawler, 1978).

As a result of using peer assessment to enrich the assessment, students will be encouraged to address their efforts to deeper levels and to engage in appropriate learning activities for meaningful learning (Gibbs et al., 2004).

Help to Make Assessment Demands Transparent to Students

Peer assessment may also support the second design principle. Assessment that communicates clear and high expectations is beneficial for learning (Gibbs et al., 2004). To obtain this effect, however, it is necessary that students understand and internalise these goals, criteria, and standards. In this regard, a special feature of peer assessment is helpful: in peer assessment students are not only assessees but also assessors. Allowing a learner to see what happens behind the curtains of an assessment, and to participate in it, supports clarification and internalisation of these goals (e.g., Rust et al.,

2003). Involving students in the process of formulating criteria gives them a better insight into the criteria and procedures of assessment. When, in addition to this, students are actually involved in the assessment process, they can experience personally (guided by an 'expert evaluator') what it means to evaluate and judge the performance against the criteria. This forms an additional support for their understanding of the expectations and the development of their self-regulation skills (Gipps, 1994; Sadler, 1998). Topping (1998) refers to this as an aspect of 'assessing for learning' (see also Gielen, Dochy, Onghena, Struyven, Smeets, & Decuyper, 2007).

Provide a Supplement to, or a Substitute for, Formative Staff Assessment

Formative assessment is important in order to provide feedback to students and to stimulate them to work during the course, not merely just before the final summative assessment (Gibbs et al., 2004). Extra assessment, however, requires extra staff time, which is not always available. Peer assessment can provide a relief by taking a supplementary, or even a substitutional, role.

To sustain and accelerate the desired learning processes, staff may integrate formative assessment tasks within the teaching. If these tasks are not marked or at least individually attended to by the teacher, however, they tend to be neglected by students. Peer assessment can be a solution to cope with the marking or reading burden of a growing number of such assignments (e.g., Forbes & Spence, 1991). The potential embarrassment of peers seeing their work, if it was of poor quality, increases the time and effort spent by students on these assignments (Cole, 1991; Gibbs et al., 2004; Pope, 2001). Pope (2001) compared the impact of the announcement that peer assessment would take place against the impact for teacher assessment, and found that students' stress levels, as well as their performances, were considerably higher when they expected a peer to correct their work.

In terms of providing feedback, peer assessment can also prove helpful. Although peers are not experts in the domain, their feedback can be a trade-off against expertise in terms of being understandable, timely, frequent, extended, individualised and reassuring. All these characteristics will be discussed in turn.

Research shows that students often perceive peer feedback as more understandable and more useful than teacher feedback, because fellow students 'are on the same wavelength' (Topping, 2003). Teachers, being

experts in the domain, often provide feedback that is based on a thorough insight into the complexities of the subject and the expectations of a domain. Although, as teachers, they should be able to translate this for their students, research shows that they do often not succeed in this. Their feedback is often not understood or is misinterpreted (Gibbs et al., 2004; Yang, Badger, & Yu, 2006). According to Higgins (2000), feedback messages like '*be more critical'* or '*your arguments need to be more academic'* do not have the same meaning for teachers as they do for students, because they are associated with a specific discourse that is not directly accessible to students (Hounsell, 1987).

Secondly, peer feedback can realise a gain in the speed of return of feedback. Teacher feedback is often provided with a considerable delay after the submission of an assignment or administration of a test. Assessing the work of a large group of students does take time, and assessment often receives a low priority in teachers' agendas. As a result, feedback sometimes is not available until after the course has finished and this feedback is likely to be a waste of time. In that case, "imperfect feedback from a fellow student provided almost immediately may have much more impact than more perfect feedback from a tutor four weeks later" (Gibbs et al., 2004, p. 19).

Thirdly, the frequency or amount of feedback can also increase with peer feedback. Gibbs and Simpson (2004) emphasise that for feedback to be useful it should be provided regularly, at each step in a learning process. Waiting until the end, and for instance only commenting on the final essay or report of a project, is not enough to support learning effectively and may provoke a lot of frustration on the part of the learner. An introduction of several 'intermediate' peer assessment sessions on draft versions of the essay or report could bring a solution, if staff are not able or not willing to increase their frequency of providing feedback.

A fourth possible advantage lies in the level of individualisation of feedback. If staff try to provide more timely and more frequent feedback, they often choose to organise it collectively to make this feasible. However, collective feedback cannot address personal needs as effectively as individual feedback. Moreover, the opportunity for personal interaction, identified as crucial by Sadler (1998) decreases: perhaps the possibility of asking questions is offered, but a student has to share the teacher's time with several other students, and in his answer a teacher will try to address the collective interest in the question at the expense of personal interest. Moreover, students

are not likely to show their ignorance or uncertainty during a collective session, so a lot of existing questions will not even be posed. Peer feedback can make it feasible to provide individual feedback and in the meantime, since the teacher does not have to provide the general feedback in front of the class, the teacher may be available for personal interaction when assessors and assessees cannot find an answer to a specific question.

A final argument in favour of peer feedback lies in the association of feedback with power issues, emotions and identity that may launch an 'emotion-defence system' in students (Kluger et al., 1996). As a consequence, students may hide their weaknesses and doubts from the teacher, rendering teachers unaware of particular student difficulties or misconceptions (Higgins, 2000). In that case, teacher feedback is less likely to connect to the learner, since it fails to address their problems or concerns. Peer feedback may by-pass some of these difficulties since it is less power-sensitive.

Support the Response to Teacher Feedback

Formative assessment and feedback (design principle 3) can only be beneficial to learning if it is received and attended to by students, and if it is acted upon by students to improve their work or their learning (Gibbs et al., 2004). A final impact of peer assessment on the consequential validity of assessment deals with assuring this response of students to the feedback that is provided by the teacher, thereby closing the 'feedback loop' (see Boud, 2000) and encouraging a 'mindful reception' of it (Bangert-Drowns et al., 1991). Some 'tactics' described by Gibbs and Simpson (2004) to address this issue are the introduction of two-stage assignments, providing only feedback on aspects that students request, and giving greater emphasis to generic feedback. These aspects, however, do not help when students have difficulties in understanding the teacher feedback, as was discussed above. Peer assessment can be helpful for this problem, since the development of students' ability to understand feedback (which is important for further learning) is expected to be one of the outcomes of the use of peer assessment. Having a clear view of the goals, criteria, and standards is necessary to understand what feedback is aiming at, and contributes to the more effective use of feedback. Making students participate in the assessment process, as assessors, enlarges their insight into it (Bloxham & West, 2004). It can help to bridge the gap between the 'discourse of the learner' and the 'discourse of

the expert' that is a source of many misunderstandings in teacher-to-student feedback. This final impact of peer assessment is again an example of 'assessing for learning' (see also Gielen, Dochy, Onghena, Struyven, et al., 2007).

Conclusion and Discussion

Traditionally we think of education as learning that takes place during a teaching period, and at the end of that period comes an assessment which measures whether the goals were attained (whether the learning has been effective). When we think of a good learning environment for supporting and steering learning, we think of curricula, teaching methods, and learning tasks. However, assessment proves to be an active player in the field too. Assessment already influences learning long before the assessment itself takes place, and its influence lasts long after that moment too.

This paper analyses how the influence of assessment on learning takes place and how it can be steered in the desired direction. It analyses three types of assessment effects: pre-, true-, and post-assessment effects. The post-assessment effect describes how judgements about the quality of student performances shape and hopefully improve the students' competences by means of feedback. Pre-assessment effects work pro-actively, since students tend to adjust their attention and learning behaviour to what they expect to be assessed. Finally, the true-assessment effect refers to what students learn when tackling the assessment tasks themselves.

The importance of these effects urges the expansion of the traditional description of 'good assessment' as providing a 'good measurement', and even of the broader definition of 'construct validity' in which 'an inquiry of the consequences of the use of test scores' is included (Messick, 1989; Messick, 1995). Within the educational theory, it is now believed that assessment should – in addition – be designed to be 'powerful assessment' that supports learning through eliciting positive pre-, true-, and post-assessment effects; referred to as having a high consequential validity from an edumetric point of view.

Assessment cannot just be a measure of goal attainment: it is jointly responsible for the goal attainment. We have, therefore, to work out what the assessment should be like in order to be able to play its role fully in a

powerful learning environment. To meet these demands, three design principles have been formulated for assessment as a tool for learning: create challenging and authentic tasks that match ambitious learning goals; make assessment demands transparent to students; and integrate a formative function into the assessment system.

These design principles are not easy to accomplish in all learning environments. Although several other strategies are possible, this paper discussed one strategy to attend to these design principles and to raise the consequential validity of an assessment system: the use of peer assessment in the learning environment. Involving students in the assessment process can bring about some relief for the teacher in empowering his learning environment. Four ways in which this can happen are described. Peer assessment can make it more feasible to include challenging and authentic tasks in an assessment system; it can help make the assessment demands clearer to the students; it can provide a supplement to, or a substitute for, formative staff assessment; and finally, it can support the response to teacher feedback.

Finally, it should be noted that, although peer assessment has the potential to raise the consequential validity of the educational assessment system, whether it realises this potential depends on the actual implementation. Furthermore, beyond not realising the intended effects, peer assessment –as with all assessment methods– may also have unintended effects that affect the consequential validity of the assessment system. Peer assessment will not function properly: if it is not accepted by students; if students do not make an effort to execute their role of assessors as effectively as possible or if they are not competent to do; if too much workload from the teacher is passed on to the students; etcetera (see, for instance, Sluijsmans & Prins, 2006; Topping, 1998). The possible unintended effects that result from these problems are not addressed in this paper, since they depend too much on the characteristics of the specific learning environment into which peer assessment is integrated (see Topping, 1998, for an elaboration of this caveat).

References

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington: American Educational Research Association, American Psychological Association, & National Council on Measurement in Education.

Askham, P. (1997). An instrumental response to the instrumental student: Assessment for learning. *Studies in Educational Evaluation, 23,* 299-317.

Bangert-Drowns, R., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61,* 213-238.

Biggs, J. (1996). Assessing learning quality: Reconciling institutional, staff and educational demands. *Assessment & Evaluation in Higher Education, 21,* 5-15.

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.

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

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5,* 7-74.

Bloxham, S. & West, A. (2004). Understanding the rules of the game: Marking peer assessment as a medium for developing students' conceptions of assessment. *Assessment and Evaluation in Higher Education, 29,* 721-733.

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

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education, 22,* 151-167.

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

Butler, D. L. & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65,* 245-281.

Cole, D. (1991). Change in Self-Perceived Competence as a Function of Peer and Teacher Evaluation. *Developmental Psychology, 27,* 682-688.

Conway, J. M. & Huffcutt, A. I. (1997). Psychometric properties of multi-source performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10,* 331-360.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58,* 438-481.

Dart, B., Burnett, P. C., Purdie, N., Boulton-Lewis, G., Campbell, J., & Smith, D. (2000). Students' conceptions of learning, the classroom

environment, and approaches to learning. *Journal of Educational Research, 93,* 262-270.

De Corte, E. (1996). Instructional psychology: Overview. In E. De Corte & F. E. Weinert (Eds.), *International encyclopedia of developmental and instructional psychology* (pp. 33-43). Oxford: Elsevier Science.

Deci, E. L. & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*.

Dierick, S. & Dochy, F. (2001). New lines in edumetrics: New forms of assessment lead to new assessment criteria. *Studies in Educational Evaluation, 27,* 307-329.

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

Dochy, F. & Moerkerke, G. (1997). Assessment as a major influence on learning and instruction. *International Journal of Educational Research, 27,* 415-431.

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.

Entwistle, N. (1991). Approaches to learning and perceptions of the learning environment. *Higher Education, 22,* 201-204.

Entwistle, N. (2000). Approaches to studying and levels of understanding: The influences of teaching and assessment. In J. Smart (Ed.), *Higher Education: Handbook of theory and research (XV)* (pp. 156-218). New York: Agathon Press.

Forbes, D. & Spence, J. (1991). An experiment in assessment for a large class. In R. Smith (Ed.), *Innovations in engineering education* (pp. 97-101). London: Ellis Horwood.

Frederiksen, J. R. & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18,* 27-32.

Gibbs, G. (1999). Using assessment strategically to change the way students learn. In S. Brown & A. Glasner (Eds.), *Assessment matters in Higher Education: Choosing and using diverse approaches* (pp. 41-53). Buckingham: SRHE & Open University Press.

Gibbs, G. & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education, 1,* 3-31.

Gielen, S., Dochy, F., & Onghena, P. (2007). An inventory of peer assessment diversity. In S. Gielen, *Peer assessment as a tool for learning* (pp. 67-94). Unpublished doctoral dissertation, Katholieke Universiteit Leuven, Leuven, Belgium.

Gielen, S., Dochy, F., Onghena, P., Struyven, K., Smeets, S., & Decuyper, S. (2007). Goals of peer assessment and their associated quality concepts. In S. Gielen, *Peer assessment as a tool for learning* (pp. 41-66). Unpublished doctoral dissertation, Katholieke Universiteit Leuven, Leuven, Belgium.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. London: Falmer.

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77,* 81-112.

Higgins, R. (2000). "Be more critical!": Rethinking assessment feedback. In *Paper presented at the British Educational Research Association Conference.* Cardiff University.

Hounsell, D. (1987). Essay writing and the quality of feedback. In J. Richardson, M. W. Eysenck, & D. W. Piper (Eds.), *Student Learning: research in education and cognitive psychology* (Milton Keynes: Open University Press.

Johnson, J., Olson, A., & Courtney, C. (1996). Implementing multiple perspective feedback: An integrated framework. *Human resource management review, 6,* 253-277.

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

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319-342.

Keller, J. M. (1983). Motivational design of instruction. In C. M. Reigeluth (Ed.), *Instructional design theories and models* (pp. 383-434). Hillsdale: Erdbaum.

Kluger, A. N. & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin, 119,* 254-284.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16,* 14-16.

Martens, R. L. & Dochy, F. (1997). Assessment and feedback as student support devices. *Studies in Educational Evaluation, 23,* 257-273.

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

Meltzer, L. & Reid, D. K. (1994). New directions in the assessment of students with special needs: the shift toward a constructivist perspective. *The journal of special education, 28,* 338-355.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3 ed., pp. 13-103). New York: American Council on Education/ Macmillan.

Miller, C. M. L. & Parlett, M. (1974). *Up to the mark: A study of the examination game*. London: SRHE.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for Performance Assessment. *Review of Educational Research, 62,* 229-258.

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

Nicol, D. J. & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education, 31,* 199-218.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know. The science and design of educational assessment*. Washington, DC: National Academy Press.

Pope, N. (2001). An examination of the use of peer rating for formative assessment in the context of the theory of consumption values. *Assessment & Evaluation in Higher Education, 26,* 235-246.

Popham, W. J. (1997). Consequential validity: Right concern, wrong concept. *Educational Measurement: Issues and Practice, 16,* 9-13.

Ramsden, P. (1992). *Learning to teach in higher education*. London: Routledge.

Rust, C., Price, M., & O'Donovan, B. (2003). Improving students' learning by developing their understanding of assessment criteria and processes. *Assessment & Evaluation in Higher Education, 28,* 147-164.

Ryan, R. M., Connell, J. P., & Deci, E. L. (1985). A motivational analysis of self-determination and self-regulation in education. In C. Ames & R. Ames (Eds.), *Research on motivation in education: Vol2. The Classroom milieu.* New York: Academic Press.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18,* 119-144.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in education, 5,* 77-84.

Sambell, K. & McDowell, L. (1998). The construction of the hidden curriculum: Messages and meanings in the assessment of student learning. *Assessment & 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.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education, 35,* 453-472.

Segers, M., Dochy, F., & Cascallar, E. (2003). *Optimizing new modes of assessment: In search of qualities and standards*. Dordrecht: Kluwer Academic.

Sluijsmans, D. & Prins, F. (2006). A conceptual framework for integrating peer assessment in teacher education. *Studies in Educational Evaluation, 32,* 6-22.

Starren, H. (1998). De toets als hefboom voor meer en beter leren. *Academia,* 26.

Struyf, E., Vandenberghe, R., & Lens, W. (2001). The evaluation practice of teachers as a learning opportunity for students. *Studies in Educational Evaluation, 27,* 215-238.

Thomas, P. R. & Bain, J. D. (1984). Contextual dependence of learning approaches: The effects of assessment. *Human Learning, 3,* 227-240.

Thomson, K. & Falchikov, N. (1998). 'Full on until the sun comes out': the effects of assessment on student approaches to studying. *Assessment & Evaluation in Higher Education, 23,* 379.

Topping, K. J. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy, & E. Cascallar (Eds.), *Optimizing new modes of assessment: In search of qualities and standards* (pp. 55-87). Dordrecht: Kluwer Academic.

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

Tynjälä, P. (1997). Developing education students' conceptions of the learning process in different learning environments. *Learning and Instruction, 7,* 277-292.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17,* 31-74.

Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15,* 179-200.

CHAPTER 3

GOALS OF PEER ASSESSMENT AND THEIR ASSOCIATED QUALITY CONCEPTS

Abstract

The output of peer assessment in higher education has been increasingly investigated in recent decades. However, this output is evaluated against a variety of quality criteria, resulting in a cluttered picture. This paper analyses the different conceptualisations of quality that appear in the literature. It is shown that discussions about the most appropriate quality criteria for the output of peer assessment should be brought back to the underlying differences in goals. The most obvious goal is its use as assessment tool, and the learning goal of peer assessment has also been well-established. Investigating the literature more closely yields three additional goals: installation of social control in the learning environment; preparation of students for self-monitoring and self-regulation in lifelong learning; and active participation of students in the classroom. Each of these goals results in different quality criteria. Only the criteria that are congruent with the goal that one is trying to achieve should be considered when evaluating the quality of peer assessment.

GOALS OF PEER ASSESSMENT AND THEIR ASSOCIATED QUALITY CONCEPTS

Introduction

Higher education is on the move. Teachers (and also students) are expected to invest energy to change and to make accommodations towards new ways of teaching, new ways of studying, new ways of assessment (De Corte, 2000). Innovations are made in curricula and classroom practices, in an attempt to align them with recent theories of learning and instruction (Biggs, 1996).

In conjunction with changes in theories about learning and teaching (De Corte, 1996), new ideas about assessment, referred to as the 'assessment culture', arise (Birenbaum et al., 1996; Segers, Dochy, & Cascallar, 2003). The role of assessment in education has been crucial, probably since the earliest approaches to formal education. However, currently this role is being broadened in educational theory and practice (Gipps, 1994; Wolf et al., 1991). Whereas in the past we have seen assessment primarily as a means to measure the achievement of goals and thus for certification and selection, there is now a belief that the potential goals of assessment are much wider and impinge on all stages of the learning process and even beyond that (Gielen, Dochy, & Dierick, 2003).

The renewed interest in peer assessment is a result of this evolution (Falchikov, 1995). Peer assessment as a concept is not new at all, and even scientific research on peer-assessment is not new (e.g., Kane et al., 1978). However, research into peer assessment has grown rapidly in the last decade. A large number of these peer assessment studies deal with issues such as the effectiveness, acceptability, fairness or reliability of peer assessment. These issues may be summarised as questions about the quality of the output of peer assessment. A problem of the current body of studies, however, is that it provides a cluttered picture due to the use of a variety of quality criteria.

The purpose of this paper is to analyse and categorise the different conceptualisations of quality that appear in the literature on peer assessment. The hypothesis is that a discussion about the most appropriate quality criteria

for the output of peer assessment can be brought back to the underlying differences in the formulation of the goals of peer assessment.

Methodology

A combined search of the major bibliographic databases (see Table 1) in humanities was performed with the search string *"peer assessment" OR "peer review" OR "peer rating" OR "peer feedback" OR "peer marking" OR "peer correction" OR "peer appraisal"*. Depending on the database, we searched in the field of *keywords, subject* or *abstract*. Restricting the search results to publications in the field of education (*AND "education" in all fields)* yielded between 174 and 1196 studies, published between 1952 and 2006, in the different databases. The ERIC database proved to be the most comprehensive; the other databases delivered mostly duplicates. All studies were examined in search of conceptualizations of quality for peer assessment and the (often implicit) goal of using peer assessment in a certain practice.

Table 1

Search results for studies on peer assessment, peer review, peer rating, peer feedback, peer marking, peer correction or peer appraisal (P), related to education, in the major relevant bibliographic databases (10/02/2007).*

Database	Date of first P*-reference	Total # P*-references	AND education *(in any field)*
ERIC	1954	1197 *(in KW)*	1196
SSCI	1971	1577 *(in TS)*	174
Acad. Search Premier (restr: acad. journals)	1970	2451 *(in AB)*	483
PsycINFO	1952	1125 *(in AB)*	257

Overview

Peer assessment can serve several goals. Sorting them on a scale from external control to autonomy support gives us the following list: peer assessment as a tool for social control; for assessment; for learning; for learning-how-to-assess; and for active participation of students. The

formulation of new goals for peer assessment, apart from the well-known assessment tool, has introduced new expectations of peer assessment and, as a consequence, new concepts concerning the quality of peer assessment.

The problem is not so much that different goals, and thus different definitions of quality, exist: the problem is that a clear view on the relationship between these goals and definitions is getting lost. Some researchers and practitioners are not explicit about their intended goals for using peer assessment, but still draw conclusions on its quality. The risk here is that one takes a certain conceptualisation of quality for granted, without questioning the alignment with the implicit goal assumptions. When concepts of quality start to take on their own lives some discussions, such as whether one set of criteria is better or more appropriate than another, or whether peer assessment should always attempt to comply to all types of requirements at the same time, become unsolvable (e.g., Dierick et al., 2001; Stefani, 1998). Without a reference to the goal, it becomes difficult to decide which quality concept is most appropriate.

This review attempts to draw the whole picture by making an inventory of the available quality concepts and establishing, or re-establishing, the link with the underlying goals. This framework should make it easier for researchers and practitioners to understand why others focus on different quality concepts, and to decide which quality concept (and the associated instruments or measures) is most suitable for their particular situation. It is comparable to realising that bothering about the right background music when receiving guests is of no importance at all when these guests are deaf.

This paper is structured around the five major goals of using peer assessment listed earlier, and their associated quality conceptualisations that are found in the literature. Some quality concepts are translated into criteria that are reasonably straightforward to measure; others are formulated on a rather abstract level and their operationalisation is more ambiguous in literature. For each stance, some example studies will be discussed to show the logic within each ‘goal - quality concept’ association.

Goal 1: Peer Assessment as a Social Control Tool

Goal

The outcome itself of an assessment by peers is not always the major concern of a teacher. It is sometimes used instead as a precautionary measure, to make sure students do not get away with being lazy. The introduction of peer assessment should assure that specific valuable learning activities take place, even if the teacher cannot control everything. Knowing that peers will assess your work, or your behaviour, may be an external motivator to work harder and perform better.

Gibbs (1999) reports two mechanisms by which peer assessment may raise performance through social control. The first is an increased time on task. Some activities, such as making problem sheets, are beneficial to learning, but if the teacher has no time for marking them, there is no social pressure in problem classes to turn up prepared. Making peers assess each others' problem sheets is an excellent way of getting students to spend time on task, Gibbs observed.

In group work, the time on task is also raised through social control by peer assessment. When students are asked to rate all group members' contributions to the group work, each group member is encouraged to participate in all the different aspects of a group project. This prevention of 'free-riding' is in its turn beneficial to learning (e.g., Abson, 1994; Segers & Dochy, 2001). In some cases, this use of peer assessment in group work is explicitly linked to curricular goals, such as learning to cooperate and to manage group projects (Topping, 2003), and so peer assessment may be a tool to assess these goals (see goal 2), or to support the acquisition of them (see goal 3). In practice it often does not go that far, however, and peer assessment is merely a tool to exert social control.

A second mechanism involved in social control by peer assessment is that students pay more attention to feedback that has a social dimension, and as a result intensify their efforts. Gibbs (1999, p. 46) states: "Students care what others think about them. A piece of work submitted confidentially and given a dreadful mark by a tutor they hardly know, may have little impact. (…) Their peers and friends, seeing and judging the same hopeless work, in public, in front of others, is likely to have quite a dramatic impact". Cole (1991) and Pope (2001a; 2005) also found that learners' self-perceived

academic competence and self-esteem were more powerfully affected by their peers' evaluation than by their teacher's. Pope (2001a, p. 243) reports that "the knowledge that the work will be rated by peers seems to induce students to write to a higher standard".

Quality Concept

The definition of quality, when conceiving peer assessment as a tool to exert social control, is whether it works efficiently: for example 'are group dynamic problems actually avoided without requiring too much intervention by the teacher?', 'do students prepare their exercises, without the lecturer having to mark them individually?'. Success is typically reported in terms of more desired behaviour or less undesired behaviour, often accompanied by a saving in resources. Whether or not this finally results in higher performance or better learning depends on the effectiveness of 'the desired behaviours' (e.g., 'time on task' or 'effort') that are controlled by the peer assessment, but this is only indirectly related to the success of the peer assessment.

Goal 2: Peer Assessment as an Assessment Tool

Goal

Using peer assessment as an assessment tool is the most obvious practice. From this perspective, students are considered to be 'surrogate' or 'assistant' teachers. They are asked to grade, rank or rate each others' products or performances, and/or to provide qualitative comments to their peers.

This can be done in two ways. On one hand, peer assessment is sometimes assumed to be a partial replacement for staff assessment, but on the other hand it often becomes part of a triangulated approach to assessment in which student learning is evaluated from multiple data sources or by multiple assessors (Breitmeyer, Ayres, & Knafl, 1993; Johnson et al., 1996; Miller, 2003). "A major issue", Miller states, "to be considered when employing triangulation is whether the purpose of it is to achieve convergence among the assessment sources, or whether it is to achieve completeness, the uncovering of multiple perspectives on the behaviour being assessed" (p. 383). This is a distinction that has important consequences for the concept of quality. Actually, the partial replacement of staff assessment and the triangulation to achieve convergence can be considered to share the

same underlying principle. Therefore, these two will be considered as a first sub-goal which has quality criteria associated with it. The second sub-goal is the achievement of completeness, which will require different quality criteria.

Quality Concept

Peer assessment as a tool for assessment should comply with all the quality requirements for assessment in general. Birenbaum (2007) summarises the current view on this issue as follows: "In assessment we draw inferences and make interpretations about what the test-taker knows and is able to do in a defined target domain from his/her observed performance on tasks designed to represent that domain. It can therefore be asserted that the quality of a given assessment practice can be judged by the appropriateness, meaningfulness, and usefulness of these inferences/interpretations" (p. 30).

The article by Birenbaum "presents a framework based on the 'unified view of validity', advanced by Cronbach (1988) and Messick (1989) over two decades ago, to assist in generating an evidence-based argument regarding the quality of a given assessment practice" (p. 29).

However, in the peer assessment literature, this general perspective on quality is not widespread. Most researchers focus on the specific feature of using peers as assessors and do not examine issues such as the content, structure or sampling quality of this assessment. In the remainder of this section, we will discuss the different views on quality that are particularly related to peer assessment.

Quality criteria for the use of peer assessment as an assessment tool can be formulated at the level of the perceptions about this assessment held by stakeholders, or at the level of the judgements by peers (their marks or comments). At the first level, quality criteria are common to both sub-goals described above. At the second level, they differ.

1. Criteria at the Level of Perceptions by Stakeholders

At the level of the users of peer assessment, the question of quality is one of confidence in peer assessment and acceptance of its results by all stakeholders. This perspective has already been acknowledged by Kane and Lawler (1978). If the result of a peer assessment is not accepted as fair and accurate by the assessee or external stakeholders such as future employers, it cannot serve its goal as an assessment tool.

Moreover, stakeholders may differ in their opinions on the validity and reliability of peer assessment: Cho, Schunn and Wilson (2006) describe several reasons why students often perceive peer assessment as unreliable and invalid when from the instructor's perspective there is no problem at all. Robinson (2002) also reports that students' perceptions of fairness dropped when the number of peer assessors was raised, although objectively this intervention resulted in an increased level of reliability. Quality at the level of perceptions is thus not necessarily the same as quality at the level of judgements.

2. Criteria at the Level of Judgements by Peers

Criteria such as the 'objective' reliability and validity of a peer assessment are formulated at the level of the judgements made by peers. These quality concepts, however, have different meanings depending on the assumptions about peer assessment's relationship to staff assessment. As already discussed above, peer assessment may be considered as a replacement or a triangulation of staff assessment. Within this latter stance, convergence or completeness might be aimed for. It is important that the quality criteria are in line with these assumptions. The goals of replacement of staff assessment and convergence with staff assessment in a triangulated approach share the same quality concept and will be discussed together.

2a. Criteria for Sub-Goal 1: Replacement or Triangulation to Achieve Convergence

When peer assessment is seen as a substitute for staff assessment, or when convergence is the purpose of integrating peer assessment in a triangulated approach to assessment, the question of quality is mostly translated in terms of 'agreement' (e.g., Orsmond, Merry, & Reiling, 2000). Similar quality concepts are 'reliability' (e.g., Topping, 1998), 'accuracy' (Topping, 2003), 'consistency' (Marcoulides & Simkin, 1995), 'similarity' and 'concurrent validity' (Saito & Fujita, 2004). These requirements are applicable to peer-marking (e.g., Falchikov & Goldfinch, 2000; Magin & Helmore, 2001), as well as to peers providing a qualitative appreciation of the behaviour of a peer (e.g., Topping, Smith, Swanson, & Elliot, 2000), although they are mostly studied in relation to the quantitative version of peer assessment (Topping, 1998).

All the above described concepts of agreement require a comparison to another assessment, to be able to measure the quality of the peer's assessment. There is still possible variation in the choice of the assessment to which a comparison is made. Magin and Helmore (2001) distinguish two possible references for comparison: the assessment of the teacher and the assessment by an equal status assessor.

Comparison with the teacher. The most obvious reference for comparison is the judgement of the teacher, tutor or another professional assessor. In this case, all divergence is attributed to the malfunctioning of the peer as assessor. Although this comparison of peer and staff marks or comments is sometimes referred to as a measure of reliability of peer assessment, Topping (2003) and Falchikov et al. (2000) clearly explain that this comparison (by means of a mean difference, a correlation, a difference in variance, or some other measure) is actually a measure of validity. This is the case because one compares the marks or comments to a normative reference instead of comparing equal status assessments by different persons or at different moments.

Comparison with other peers. A second possible reference is the assessment by an equal status assessor: one or several other peer assessors who comment on or mark the same product, performance or process. In such cases the inter-observer or inter-rater reliability can be studied (Magin et al., 2001), or a generalisability coefficient for several numbers of assessors can be calculated (Segers et al., 2001). A large agreement (i.e., a small variance) between peer assessors is considered a sign of high quality. If the variance is large due to an outlying measurement, this is considered a bad measurement; if the variance is large due to a wide spread of observations, the reliability of the whole peer assessment is doubted. Rada and Hu (2002) have designed software to automatically perform this type of 'quality control' in an online peer assessment setting by detecting suspicious patterns of large variation (high range of individual scores) or outlying mean scores, called 'out-of-control assessments'.

Comparison with another episode. A third possible reference is an assessment by the same assessor at another moment in time. This type of comparison, mentioned in the review study of Kane et al. (1978), may in fact

also be considered as a comparison to an equal status assessor. Stability or test-retest reliability measures compare a peer assessment score or comment with a score or comment on the same performance collected during another peer assessment episode.

Comparison with self assessment. A fourth possible reference is described by Falchikov (1993) in a study where group members assessed the process of working together on a small group project. Since both self and peer assessment took place, she studied the consistency in marking when comparing self ratings with peer ratings. Even when there are no self assessments available from the assessee, asking them if they think the peer assessment is 'fair' refers in fact to the same comparison.

2b. Criteria for Sub-Goal 2: Triangulation to Achieve Completeness

When peer assessment is part of multi-source assessment (Conway et al., 1997; Johnson et al., 1996), and serves to 'uncover the presence of multiple perspectives about the performance being assessed', assessors do not necessarily have to agree; they do not need to converge on a 'single truth about the quality of the performance' (Miller, 2003, p. 390). The question of quality here is not 'who is right?' but (1) 'are all different opinions an enrichment for the final assessment (do they contribute to its construct validity)?' and (2) 'are they transparent to all participants in the assessment process and are the underlying differences in conceptual frameworks and evaluation schemes becoming clear' (Liu & Tsai, 2005)? .

The presence of 'multiple perspectives' has been interpreted in the literature in two ways: as multiple expectations and as multiple information sources. These lead towards slightly different conceptualisations of 'enrichment and transparency'.

Completeness in expectations. "It is conceivable", as Miller (2003) says, "that different groups of assessors may have different expectations of a performance, and this would affect their assessment. (…) Different expectations would likely diminish the convergent validity of an assessment, but can strengthen the 'completeness' validity, as long as the differing perspectives are identified" (pp. 390-391). Peers, for example, might have different expectations of a presentation or a poster from teachers, or from practitioners, because they are interested in different topics or they have a

different level of prior knowledge. For instance, students might think that a real life example of a theory is important to mention, while teachers might not value this and prefer to see a comparison of theories at a more abstract level. Or students might appreciate specific references to the constructs and definitions taught in class when analysing an everyday problem, while practitioners might consider these of minor importance. However, if the presentation or the poster is meant to address both groups, both sets of expectations are valid (Conway et al., 1997). The quality criteria are whether peer assessment realises an enrichment of the expectations on which the judgement is based, and whether peers are transparent about the expectations they use for their judgement.

One may distinguish between assessors who focus on different aspects of the performance and assessors who have varying, or even opposing, opinions concerning the same aspects of the performance (Topping et al., 2000). Topping and his colleagues (2000) seem to conclude that differences in the criteria (or the interpretation of these criteria) between the assessors will enrich the multi-source assessment, but that differences in judgements on the same criteria, due to a disagreement on standards, need to be avoided.

Completeness in information. Another possible factor that may explain differences in opinions between assessors concerning the same aspect of performance, aside from the difference in standards or interpretation of the criterion, is a difference in the information to which one has access. This issue, amongst others, is discussed by Kane and Lawler (1978). An advantage of using peer assessment for the assessment of group work is that the assessment method might be sensitive to information about each group member that is only accessible to other group members. If this is the case, Kane et al. (1978) argue, we would expect lower validities for outsiders' (i.e., non group members') judgements, such as those from teachers, not because they focus on other aspects but because they have to base their judgement on other (more limited) information about the performance. In case of the assessment of complex behaviours, it is possible that not all assessors can observe every relevant aspect. So multiple assessors, and multiple types of assessors, are needed to collect the different pieces of the puzzle. The quality question then becomes whether peer assessment realises an enrichment of the

information on which the judgement is based, and whether peers are transparent about the information they use for their judgement.

Goal 3: Peer Assessment as a Learning Tool

Goal

Although the use as a control instrument might be considered as a support for learning too, the paradigm shift from a testing culture to an assessment culture (Birenbaum et al., 1996) showed us that assessment can do much more to support learning. Many scholars support the idea that peer assessment should also be considered as a tool for learning (e.g., Dochy & McDowell, 1997; Gielen et al., 2003). Many studies introduce peer assessment mainly for its beneficial impact on learning.

Before proceeding to the quality concept that is used in these studies, we first focus on the processes inherent in peer assessment that are able to initiate learning. Peer assessment includes the following processes: a student undergoing an assessment; a student assessing someone; and an interaction between peers that is the consequence of both. These three activities give rise to three sub-goals of peer assessment as a tool for learning.

Sub-Goal 1: Assessment for Learning

The first sub-goal is the well-known 'assessment for learning' (Taras, 2002), also referred to as learning-oriented assessment (Carless, Joughin, & Mok, 2006) or formative assessment (Black et al., 1998). In this use of peer assessment, the learning of the assessee is central.

This process is not unique to peer assessment: staff assessment may also be used as assessment for learning. Nevertheless, several mechanisms are described in the literature to explain why being assessed by a peer, or receiving feedback from a peer, may be particularly effective for learning. Peer feedback is often perceived as better understandable and more useful by students, because fellow students 'are on the same wavelength' (Topping, 2003), and share the same discourse (Hounsell, 1987). Moreover, it leaves room for discussion leading to deeper understanding (Topping, 2003). Furthermore, it can realise a gain in speed of return compared to staff feedback: "Imperfect feedback from a fellow student provided almost immediately may have much more impact than more perfect feedback from a

tutor four weeks later" (Gibbs et al., 2004, p. 19). The same trade-off works for the frequency or amount of feedback, which can also be increased when peers provide feedback instead of or complementary to staff. An extra advantage lies in the level of individualisation of feedback. If staff tries to provide more timely and more frequent feedback, they often choose to organise it collectively to keep it feasible, at the expense of a more personal guidance. Formative peer assessment can compensate this lack to a certain extent. A final argument in favour of peer feedback lies in the association of feedback with power issues, emotions and identity that may launch an 'emotion-defence system' in students (Higgins, 2000). As a consequence students may hide their weaknesses and doubts for the teacher, rendering teachers unaware of particular student difficulties or misconceptions. In that case, teacher feedback is less likely to connect to the learner, since it fails to address their problems or concerns. Peer feedback may by-pass some of these difficulties since it is less power-sensitive.

Sub-Goal 2: Assessing for Learning

A second sub-goal of peer assessment is to raise the learning of the *assessor* through the peer assessment activity: assessing for learning (e.g., Topping, 1998). Reasons for these learning effects are twofold. Firstly, students discover interesting ideas or alternative approaches to the task when reading others' work or observing others' performances, and will incorporate these in their own work. In addition, they will probably also detect some weaknesses or mistakes by the others that will stimulate self-reflection and probably lead to a correction of similar flaws in their own work. Secondly, Pryor and Lubisi (2002) mention that the assessment activity engages students to cognitively operate at an evaluative level and to pose metacognitive questions. These are higher order learning activities that help the assessor acquire a deeper insight into the subject. Sluijsmans and Prins (2006) also found that training student assessors in assessment skills has positive effects on their development of content related skills. An explanation for this is provide by Stiggins (1991, p. 38): "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".

Sub-Goal 3: Peer Learning

Finally, the assessment act in itself can even become a side issue, used to initiate an interaction between peers and to give rise to peer learning processes. In his review of peer learning, Topping (2005) explicitly names peer assessment as an extension of the forms of peer learning from the traditional peer tutoring and cooperative learning. Peer assessment may actually turn into a collaborative learning experience, especially when the assessor is expected to give formative and qualitative feedback and when assessor and assessee are encouraged to discuss differences in opinions and look for implications and solutions together. This feedback is more cognitively demanding of the assessor and more useful to the assessee than just marking each other's performance.

Quality Concept

The quality question in the case of peer assessment as a tool for learning differs from the previous requirements, and can be summarised by the concept of 'consequential validity' (Boud, 1995; Gielen et al., 2003; Saito et al., 2004). The quality criteria congruent with the third goal of peer assessment refer to the effects that performing and undergoing peer assessment, or giving and receiving peer feedback, have on the student and his learning.

The specific criteria to decide on the quality of peer assessment as a learning tool are diverse, since they are closely related to the learning goals themselves. If peer assessment is used in a mathematics course for instance, the appropriate criteria will express an improvement in the mathematical competence of a student. This improvement may take place on several levels: for example a correction of a certain misconception, a more fluent application of a heuristic, or a growing awareness of what aspects to include in an answer to a mathematical problem task (in fact, an awareness of the assessment criteria for a problem task). In another subject domain, these criteria will be different.

Depending on how the learning effects are measured, one even measures a different learning goal, Yorke (2003) argues. Assuming that students get a chance to revise their work or their performance after the peer assessment experience, one can measure the learning effect in two ways. On one hand, the improvement on the current assessment task can be measured. Learning from feedback by peers, learning by discovering alternative

approaches to handle learning tasks through the role of assessor, or learning from cooperative thinking in a peer group, are all means to reach a higher performance on a certain task than a student could attain alone. If task results improve after an intermediate peer assessment, one knows that students learned in the short term. Nevertheless, one cannot be sure that they will be able to perform at the same level if they were to handle an analogous task independently. The success may be in part attributable to the feedback that students receive on the drafts, or the ideas that they borrow from giving feedback to peers or discussing the assignment with peers. "In theoretical terms", Yorke (2003, p. 482) adds, "it cannot be said whether the student has moved his or her 'zone of proximal development' up the developmental gradient". So, a second way to measure learning effects is to provide an extra, independent measure to examine the quality of the learning without help from peers. Yorke (2003) refers to this type of effect as 'learning effects on the long term'.

Goal 4: Peer Assessment as a 'Learn-How-To-Assess-Tool'

Goal

Students learn to become assessors through peer assessment. This is learning on a meta-level, beyond the immediate learning gains from receiving feedback and assessing someone else's work. Learning-how-to-assess is an important part of becoming a lifelong learner, since students have to be able to undertake assessment of learning tasks they face throughout their lives (Boud, 2000). They have to learn how to define appropriate criteria, and to determine themselves whether or not they meet these. Moreover, they should learn how to seek feedback from their environment, when a teacher is no longer available.

The experience of being a peer assessor can be considered as a precursor to becoming a skilled self assessor (Sambell & McDowell, 1997). Sluijsmans (2002) distinguishes a first order course goal (content-related skills) from a higher order course goal (acquiring peer assessment skills). She uses peer assessment tasks, which are embedded in the study tasks of a

course, explicitly as tools to reach the higher order course goal of developing the peer assessment skills of students.

In traditional assessment, students may develop 'learned dependence', which does not go as far as 'learned helplessness' according to Yorke (2003), but which nevertheless discourages students from developing to their full potential, because students remain dependent on the teacher or the examiner to make decisions about what they know. Involving students in assessment may enable them to learn to recognise cues from the context of study which indicate what is good quality work and to develop criteria to assess a certain performance (Stefani, 1998). Experiencing the value of peer feedback may teach them to construct formative assessment processes for themselves in situ, using colleagues, peers and friends (Boud, 2000).

Quality concept

Whether or not students develop the above described abilities is the core question of quality for this fourth goal of peer assessment, and is also considered as an important element of its consequential validity (Boud, 1995).

No empirical study so far has addressed this effect on lifelong learning, after formal education. This will not be easy either, since the impact is not immediate but should be built in the long term through several experiences of being an assessor, and by avoiding an undermining impact of various other assessment experiences in the curriculum. The design of such a study will have to be longitudinal. Some studies (Gielen, Peeters, & Tops, 2007; Sluijsmans et al., 2006; Sluijsmans, Brand-Gruwel, & van Merriënboer, 2002; Sluijsmans, Brand-Gruwel, van Merriënboer, & Martens, 2004) try to get a glimpse of these future effects by monitoring peer assessment skills or metacognitive growth in the short term, during formal education. Nevertheless, studies of the long term effect on self-regulated learning are needed.

Goal 5: Peer Assessment as an Active Participation Tool

Goal

Finally, we distinguish a fifth goal that is quite different from the other goals. In the previous sections, the involvement of students in the assessment always served a higher goal: making sure that certain actions take place or others are avoided; delivering high quality assessment information (perhaps combined with a time gain for staff); creating learning gains; or developing competent lifelong learners. Engaging students in assessment is, in those instances, only a means to reach that end. In this fifth section, engaging students as active participants in their own learning and assessment is the goal itself. Magin et al. (2001) suggest that developing student autonomy and empowering students to make judgements that count are arguments that support the use of peer assessment, and even the use of it in a summative way (see also Langan et al., 2005). Peer assessments "can be viewed as vehicles for student empowerment" (Stanier, 1997, p. 95). Additionally, peer assessment can be considered as part of the self assessment process and serves to inform self assessment (Boud, 1986; Somervell, 1993).

Current assessment practices, Boud (2000) contends, too often "provide a mechanism of control exercised by those who are guardians of particular kinds of knowledge - teachers, educational institutions, professional bodies and occupational standards organisations - over those who are controlled by assessment - students, novices and junior employees" (p. 155). Peer assessment, together with self assessment, should be an aid in the liberation of the student, instead of serving as a new mechanism for oppression (Boud, 1994). The shift in responsibility for assessment from the teacher to the student leads to a greater democracy within the educational community (Searby & Ewers, 1997; Somervell, 1993).

Boud and Brew (1995) refer to the 'emancipatory knowledge interest' described by (Habermas, 1987) in their discussion of different ways in which self assessment may be used; this is also applicable to peer assessment. Involving students in assessment in an emancipatory way means that teachers and students do not discuss whether the student has met a set of

criteria, but that they develop the criteria and students' understanding of it together. Criteria and standards are not accepted as a given, but are the subject of a critique. This way of teaching truly values students' opinions and it takes a student-centred position in discussions and negotiations on the nature of assessment criteria (Stefani, 1998). Stefani adds: "Teaching staff can bring their expertise to bear on the management of student learning, but we need to show a willingness to share the learning goals and the assessment criteria in a meaningful way with our students. There is a strong need for academic staff to recognise differences in interpretation which come from the social, cultural and political diversity within groups of students. We need to move further from the rhetoric to the reality of student-centred teaching and learning" (p. 346).

Quality Concept

Concerning this last goal, the quality criterion may be whether one accomplishes the creation of a 'sense of shared ownership' (Sadler & Good, 2006) of the learning and assessment processes for each learner. This involves installing a new classroom culture in which the traditional asymmetrical relationship between teachers and learners is fading. In such a classroom, the current boundaries of knowledge can be crossed, and teachers and students can explore new undiscovered fields together. Knowledge is not held by the teacher, but is socially distributed and is constructed through interaction. However, as Stanier (1997, p. 95) states: "The nature of student empowerment associated with the use of these methods is difficult to monitor and, indeed, the benefits may be delayed". Qualitative research methods seem most appropriate to grasp this kind of change in learning environments as an indicator of 'quality'. However, no specific inquiry methods or criteria are found in the peer assessment literature yet.

Conclusion

This paper sought a pattern in the cluttered picture of quality concepts and quality criteria regarding peer assessment in higher education. It revealed that a discussion about the most appropriate quality criteria for the

output of peer assessment should be brought back to the underlying differences in goal formulation of peer assessment.

Peer assessment may be chosen for very different reasons. Five distinctive goals were identified in the literature. The five goals are sorted from more external control to more support of student's autonomy by peer assessment (see Figure 1 for a summary).

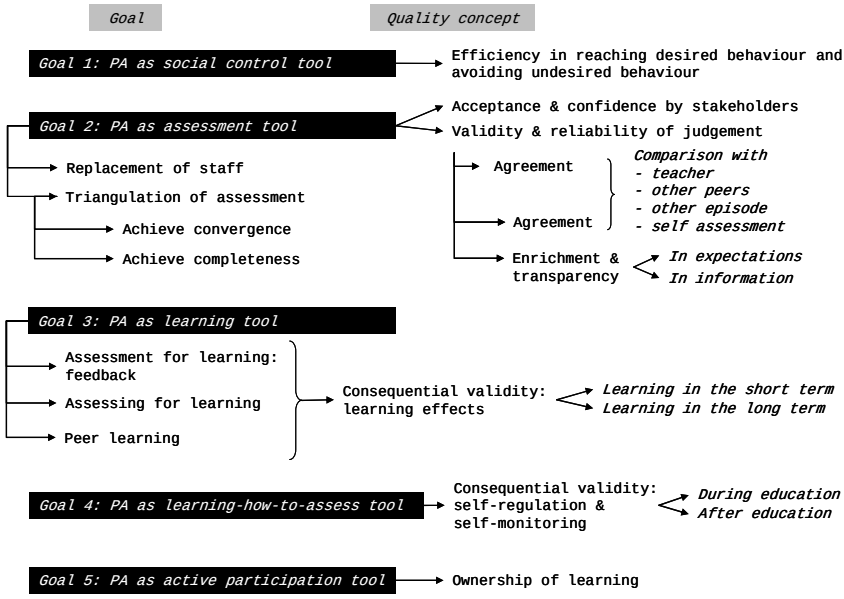


Figure 1. Overview of goals of PA and associated quality concepts.

The first goal of peer assessment is as a tool for social control. The fact that students know that a peer will assess them on a certain task or performance encourages them to work harder and perform better. When peer assessment is used for this purpose, quality is defined as how efficient it is in reaching the desired behaviour in students, and avoiding undesired behaviour.

The second and most well-known goal of peer assessment is its use as an assessment tool. Prerequisite, and thus the general quality criterion to achieve this end, is that the stakeholders have confidence in the results of this assessment and thus accept it. Furthermore, some specific quality criteria can be formulated for the quality of the judgements by peers. These judgements

have to be valid and reliable, but these concepts have different meanings depending on the sub-goal of peer assessment as an assessment tool. If peer assessment has to be able to replace staff assessment, agreement or concurrent validity is important. Also, when peers are used in a triangulation approach where convergence between assessors is sought, sub-goal 2 (agreement) is an important quality concept. Agreement, however, requires a reference for comparison. On this issue, studies disagree on what the most appropriate reference is: assessment by teachers or other peers, other assessment episodes by the same assessor or self assessment by the assessee. A third sub-goal of using peer assessment as an assessment tool is to use it to uncover multiple perspectives on the product or performance, to reach 'completeness validity'. In this case, the appropriate quality concept is 'enrichment and transparency'.

Thirdly, peer assessment can be used as a tool for learning. Three types of processes are able to produce or support this learning: learning by the assessee through assessment for learning and feedback; learning by the assessor through assessing for learning; and learning by both through peer learning processes. The quality of all three is expressed in the consequential validity concept that refers to the achieved learning effects, in the short and in the long term. Depending on the subject matter, different direct and indirect measures can be developed to measure these learning effects.

The fourth goal of peer assessment is to help students to learn how to assess themselves as lifelong learners. Peer assessment succeeds in this goal if students become independent learners who are able to self-regulate and self-monitor their learning in the learning society. Criteria to decide on the achievement of this success are not yet discussed explicitly in the literature. Direct and indirect measures may be possible.

And finally the fifth goal of peer assessment is most directly linked to autonomy support in the classroom. Peer assessment becomes a tool to realise active participation of students in their learning, and to create student-centred learning environments where teachers do not control what knowledge or good performance is. Quality, from this point of view, is conceptualised as the development of a 'sense of ownership' of the learning and assessment for each learner.

Quality concepts related to each goal or sub-goal are different, and a discussion about the appropriateness of a specific quality criterion is pointless

unless a reference is made to the goal that one is trying to achieve. Topping (1998) was right when he stated that "the role and function of teacher assessment might differ from that of peer assessment, so high reliability might not actually be necessary (for peer assessment)" (p. 257). However, this insight did not lead to a fully clear picture, since peer assessment may in some circumstances still fulfil the role of being an assessment tool, thus requiring a reliable judgement. On the other hand, teacher assessment may also serve, for instance, as a tool for learning, thus requiring quality criteria other than high reliability.

Each goal or sub-goal defines different expectations regarding the tool of peer assessment, so it is not likely that an 'ideal' version of peer assessment can be designed that can comply with all wishes at the same time. However, it is likely that teachers want to combine some of the goals of peer assessment. In that case, the challenge will be to find a delicate balance between the different expectations and quality concepts associated with these different goals. Sometimes, strategies to increase the quality of peer assessment from the point of view of one goal might act detrimentally on the quality in the light of another goal. Boud and Falchikov (2006) give the following example of these 'new assessment traps'. New strategies, such as providing students with criteria for assessment, which have positive effects on their current learning (goal 3), may have unintended longer-term consequences and counteract the fourth goal: they "portray to students the idea that the specification of standards and outcomes is a given and that learning only proceeds following such a specification by others. Yet in the learning that professionals do outside the academy, learning outcomes are rarely specified in explicit terms" (pp. 403-404).

On one hand, this overview should help researchers and practitioners to be more explicit about their goals of using peer assessment, and should clarify the relationship with appropriate quality criteria. On the other hand, this overview might inspire practitioners who have always used peer assessment for a certain goal to extend its goal or to replace it with another goal in their teaching practice.

References

Abson, D. (1994). The effects of peer evaluation on the behaviour of undergraduate students working in tutorless groups. In H. C. Foot, C. J. Howe, A. Anderson, A. K. Tolmie, & D. A. Warden (Eds.), *Group and interactive learning* (pp. 153-158). Southampton: Computational Mechanics.

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

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

Birenbaum, M. (2007). Evaluating the assessment: Sources of evidence for quality assurance. *Studies in Educational Evaluation, 33,* 29-49.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5,* 7-74.

Boud, D. (1986). *Implementing student self-assessment*. Sydney: Higher Education Research and Development Society of Australia.

Boud, D. (1994). The move to self-assessment: Liberation or a new mechanism for oppression? In P. Armstrong, B. Bright, & M. Zukas (Eds.), *Reflecting on Changing Practices, Contexts and Identities* (pp. 10-14). Leeds: Department of Adult Continuing Education, University of Leeds.

Boud, D. (1995). Assessment and learning: Contradictory or complementary? In P. Knight (Ed.), *Assessment for Learning in Higher Education* (pp. 35-48). London: Kogan Page.

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education, 22,* 151-167.

Boud, D. & Brew, A. (1995). Developing a typology for learner self assessment practices. *Research and Development in Higher Education, 18,* 130-135.

Boud, D. & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education, 31,* 399-413.

Breitmeyer, B. J., Ayres, L., & Knafl, K. A. (1993). Triangulation in qualitative research: evaluation of completeness and confirmation purposes. *IMAGE Journal of Nursing Scholarship, 25,* 237-243.

Carless, D., Joughin, G., & Mok, M. M. C. (2006). Learning-oriented assessment: principles and practice. *Assessment & Evaluation in Higher Education, 31,* 395-398.

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98,* 891-901.

Cole, D. (1991). Change in self-perceived competence as a function of peer and teacher evaluation. *Developmental Psychology, 27,* 682-688.

Conway, J. M. & Huffcutt, A. I. (1997). Psychometric properties of multi-source performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10,* 331-360.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale: Erlbaum.

De Corte, E. (1996). Instructional psychology: Overview. In E. De Corte & F. E. Weinert (Eds.), *International encyclopedia of developmental and instructional psychology* (pp. 33-43). Oxford: Elsevier Science.

De Corte, E. (2000). Marrying theory and the improvement of school practice: a permanent challenge for instructional psychology. *Learning and Instruction, 10,* 249-266.

Dierick, S. & Dochy, F. (2001). New lines in edumetrics: New forms of assessment lead to new assessment criteria. *Studies in Educational Evaluation, 27,* 307-329.

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

Falchikov, N. (1995). Improving feedback to and from students. In P. Knight (Ed.), *Assessment for Learning in Higher Education* (pp. 157-166). London: Kogan Page.

Falchikov, N. (1993). Group-process analysis - Self and peer assessment of working together in a group. *Educational & Training Technology International, 30,* 275-284.

Falchikov, N. & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70,* 287-322.

Gibbs, G. (1999). Using assessment strategically to change the way students learn. In S. Brown & A. Glasner (Eds.), *Assessment matters in Higher Education: Choosing and using diverse approaches* (pp. 41-53). Buckingham: SRHE & Open University Press.

Gibbs, G. & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education, 1,* 3-31.

Gielen, S., Dochy, F., & Dierick, S. (2003). Evaluating the consequential validity of new modes of assessment: The influence of assessment on learning, including pre-, post-, and true assessment effects. In M. Segers, F. Dochy, & E. Cascallar (Eds.), *Optimizing new modes of assessment: In search of qualities and standards* (pp. 37-54). Dordrecht: Kluwer Academic.

Gielen, S., Peeters, E., & Tops, L. (in preparation). The impact of a peer review experience on the use of self-regulation strategies.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. London: Falmer.

Habermas, J. (1987). *Knowledge and human interests*. (Translated by Shapiro, J. ed.) London: Polity Press.

Higgins, R. (2000). "Be more critical!": Rethinking assessment feedback. In *Paper presented at the British Educational Research Association Conference.* Cardiff University.

Hounsell, D. (1987). Essay writing and the quality of feedback. In J. Richardson, M. W. Eysenck, & D. W. Piper (Eds.), *Student Learning: research in education and cognitive psychology* (Milton Keynes: Open University Press.

Johnson, J., Olson, A., & Courtney, C. (1996). Implementing multiple perspective feedback: An integrated framework. *Human resource management review, 6,* 253-277.

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

Langan, A. M., Wheater, C. P., Shaw, E. M., Haines, B. J., Cullen, W. R., Boyle, J. C. et al. (2005). Peer assessment of oral presentations: effects of student gender, university affiliation and participation in the development of assessment criteria. *Assessment & Evaluation in Higher Education, 30,* 21-34.

Liu, C.-C. & Tsai, C.-M. (2005). Peer assessment through web-based knowledge acquisition: tools to support conceptual awareness. *Innovations in Education & Teaching International, 42,* 43-59.

Magin, D. & Helmore, P. (2001). Peer and teacher assessments of oral presentation skills: How reliable are they? *Studies in Higher Education, 26,* 287-298.

Marcoulides, G. A. & Simkin, M. G. (1995). The consistency of peer review in student writing projects. *Journal of Education for Business, 70,* 220-224.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3 ed., pp. 13-103). New York: American Council on Education/ Macmillan.

Miller, P. (2003). The effect of scoring criteria specificity on peer and self-assessment. *Assessment and Evaluation in Higher Education, 28,* 383-394.

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

Pope, N. (2001). An examination of the use of peer rating for formative assessment in the context of the theory of consumption values. *Assessment & Evaluation in Higher Education, 26,* 235-246.

Pope, N. K. L. (2005). The impact of stress in self- and peer assessment. *Assessment & Evaluation in Higher Education, 30,* 51-63.

Pryor, J. & Lubisi, C. (2002). Reconceptualising educational assessment in South Africa - testing times for teachers. *International Journal of Educational Development, 22,* 673-686.

Rada, R. & Hu, K. (2002). Patterns in student-student commenting. *IEEE Transactions on Education, 45,* 262-267.

Robinson, J. M. (2002). In search of fairness: An application of multi-reviewer anonymous peer review in a large class. *Journal of Further and Higher Education, 26,* 183-192.

Sadler, P. & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11,* 1-31.

Saito, H. & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research, 8,* 31-54.

Sambell, K. & McDowell, L. (1997). 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: Oxford Centre for Staff and Learning Development.

Searby, M. & Ewers, T. (1997). An evaluation of the use of peer assessment in higher education: A case study in the School of Music, Kingston University. *Assessment & Evaluation in Higher Education, 22,* 371-383.

Segers, M., Dochy, F., & Cascallar, E. (2003). *Optimizing New Modes of Assessment: In Search of Qualities and Standards*. Dordrecht: Kluwer Academic.

Segers, M. & Dochy, F. (2001). New assessment forms in problem-based learning: The value-added of the students' perspective. *Studies in Higher Education, 26,* 327-343.

Sluijsmans, D. (2002). *Student involvement in assessment. The training of peer assessment skills*. Unpublished doctoral dissertation. Open Universiteit Nederland, Heerlen.

Sluijsmans, D. & Prins, F. (2006). A conceptual framework for integrating peer assessment in teacher education. *Studies in Educational Evaluation, 32,* 6-22.

Sluijsmans, D. M. A., Brand-Gruwel, S., & van Merriënboer, J. J. G. (2002). Peer assessment training in teacher education: Effects on performance and perceptions. *Assessment & Evaluation in Higher Education, 27,* 443-454.

Sluijsmans, D. M. A., Brand-Gruwel, S., van Merriënboer, J. J. G., & Martens, R. L. (2004). Training teachers in peer-assessment skills: effects on performance and perceptions. *Innovations in Education & Teaching International, 41,* 60-78.

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

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

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

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

Taras, M. (2002). Using assessment for learning and learning from Assessment. *Assessment & Evaluation in Higher Education, 27,* 501-510.

Topping, K. J. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy, & E.

Cascallar (Eds.), *Optimizing new modes of assessment: In search of qualities and standards* (pp. 55-87). Dordrecht: Kluwer Academic.

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment & Evaluation in Higher Education, 25,* 149-169.

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

Topping, K. J. (2005). Trends in peer learning. *Educational Psychology, 25,* 631-645.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17,* 31-74.

Yorke, M. (2003). Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher Education, 45,* 477-501.

CHAPTER 4

AN INVENTORY OF PEER ASSESSMENT DIVERSITY

Abstract

In 1998 Topping stated: “The next decade should bring a major expansion in the peer assessment literature. Ten years later Topping has proved to be right about the expansion of peer assessment research, and the use of peer assessment in practice: since the review of Topping in 1998, the number of studies on this subject has doubled, if not tripled. Together with this expansion, however, also the diversity of peer assessment practices has increased exponentially. Although all these practices are members of the same family, they differ in certain aspects at least as much as they are similar in other aspects. This diversity poses difficulties for practitioners as well as researchers. In the current paper an inventory of peer assessment diversity has been developed that may be of interest to practitioners, as a checklist of important decisions to take or an overview of possible alternatives to a specific practice, and to researchers as a guideline of what information to provide on the particularities of their peer assessment design. Finally the framework that has been developed in this paper may help to clarify the confusion that originates from the use of a single term to cover a multitude of sometimes incompatible practices. Based on a review of the recent literature, the inventory of peer assessment diversity provides an update of Topping’s typology. Eight new variables were added and another eight variables were extended with extra subdimensions. Five original variables of Topping were absorbed in larger entities, and also the implementation factors of Topping were given a place within the variables of the inventory. Finally, the 20 resulting variables were grouped into five clusters, building on an earlier clustering by van den Berg, Admiraal and Pilot (2006b).

AN INVENTORY OF PEER ASSESSMENT DIVERSITY

The Need for an Inventory of Peer Assessment Diversity

Peer assessment is "an arrangement in which individuals consider the amount, level, value, worth, quality or success of the products or outcomes of learning of peers of similar status" (Topping, 1998, p. 250). However, the definition of peer assessment does not describe one unifying design of peer assessment. A diversity of peer assessment practices is described in the literature. Although all these practices are members of the same family, they differ in certain aspects at least as much as they are similar in other aspects.

This diversity poses difficulties for practitioners as well as researchers. Practitioners who took the decision to introduce peer assessment in their teaching, do not know where to start in designing of their peer assessment practice. A checklist of important decisions to make can support this design process.

Secondly, practitioners who work already with peer assessment, might not be aware of a lot of implicit decisions in the design of their peer assessment practice, and of the possible alternatives for several choices. They might not experience this as a problem, until they notice that their peer assessment design is not perfectly working as they intended, and they do not know where to start looking for an explanation or a solution. An overview of the important variables in assessment that might influence its effectiveness would be helpful for these practitioners.

A third problem is one faced by researchers who want to compare different peer assessment practices reported in the literature: some reports mention some aspects of the peer assessment design, and other reports mention others. As a consequence, there is only a limited overlap in the type of information provided in the individual studies. This makes it difficult to combine studies in a critical review, a best-evidence synthesis, or a meta-analysis. To enable a comparison between different peer assessment studies,

and in the end a synthesis of their results, all individual studies should address more or less the same topics of a peer assessment design. For the researchers of these individual studies, it would therefore be helpful to have a guideline of the characteristics of a peer assessment design that they should describe in their report.

Finally, a fourth problem is relevant to all those who talk or write about peer assessment: when a single term covers a multitude of sometimes incompatible practices, confusion may rise. A framework that indicates where to look for similarities and differences between two instances carrying the same name, helps to clarify or avoid this confusion.

In the current paper an inventory of peer assessment diversity will be developed that can function as a checklist, an overview, a guideline or a framework as described above. We will, however, not start from scratch. In 1998, Topping already delivered a significant contribution to the field of peer assessment by developing a typology and a list of implementation factors that provided an index of the variables on which peer assessment applications vary. However, when designing peer assessment applications for our empirical studies (Gielen, Dochy, Onghena, Struyven, Smeets, & Decuyper, 2007; Gielen, Dochy, Onghena, & Smeets, 2007), we experienced several lacunas in the typology. This was the impetus to start a literature review in search of the important variables that should be included in a design checklist; that are important for the effectiveness; that differentiate between several peer assessment designs; and that help to provide a clear description of a specific practice. Before proceeding to the methodology and the results of the literature review, first the existing typology of Topping will be discussed.

Topping's Typology of Peer Assessment

Topping (1998) immersed himself in the literature on "peer assessment, peer marking, peer correction, peer rating, peer feedback, peer review and peer appraisal" between 1980 and 1996 and summarized some of the main variables that explain the variation between peer assessment projects reported in the literature. Each variable has several possible values, and each combination of values on the different variables results in a different 'type' of

peer assessment. That is why his scheme may be called a 'typology'. Recently, van den Berg et al. (2006b) published a paper on course designs for peer assessment, in which they proposed a grouping of Topping's variables into four clusters. An extra column is added to the typology to represent these clusters.

Table 1

Typology of peer assessment (adapted from Topping, 1998b, p. 252).

Cluster (van den Berg, Admiraal, & Pilot, 2006b)	**Variable**	**Range of Variation**
Cluster I The function of PA as an assessment instrument	1) Curriculum area/subject	All
	2) Objectives	Of staff and/or students? Time saving or cognitive/affective gains?
	3) Focus	Quantitative/summative or qualitative/formative or both?
	4) Product/Output	Tests/marks/grades or writing or oral presentations or other skilled behaviours?
	5) Relation to staff assessment	Substitutional or supplementary?
	6) Official weight	Contributing to assessee final official grade or not?
Cluster II Interaction between peers	7) Directionality	One-way, reciprocal, mutual?
	8) Privacy	Anonymous/confidential/public?
	9) Contact	Distance or face to face?
Cluster III Composition of the feedback group	10) Year	Same or cross year of study?
	11) Ability	Same or cross ability?
	12) Constellation Assessors	Individuals or pairs or groups?
	13) Constellation Assessed	Individuals or pairs or groups?
	14) Place	In/out of class?
	15) Time	Class time/free time/informally?
Cluster IV Requirement & reward	16) Requirement	Compulsory or voluntary for assessors/ees?
	17) Reward	Course credit or other incentives or reinforcement for participation?

In addition to this typology, Topping (1998) also listed some general organizational factors that should be taken into account when implementing peer assessment. What the exact organisational arrangements are varies according to the type of peer assessment that is deployed, particularly the type of object that is the subject of peer assessment. Therefore, these factors do not have a column 'range of variation'. However, for each of the factors one can check whether or not they are present and if so, how they are handled (see Table 2).

Table 2
Implementation factors (adapted from Topping, 1998, pp. 265-267).

Implementation factors
- Clarifying expectations, objectives and acceptability
- Matching participants and arranging contact
- Developing and clarifying assessment criteria
- Providing quality training
- Specifying activities
- Monitoring the process and coaching
- Moderating reliability and validity
- Evaluating and providing feedback.

Since Topping's literature search on peer assessment, that collected all relevant papers between 1980 and 1996, the number of studies of peer assessment research has increased fast. A combined search of the major bibliographic databases (see Table 3) in humanities with the search string *(("peer-assessment" OR "peer assessment") AND "education") in all fields* learns that at least two thirds of the studies published on peer assessment (since the beginning of these databases in the seventies) were published between 1997 and the end of 2006. Since peer assessment research has been published under a wide variety of descriptors, we repeated this search with the search string of Topping (*"peer assessment" OR "peer review" OR "peer rating" OR "peer feedback" OR "peer marking" OR "peer correction" OR "peer appraisal") AND "education"*. Depending on the database, we searched in the field of *keywords, subject or abstract* (see Table 3). The addition of some 'older' terminology on peer assessment decreases

the proportion of recent literature somewhat, but still at least 40% was published after Topping's review study. These data clearly suggest that there is a need to check Topping's typology study against the recent literature, and to update it if necessary.

Table 3

Proportion of recent references in the major relevant bibliographic databases (10/02/2007).

Database	Date first PA[1]-ref	Total # PA-ref	# PA-ref 1997-2006	Date first P*[2]-ref	Total # P*-ref	# P*-ref 1997-2006
ERIC (CSA) (in KW)	1970	195	130 (66%)	1954	1196	467 (39%)
SSCI (in TS)	1977	36	28 (78%)	1971	174	121 (70%)
Acad. Search Premier (in AB, restr. acad. journals)	1984	181	166 (92%)	1970[3]	483[3]	402[3] (83%)
PsycINFO (in AB)	1970	90	82 (91%)	1952	257	159 (62%)

[1] PA= peer assessment
[2] P*= peer assessment, peer review, peer rating, peer feedback, peer marking, peer correction or peer appraisal
[3] Search restricted to 'academic journals'

Methodology

Based on our literature search with the extended search string as described above, those studies, published between 1997 and 2006, that addressed peer assessment by learners (in formal and non-formal education) were selected, removing the studies about peer assessment among scholars (for scientific publication); employees (for personnel evaluation); institutions (for quality assurance and accountability) or friends (for sociometrics). Since the number of duplicates between the databases was high, one database (Academic search premier) was taken as a starting point and later complemented with the non-duplicates from the other databases. Furthermore, since authors in our sample often referred to specific peer assessment designs that were published before 1997, those sources were consulted to (the so-called snowball method).

In line with the purpose of this study, the collected literature was explored in search of variables that were necessary to describe a specific peer assessment practice. Almost all sections of a publication could contain useful information for our review. We focused on the description of the peer assessment designs in the procedure sections; manipulated variables in method sections of (quasi-)experimental studies; reported adaptations in action research studies; and finally aspects of the design mentioned as possible explanations for positive or negative results, or as suggestions for future applications in discussion sections of publications. In a cumulative and iterative process, each additional variable, dimension within a variable or value within a dimension was indexed. To justify a new variable or dimension in the inventory, a reference to an exemplary publication, in which this variable or dimension is mentioned, will be added in the results section. For our research goal, it was not necessary to classify the studies, or to count certain occurrences. It is not our aim to give an overview of all studies that mention a certain variable, since that would not contribute in any way to our goal of developing a useful checklist, overview, guideline or framework for practitioners and researchers in the field of peer assessment.

At the end of the search and indexing process, variables were ordered within a few higher order clusters, thereby extending the original clustering by van den Berg et al. (2006b), in order to make the inventory more practical in use by reducing the number of main themes.

We chose to abandon the term 'typology' for this classification framework. In the original typology, variables are discrete and each variable has a list of 'multiple choice' options associated with it, which makes it – in theory – possible to define a certain number of 'types' of peer assessment by crossing the different variables. However, in our indexation process of variables the range of variation appeared no longer to be a list of separate values but is a continuum, or sometimes even a multi-dimensional space, with some exemplary values listed for clarification. We therefore suggest calling the result 'an inventory of diversity of peer assessment'.

Results

In this section all components of the inventory of peer assessment diversity will be described. The order of the variables does not indicate their importance: variables are ordered to allow for a meaningful clustering, but within a cluster the sequence of variables is random. The headings of the variables contain a comparison to the typology of Topping, represented by a set of prefixes and symbols. The meaning of these indicators is presented in Table 4.

Table 4

Overview of the indicators used in the results section to compare the inventory of peer assessment diversity to the typology of Topping (1998)

Indicator	Meaning
New: (new label)	This variable is added
Extended: (old label)	This variable is widened with extra dimensions
(old label) → (new label)	This variable is widened (or narrowed) to such an extent that the original label was not applicable anymore
(old label) < (other label)	This variable has disappeared, since it is absorbed by another variable
(old label)	This variable is the same as in Topping's typology

Cluster I: Decisions Concerning the Use of Peer Assessment

Originally, the first cluster of van den Berg et al. (2006b), "the function of PA as an assessment instrument", referred to the goal of peer assessment as an assessment tool. After the revision, this cluster has expanded. The cluster will now not only contain information on "the function" of peer assessment, however, but also on several other basic decisions or entry data. We suggest changing the name of the first cluster to *'Decisions concerning the use of peer assessment'*. Several contextual variables with background information are grouped in this cluster.

Curriculum area/subject → Setting. The variable curriculum area/subject is found to be too narrow to cover all differences in settings that may influence the success of peer assessment. It is considered to be one dimension of the 'Setting' variable. Some important dimensions are added, such as 'educational or non-educational use', 'formal or informal learning',

'level of education', 'characteristics of participants', and 'class size' (e.g., Ballantyne, Hughes, & Mylonas, 2002; Falchikov et al., 2000).

Product/Output → Object. The products of the assessment (the type of tasks or performances that are assessed), and the output of the assessment (marks/grades = quantitative vs. open-ended = qualitative) are entangled in Topping's typology. Moreover, it is confusing to talk about 'products' if not all the objects of peer assessment are products (e.g. behaviour, processes). In the full description Topping makes a distinction between objects that are subject to scoring, marks and grades, being one type of output, and those to which detailed open-ended assessment and feedback are more frequently applied: in fact a second type of output. In the inventory "Objects" is separated from "Output", since in principle all combinations are possible. The latter will be considered to be a separate variable.

Possible objects of peer assessment are artefacts on one hand and observed behaviour on the other hand. Examples of artefacts are answers to a test (Sadler et al., 2006), products such as writing (e.g., Venables & Summit, 2003), posters (e.g., Orsmond, Merry, & Reiling, 2002; Smith, Cooper, & Lancaster, 2002), presentations (e.g., Langan et al., 2005), or reports of individual or group work projects (e.g., Malcolmson & Shaw, 2005; Prins, Sluijsmans, Kirschner, & Strijbos, 2005). Examples of observed behaviour are what Topping calls "other skilled professional behaviors" (e.g., in medicine, Norcini, 2003; or in music, Searby & Ewers, 1997). These behaviours can be observed in real life or from a videotape (in fact, rendering it a type of artefact) (Prins, Sluijsmans, & Kirschner, 2006; Trahasch, 2004).

Although it may be hidden in Topping's group of 'other professional behaviors', the inventory explicitly adds 'group work skills' such as cooperative skills, contribution to the group product, communication skills and social skills to the list of objects of peer assessment that are not artefacts. These types of 'process skills' are often subjected to peer assessment (e.g., Cheng & Warren, 2000; Segers et al., 2001). Sivan (2000) refers to this last type as 'intra-group assessment', in contrast to 'inter-group assessment' when class members assess the product of a certain group.

Concerning the object, the formal description of the object (such as Topping's examples) is extended with a deeper description of the content of these questions, tasks or observation settings. What type of performance is expected of students? For instance: reproduction of knowledge; personal

construction of knowledge; application; critical thinking; or self-reflection. And what information is taken into account? Should only the outcome (final results, answers or solutions) be considered, or also the ways they were achieved? In the study of Sadler et al. (2006) for example, some of the test questions that are peer assessed are multiple choice items measuring reproduction of factual knowledge.

Finally, an important addition, in the case of a formative use of peer assessment, is whether the object is still a draft version or is the final version (van den Berg et al., 2006b). This is important because the first type indicates that assessees still have the opportunity to revise their work or behaviour before the final assessment, thus applying the formative feedback to their current performance (e.g., Saito et al., 2004). In terms of Sadler (1989), also discussed by Boud (2000), this refers to the question of whether or not the feedback loop is closed. Yorke (2003) refers to the same issue in his discussion of the difference between short term and long term learning.

New: Frequency & Experience. Frequency & Experience is a new variable. It questions whether peer assessment takes place only once or sporadically during a course or curriculum, or more frequently, and what the amount of previous experience is that students already have with peer assessment (e.g., Cheng & Warren, 1999; Oldfield & MacAlpine, 1995). This is important in order to understand how familiar students are with peer assessment (Sluijsmans et al., 2006), but also to anticipate to the risk of students resenting having to participate in peer assessment too often (Ballantyne et al., 2002). Falchikov et al. (2000) point to the necessity for more research on the effects of repeated experience of peer assessment.

Extended: Objectives (goal of peer assessment). It is unclear to which dimension the question "Of staff and/or students?" in the original typology refers, and it is not discussed in Topping's description either. Therefore, it is left out. The dimension "Time saving or cognitive/affective gains?" is replaced by a more comprehensive list of potential goals that peer assessment may serve. Peer assessment may be used as a tool for social control (e.g., Gibbs, 1999), for assessment (e.g., Norcini, 2003; Robinson, 2002; Segers et al., 2001), for learning (e.g., Purchase, 2000; Venables et al., 2003), for learning-how-to-assess (e.g., Bloxham et al., 2004) or for active participation (e.g., Stefani, 1998), or any combination of these. These five

goals of peer assessment are discussed more profoundly in Gielen, Dochy, Onghena, Struyven, Smeets, and Decuyper (2007).

Time saving is an effect that may be reached for several of these goals. As Sluijsmans (2002) notes, time saving is more of a side effect, that may or may not happen, than a goal of peer assessment.

Focus → Function. Although not in the full description of the variable 'Focus', Topping associates quantitative with summative and qualitative with formative in his summary table of the typology. The difference between quantitative and qualitative is unjustly associated with summative and formative use of peer assessment, since a summative assessment may also comprise open-ended comments and, on the other hand, formative assessment may also be restricted to a quantitative appreciation. The quantitative/qualitative dimension is removed from the Function variable and replaced by the new 'Output' variable. Thus, function only refers to summative/formative in our inventory. In the meantime, however, it absorbs also another variable of Topping's original typology, namely 'Official weight'. The question "Contributing to assessee final grade or not?" is subordinate to the use of peer assessment as a summative or formative assessment.

Official weight < Function. Official weight (Contributing to assessee final official grade or not?) is a variable that disappears in the new inventory, since it is absorbed in the Function variable.

Cluster II: Link between Peer Assessment and Other Elements in the Learning Environment

'Relationship to other assessments', together with the two new variables 'Alignment' and 'Scope', deal with the *'Link between peer assessment and other elements in the learning environment'*. This is a new cluster that was not recognised by van den Berg et al. (2006b).

New: Alignment. This variable is added, and it concerns the object and use of peer assessment from a broader perspective. It examines their degree of alignment with curriculum, learning goals and teaching. This refers to the extent and the ways in which a peer assessment application really 'fits' into its learning environment, and is not an artificial add-on. The dimensions

that might be considered within this variable are numerous and diverse, since they depend on which other components are present in the learning environment, what the learning goals are, and what choices are made on the other variables of the inventory.

We provide some examples: If peer assessment is used as a tool for learning, it is important that the object of peer assessment is aligned with the learning goals and that the output is integrated within the teaching. For instance van den Berg et al. (2006b) discusses several designs of peer assessment that differ – amongst other aspects – in whether or not a peer assessment exercise is followed by a teacher-led plenary discussion of themes brought in by the feedback groups. If teachers do not give enough weight to the importance of peer feedback, students might consider it as inferior to other assignments.

A second example: if peer assessment is used as a summative assessment tool, it should be aligned to the official assessment criteria, and it should be treated with the appropriate care that is required of a summative assessment. Peer assessment of social skills of each group member in a group project, for instance, may only be justified if the acquisition of social skills is treated as a real learning goal in the learning environment (e.g., Schelfhout, Dochy, & Janssens, 2004).

Furthermore, more practical arrangements such as output, privacy and contact should also fit into the current setting (e.g., feasibility of timing, of workload). An example is mentioned in the discussion of Prins et al. (2005), namely the alignment with other assignments in a course, and their workload. Their course contained several assessment assignments that needed substantial investments of time and effort while the content-related assignments also happened to be very time consuming. The authors suggest that "the ratio between time available for the course and time needed for the assessment has to be guarded" (p. 436).

The 'Alignment' variable is certainly not the most easy variable to discuss in the context of a peer assessment design, but nevertheless it is a crucial one.

Extended: Relationship to other assessments. The most obvious group of other assessors are, of course, the teachers. The relationship between peer assessment and staff assessment may be (partially) substitutional (e.g., Sitthiworachart & Joy, 2003) or supplementary (e.g., Oldfield et al., 1995).

Substitutional means that students become 'surrogate' teachers; or 'assistant' teachers in case of a partially substitutional or complementary use (e.g., Cheng et al., 2000). Supplementary refers to peers and staff both assessing the same performance (e.g., Orsmond, Merry, & Reiling, 1996).

Somewhere between these positions are studies where teachers offer to re-mark assessment pieces in cases where there is a large discrepancy between peer marks or between self-assessment and peer assessment (e.g., Ballantyne et al., 2002).

An additional issue is raised when peer assessment is supplementary to staff assessment: namely does peer assessment take place before, simultaneously with or after staff assessment, and are they aware of each other's judgements? For instance, van den Berg et al. (2006b) reports: "the teacher's strategy was to give his comment only after peer feedback had been given" (p. 22).

Finally, other types of assessment, such as self assessment or external assessment, may be present in the course too. The relationship of these assessments to peer assessment might also be important to discuss (Ballantyne et al., 2002; McGourty, 2000).

New: Scope of involvement. Scope of involvement is a new, and important, variable in the typology. Peer assessment refers to the involvement of peers in assessment, but there is a large variance in the extent to which peers are involved. Peer assessment may include:

- the involvement of students in the definition of desired learning outcomes (course objectives, see for instance Sluijsmans et al., 2006);
- and/or the design of assessment tasks (e.g., Ballantyne et al., 2002);
- and/or the development of assessment criteria and standards (e.g., Orsmond et al., 2000; Smith et al., 2002), also mentioned as a factor of implementation quality by Topping;
- and/or the development of assessment procedures (e.g., Ballantyne et al., 2002);
- and/or the judgements in terms of grading/marking/commenting (e.g., Falchikov, 1995);
- and/or decision taking (e.g., Cheng et al., 2000);
- and/or the providing of knowledge of results/feedback to a peer (e.g., Falchikov, 1995);

- and/or the monitoring/guiding of a peer's progress (no studies found).

The judgement itself is the most obvious aspect that is shared with students. Without this, the term 'peer assessment' would probably be inappropriately used. But some applications go a lot further.

For all these activities that are part of the assessment process, involvement may also differ in degree: students are 'informed' about it, it is 'discussed' with them, they 'participate' in it, or it is 'their responsibility' (see also Topping's first factor of implementation quality). The term co-assessment is often used for designs where students collaborate in the different steps but do not take full responsibility for it (Somervell, 1993).

Cluster III: Interaction between Peers

The second cluster by van den Berg et al. (2006b) is extended with two new variables that also determine the type of interaction that takes place between assessor and assessee: the type of output that is requested from peer assessment and the role of the assessee.

New: Output. As explained above (see 'Object'), this variable refers to the type of information that is the product of an assessment by peers. This may be: a pass/fail message; a ranking; a grade or mark; a score profile; a diagnosis of strengths and weaknesses; a suggestion for remedial actions; a personal interpretation; reflective questions; an offer for help; or, for instance, an interactive dialogue between learners (e.g., Falchikov et al., 2000; Purchase, 2000; Topping, Smith, Swanson, & Elliot, 2000).

These possibilities differ in the nature of the information (quantitative and/or qualitative), the extent to which the information is condensed (holistic, global or at the level of single criteria) (e.g., Miller, 2003; Pope, 2005; Trahasch, 2004), and finally the 'feedback stance' that is taken (authoritative, interpretive, probing or collaborative, see Lockhart and Ng, 1995; and van den Berg, Admiraal and Pilot, 2006a).

Directionality. Topping defines three values for directionality: unidirectional (from assessor to assessee but not the reverse) (e.g., Sitthiworachart et al., 2003); reciprocal (assessment of each other between two people or two groups) (e.g., Topping et al., 2000); or mutual (assessment of each other between more than two people or groups) (e.g., Pâquet & Des

Marchais, 1998). This variable is closely linked to the 'Constellation of assessors and assessees' variable as well.

Extended: Privacy. The variable Privacy, together with Contact, concerns the modalities in which the assessment takes place and the output is communicated. Does the assessor know who the assessee is, and does the assessee know by whom he is being assessed (dimension of anonymity) (e.g., Saito et al. (2004) use a double-blind procedure)? Is the output communicated in a confidential way (e.g., Pâquet et al., 1998) between assessor and assessee? Is the teacher present at the assessment, or does he or she have access to the assessment results (e.g., in some courses in the study of van den Berg et al. (2006b) the teacher participates in the feedback groups)? Is the output publicly reported in the presence of others (e.g., the oral feedback given immediately after a presentation in class in the study of Falchikov (1995) or the feedback sessions on students' reflection papers in Sluijsmans, Brand-Gruwel, van Merriënboer, and Bastiaens (2002))?

Extended: Contact. The variable Contact is an additional description of the modalities in which the assessment takes place and the output is communicated. Does the assessment take place in the presence of the assessee (e.g., Ballantyne et al., 2002; Purchase, 2000), or at distance (e.g., Sitthiworachart et al., 2003)? And is the output communicated face to face in a conversation (Zhang, 1995), in an online discussion in an electronic learning environment – synchronously or asynchronously – (Trahasch, 2004), or is it provided in writing without any direct interaction (Prins et al., 2005)? The latter may be on paper or web-based. Does it happen 'one-way' or interactively? Combinations of these modalities are also possible; for instance a written preparation of peer feedback that is later orally explained to the assessee (van den Berg, Admiraal, & Pilot, 2006a).

New developments in information and communication technology enable students to have contact with each other in a virtual environment, and this of course also has an impact on the 'Place' and 'Time' variable. Due to a considerable overlap between the 'Contact' variable and Topping's 'Time' and 'Place' variables, they are combined in the inventory. The traditional two possibilities for place, namely in or out of class, are nowadays expanded to an intermediate place: the electronic learning environment. This is an environment that can be reached at any time and from any location (if a

computer is available), but it still is under a certain degree of control by staff. To avoid losing 'contact time' for peer assessment, but still keeping an eye on what is happening (e.g., to make sure peer assessment actually happens), peer assessment is moved to internet-based learning environments in many cases (e.g., the SWORD software by Cho, Schunn and Wilson, 2006).

Place < Contact. Place ("In or out of class?") is a variable that disappears in the new inventory, since it is absorbed in the Contact variable.

Time < Contact. Time ("Class time/free time/informally?") also disappears as a variable in the new inventory, since it too is absorbed in the Contact variable.

New: Role of assessee. This is an extra variable that deals with the way students receive their assessment or feedback. Peer assessment may be considered as something the assessee is subjected to, and passively undergoes. The assessee may, on the other hand, also be assigned an active role, rendering him or her partially responsible for the output or result of a peer assessment experience.

Students may, for instance, only be formatively assessed on request, or be expected to indicate themselves the aspects on which they would like to receive feedback (e.g., Nicol et al., 2006), or to communicate their preference for a certain style of feedback (Prins et al., 2006).

Different opinions concerning the desired response to the assessment or feedback are also found in the literature. Some ask assessees to absorb the feedback and reflect on it without immediate reaction. For instance, when the output of the peer assessment is merely quantitative, often no reaction from the assessee is expected (e.g., Sitthiworachart et al., 2003). Others expect students to actively engage in the interaction: to ask questions for clarification; to discuss differences in opinion; to search for improvements together, etc. (van den Berg et al., 2006b).

When considering the 'Object' variable, we have already discussed the difference between feedback on a draft or on a final version. This difference is also related to the role of the assessee: in the first case the assessee is expected to revise his work or performance based on the assessment (see two-stage assignments, discussed by Nicol et al., 2006); in the second he is not (Trahasch, 2004). Additionally, the assessee might be

asked to write a reply to the assessor indicating how the object of assessment was revised in view of the results of the assessment (e.g., Prins et al., 2005).

Cluster IV: Composition of Assessment Groups

In our opinion, the 'assessment groups' are inappropriately restricted to 'feedback groups' in van den Berg's framework. Peer feedback is just one type of peer assessment.

Year & Ability → Matching. The assignment of students to assessment groups in the seven designs by van den Berg et al. (2006b) differs in a way that is not represented in the variables of the original typology: in some courses the teacher assigns groups by putting together students with a related subject; in other courses, students were grouped at random. Another possibility would have been that students could choose their own groups (with or without certain constraints of subject similarity etc.) (e.g., Ballantyne et al., 2002; Strachan & Wilcox, 1996). Hence, a variable is created that refers to the way students are matched from a broader perspective than just the similarity (or dissimilarity) in year or ability. The central question is: on the basis of what principles does matching take place, and by whom?

In his 'implementation factors', Topping has already described another principle, beyond year and ability, for matching: the "social constellation" of peer assessment. "Students might be matched with peer assessors whom they found credible or with whom they were already friends, or simply by random allocation" (p. 266).

A final dimension is related to the consistency of a specific matching. If students will repeatedly perform peer assessments, will the matching of assessors to assessees remain fixed, or is it variable (Trahasch, 2004)? For instance in Sluijsmans et al. (2002), the 'who assesses whom' scheme altered after each course.

Extended: Constellation assessors & assessed. Both of Topping's constellation variables are taken together because all options apply to both sides of the relationship. Moreover, different dimensions of a constellation are distinguished to allow a more precise description of what is going on in a specific classroom.

In the first place, the unit of what counts as an assessor or an assessee may differ. For example, if a group project is being assessed by

another group, the unit of both the assessor and assessee is the 'group', instead of an individual or pair. It is also possible that the group of assessors all assess the group project of a fellow group individually (e.g., Prins et al., 2005): in that case the unit of assessor is the individual and the unit of assessee remains the group. In the study of Bloxham et al. (2004) on the other hand, the unit of assessor is a pair. A special situation is encountered in the studies of Sadler et al. (2006) and Sitthiworachart et al. (2003): peers assess individually, but get the opportunity to discuss their doubts with other peers in class.

It is, however, not only the units that may differ in extent. The number of assessors that are assigned to each unit of assessee may also vary. To stay with the example of group work, it may be not just one other group that assesses the project, but several other groups (1, 2, more, or all other groups in a course). Magin and Helmore (2001) and also Robinson (2002) studied the effect of additional peer assessors on the reliability, and perceived reliability, of the final assessment.

Finally, the number of assessees per unit of assessor may also differ. This refers to how many different projects an assessor (which may be a group) will have to assess (1, 2, more, or all other projects in a course). For example, in Purchase's (2000) study, students assessed 3 to 4 peers' computer interface designs.

Cluster V: Management of the Assessment Procedure

In the clustering by van den Berg et al. (2006b), it is obvious that the last cluster is not as comprehensive as the other clusters. Topping (1998) identified two variables (Requirement and Reward) that do not fit into the other categories, but they are not really a cluster on their own either. They actually seem to belong to a larger cluster, referring to the management of the assessment procedure. Certain other procedural issues were lacking in Topping's typology. We therefore added the variables 'Format', 'Training/guidance' and 'Quality control'.

New: Format. An extra variable was added that deals with the way peers assess or provide feedback, or perform the other aspects of the assessment process in which they are involved (depending on the scope).

The first option is that students are free to perform the assessment in a way they think is best. An example of this is when a teacher of a foreign

language class, at the end of a presentation by a student, asks the class what they think about it. This type of peer assessment is rather informal, and is not likely to be systematically studied.

The second option is that the staff provide certain guidelines. Examples are a criterion list for judgement (Omelicheva, 2005), the need to 'flag' comments as positive or negative in feedback (Topping et al., 2000), and general feedback rules (Prins et al., 2006; Prins et al., 2005).

The third option is that the staff provide a fixed format. Examples are a checklist for quantitative assessment (Purchase, 2000), a feedback form (possibly online) (Miller, 2003; Trahasch, 2004), or a template for making an assessment form when students are involved in the formulation of criteria (Prins et al., 2005).

An example of a format and its justification is given in the study of Purchase (2000). Students first had to mark the products of three peers using a checklist, and then they were required to rank the products with 'a gold, a silver and a blue star'. The author argues: "The stars did not translate into marks, but were important in ensuring that students reflected about what they had seen. Without this reflection, students may merely have marked the assessment criteria as either present or absent in an automatic, unthinking manner. The stars also encouraged them to consider qualitative and subjective judgements: the program that got the highest quantitative score may not necessarily have been the one that they thought was the best" (p. 345).

Requirement. This is the same variable that Topping defines: is peer assessment compulsory (e.g., in Purchase (2000) students would not get a mark for their own assignment if they did not mark their peers) or voluntary for assessors and/or assessees?

Reward. Reward is also maintained as a variable in the inventory. Do students receive course credit, other incentives, or reinforcement, for participation? Topping only applies this to the assessor; the inventory opens it up to the assessee too, in order to be comprehensive. When the reward is conditional on certain quality requirements we refer to the last variable, 'Quality control'.

New: Training/Guidance. In her research, Sluijsmans (2002) focuses explicitly on the need for, and impact of, the training of peer assessment skills. "It should be noted that peer assessment skills are not easily and automatically acquired. Peer assessment is considered a complex skill that needs to be developed" (Sluijsmans et al., 2006, p. 9). The extent to which students are prepared for and guided in their role as assessor and assessee is therefore added as an important variable in the inventory. Topping discussed the provision of training and guidance (monitoring the process and coaching) in his general organisational factors. By incorporating it into the inventory, however, its crucial role in the design of a peer assessment application is stressed.

New: Quality control. Quality of peer assessment is a broad issue and may, in fact, refer to all the dimensions of peer assessment discussed above. Moreover, as is extensively discussed in Gielen, Dochy, Onghena, Struyven, et al. (2007), the definition of how quality should be conceived is largely dependent on the goal (objective) one has for peer assessment. In this variable, it is limited to the type of 'control' that is exercised directly on the students and their assessments. Quality control may act pro-actively or reactively. Reactive examples are staff who run quality checks of scores awarded by students by assessing some of the work themselves (see also Topping, p. 267), or staff who introduce an automated quality control system to detect outlying scores submitted by peer assessors (e.g., Rada & Hu, 2002; Sitthiworachart et al., 2003). An example of pro-active quality control is related to the reward variable. In some contexts, assessors are stimulated to do their best to give extended and constructive feedback by means of a reward or sanction for feedback of high or low quality. For instance, in the study by Bloxham et al. (2004), a quarter of the total mark of the assessor is based on the quality of the peer assessment. In Searby et al. (1997), the same principle is applied to the quality of the open-ended feedback provided by students. Finally, in the study of Sitthiworachart et al. (2003), an extra peer is used to mark the quality of the feedback provided by peer assessors.

Conclusion

In 1998 Topping stated: "The next decade should bring a major expansion in the peer assessment literature. A more critical review, a best-evidence synthesis, and a meta-analysis should then become possible. Since peer assessment practices are so varied, future reports should include information on all 17 parameters in the typology and all 8 implementation factors, giving the basis for subsequent meta-analytic blocking" (p. 268).

Topping was right about the expansion of peer assessment research; the number of studies since his review has doubled, or even tripled. Despite the large number of studies available today, however, the type of review, synthesis or meta-analysis that Topping anticipated is still not feasible today. He hoped that his review and typology would "encourage fuller and more consistent reporting in the future and help promote more orderly, focused, coherent, and cost-effective onward research" (p. 267). This is where things went wrong. Most studies still do not provide a full description of their practice: they do not address the 17 parameters of Topping's typology, nor his 8 factors for implementation quality. And the variation in peer assessment practices has only expanded, making the typology in fact even inadequate to capture the diversity of peer assessment today.

This paper developed a more extended inventory of peer assessment diversity, based on a new review of the literature. Although a list of distinguishing features will never be exhaustive, this study was able to add some important variables and dimensions within variables to the typology. Eight new variables were added and another eight variables were extended with extra subdimensions. Five original variables were absorbed in larger entities, and also the implementation factors of Topping were given a place within the variables of the inventory. Finally, we grouped the 20 resulting variables into five clusters, building on an earlier clustering by van den Berg et al. (2006b). In Table 5, an overview of the new inventory is provided.

Finishing this work, we can only repeat Topping's plea to use this inventory as a guideline for the description of peer assessment practices in future studies. When this happens, it should become more straightforward to perform replication studies, to study interaction effects between certain variables, and finally to compare and synthesise findings on peer assessment. Beyond this scholarly interest, we hope this inventory also proves helpful for

practitioners, as a checklist of important decisions to take or an overview of possible alternatives to a specific practice. And finally, we aim at contributing to more clarity in the diversity of peer assessment practices.

Table 5

Summary table of the inventory of PA diversity

Cluster	**Variable**		**Dimensions & range of Variation**
Cluster I Decisions concerning the use of peer assessment	1)	Setting	Educational or non-educational use, curriculum area/subject, formal or informal learning, level of education, characteristics of participants, class size?
	2)	Object	Artefact or observed behaviour? (e.g. test, report, presentation, group work skills) Type of performance expected of learner? (e.g. reproduction, reflection) Information taken into account? (e.g. outcome, approach) Draft or final version?
	3)	Frequency & Experience	Once, sporadically or more frequently? Novel or familiar?
	4)	Objectives (goal)	Tool for social control, assessment, learning, learning-how-to-assess or active participation? Or a combination?
	5)	Function	Summative or formative?
Cluster II Link between peer assessment and other elements in the learning environment	6)	Alignment	Degree of alignment with curriculum, learning goals and teaching?
	7)	Relationship to other assessments	Other assessments? (Partially) substitutional or supplementary? Re-marking possible? If supplementary: before, simultaneous with or after staff assessment? Knowledge of other's judgement?
	8)	Scope of involvement	Aspects of involvement (e.g. definition of desired learning outcomes, design of assessment tasks, development of assessment criteria & standards, development of assessment procedures, judgements, decision taking, providing of knowledge of results/ feedback, monitoring/ guiding of a peer's progress) Extent of involvement (e.g. informed, discussed, participate or responsibility)
Cluster III Interaction between peers	9)	Output	Nature of information: quantitative and/or qualitative? Extent of 'condensation': at level of single criteria or global/holistic? Feedback stance: authoritative, interpretive, probing or collaborative?

	10)	Directionality	Unidirectional, reciprocal or mutual?
	11)	Privacy	Anonymity of assessor/ee? Teacher present? Output confidential or public?
	12)	Contact	In the presence of the assessee or at distance? Output face to face, in online discussion, or in writing? One-way or interactively? Time and place?
	13)	Role of assessee	Passive or active? Examples of active role: request, questions, preferences, immediate response, revision, reply.
Cluster IV Composition of the assessment groups	14)	Matching	Principle for matching? (e.g. random, year, ability, subject, friendship) Responsibility for matching? (e.g. teacher, students) Consistency of matching? (e.g. fixed or variable)
	15)	Constellation of assessors & assessees	Unit of assessor? (e.g. individual, pair or group) Unit of assessee? (e.g. individual, pair or group) Number of assessors per unit of assessee? (e.g. 1, 2, more, or all) Number of assessees per unit of assessor? (e.g. 1, 2, more, or all)
Cluster V Management of the assessment procedure	16)	Format	Freestyle, guidelines or fixed format?
	17)	Requirement	Compulsory or voluntary for assessor/ee?
	18)	Reward	Course credit, other incentive or reinforcement for participation to assessor/ee?
	19)	Training/ Guidance	Extent of training and guidance for assessor/ee?
	20)	Quality control	Presence of pro-active or reactive quality control?

References

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment & Evaluation in Higher Education, 27,* 427-441.

Bloxham, S. & West, A. (2004). Understanding the rules of the game: Marking peer assessment as a medium for developing students' conceptions of assessment. *Assessment and Evaluation in Higher Education, 29,* 721-733.

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education, 22,* 151-167.

Cheng, W. & Warren, M. (1999). Peer and teacher assessment of the oral and written tasks of a group project. *Assessment & Evaluation in Higher Education, 24,* 301-314.

Cheng, W. & Warren, M. (2000). Making a difference: Using peers to assess individual students' contributions to a group project. *Teaching in Higher Education, 5,* 243-255.

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98,* 891-901.

Falchikov, N. (1995). Improving feedback to and from students. In P. Knight (Ed.), *Assessment for Learning in Higher Education* (pp. 157-166). London: Kogan Page.

Falchikov, N. & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70,* 287-322.

Gibbs, G. (1999). Using assessment strategically to change the way students learn. In S. Brown & A. Glasner (Eds.), *Assessment matters in Higher Education: Choosing and using diverse approaches* (pp. 41-53). Buckingham: SRHE & Open University Press.

Gielen, S., Dochy, F., Onghena, P., Janssens, S., Schelfhout, W., & Decuyper, S. (2007). A complementary role for peer feedback and staff feedback in powerful learning environments. In S. Gielen, *Peer assessment as a tool for learning* (pp. 157-199). Unpublished doctoral dissertation, Katholieke Universiteit Leuven, Leuven, Belgium.

Gielen, S., Dochy, F., Onghena, P., Struyven, K., Smeets, S., & Decuyper, S. (2007). Goals of peer assessment and their associated quality concepts. In S. Gielen, *Peer assessment as a tool for learning* (pp. 41-66). Unpublished doctoral dissertation, Katholieke Universiteit Leuven, Leuven, Belgium.

Gielen, S., Tops, L., Dochy, F., Onghena, P., & Smeets, S. (2007). Peer feedback as a substitute for teacher feedback. In S. Gielen, *Peer assessment as a tool for learning* (pp. 95-124). Unpublished doctoral dissertation, Katholieke Universiteit Leuven, Leuven, Belgium.

Langan, A. M., Wheater, C. P., Shaw, E. M., Haines, B. J., Cullen, W. R., Boyle, J. C. et al. (2005). Peer assessment of oral presentations: effects of student gender, university affiliation and participation in the development of assessment criteria. *Assessment & Evaluation in Higher Education, 30,* 21-34.

Lockhart, C. & Ng, P. (1995). Analyzing talk in ESL peer response groups: Stances, functions and content. *Language Learning, 45,* 605-655.

Magin, D. & Helmore, P. (2001). Peer and teacher assessments of oral presentation skills: how reliable are they? *Studies in Higher Education, 26,* 287-298.

Malcolmson, C. & Shaw, J. (2005). The use of self- and peer-contribution assessments within a final year pharmaceutics assignment. *Pharmacy Education, 5,* 169-174.

McGourty, J. (2000). Using multisource feedback in the classroom: A computer-based approach. *IEEE Transactions on Education, 43,* 120-124.

Miller, P. (2003). The effect of scoring criteria specificity on peer and self-assessment. *Assessment and Evaluation in Higher Education, 28,* 383-394.

Nicol, D. J. & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31,* 199-218.

Norcini, J. J. (2003). Peer assessment of competence. *Medical Education, 37,* 539-543.

Oldfield, K. A. & MacAlpine, J. M. K. (1995). Peer and self-assessment at tertiary level - an experiential report. *Assessment & Evaluation in Higher Education, 20,* 125-131.

Orsmond, P., Merry, S., & Reiling, K. (1996). The importance of marking criteria in the use of peer assessment. *Assessment & Evaluation in Higher Education, 21,* 239-250.

Orsmond, P., Merry, S., & Reiling, K. (2002). The use of exemplars and formative feedback when using student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education, 27,* 309-323.

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

Pâquet, M. R. & Des Marchais, J. E. (1998). Students' acceptance of peer assessment. *Education for Health: Change in Training and Practice, 11,* 25-35.

Pope, N. K. L. (2005). The impact of stress in self- and peer assessment. *Assessment & Evaluation in Higher Education, 30,* 51-63.

Prins, F., Sluijsmans, D., & Kirschner, P. A. (2006). Feedback for general practitioners in training: Quality, styles, and preferences. *Advances in Health Sciences Education, 11,* 289-303.

Prins, F., Sluijsmans, D. M. A., Kirschner, P. A., & Strijbos, J. W. (2005). Formative peer assessment in a CSCL environment: a case study. *Assessment & Evaluation in Higher Education, 30,* 417-444.

Purchase, H. C. (2000). Learning about interface design through peer assessment. *Assessment & Evaluation in Higher Education, 25,* 341-352.

Rada, R. & Hu, K. (2002). Patterns in student-student commenting. *IEEE Transactions on Education, 45,* 262-267.

Robinson, J. M. (2002). In search of fairness: An application of multi-reviewer anonymous peer review in a large class. *Journal of Further and Higher Education, 26,* 183-192.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18,* 119-144.

Sadler, P. & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11,* 1-31.

Saito, H. & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research, 8,* 31-54.

Schelfhout, W., Dochy, F., & Janssens, S. (2004). The use of self, peer and teacher assessment as a feedback system in a learning environment aimed at fostering skills of cooperation in an entrepreneurial context. *Assessment & Evaluation in Higher Education, 29,* 177-201.

Searby, M. & Ewers, T. (1997). An evaluation of the use of peer assessment in higher education: A case study in the School of Music, Kingston University. *Assessment & Evaluation in Higher Education, 22,* 371-383.

Segers, M. & Dochy, F. (2001). New assessment forms in problem-based learning: The value-added of the students' perspective. *Studies in Higher Education, 26,* 327-343.

Sitthiworachart, J. & Joy, M. (2003). Deepening computer programming skills by using web-based peer assessment. In *Proceedings of the 4th Annual Conference of the LTSN Centre for Information and Computer Sciences.* NUI Galway (Ireland): LTSN-ICS.

Sivan, A. (2000). The implementation of peer assessment: An action research approach. *Assessment in Education: Principles, Policy & Practice, 7,* 193-213.

Sluijsmans, D. (2002). *Student involvement in assessment. The training of peer assessment skills.* Unpublished doctoral dissertation. Open Universiteit Nederland, Heerlen.

Sluijsmans, D., Brand-Gruwel, S., van Merriënboer, J. J. G., & Bastiaens, T. (2002). The training of peer assessment skills to promote the development of reflection skills in teacher education. *Studies in Educational Evaluation, 29,* 23-42.

Sluijsmans, D. & Prins, F. (2006). A conceptual framework for integrating peer assessment in teacher education. *Studies in Educational Evaluation, 32,* 6-22.

Smith, H., Cooper, A., & Lancaster, L. (2002). Improving the quality of undergraduate peer assessment: A case for student and staff development. *Innovations in Education & Teaching International, 39,* 71-81.

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

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

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.

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment & Evaluation in Higher Education, 25,* 149-169.

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

Trahasch, S. (2004). Towards a flexible peer assessment system. In *Proceedings of the Fifth International Conference on Information Technology Based Higher Education and Training, 2004.* (pp. 516-520).

van den Berg, I., Admiraal, W., & Pilot, A. (2006b). Peer assessment in university teaching: Evaluating seven course designs. *Assessment & Evaluation in Higher Education, 31,* 19-36.

van den Berg, I., Admiraal, W., & Pilot, A. (2006a). Designing student peer assessment in higher education: analysis of written and oral peer feedback. *Teaching in Higher Education, 11,* 135-147.

Venables, A. & Summit, R. (2003). Enhancing scientific essay writing using peer assessment. *Innovations in Education & Teaching International, 40,* 281-290.

Yorke, M. (2003). Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher Education, 45,* 477-501.

Zhang, S. (1995). Reexamining the affective advantage of peer feedback in the ESL writing class. *Journal of Second Language Writing, 4,* 209-222.

CHAPTER 5

PEER FEEDBACK AS A SUBSTITUTE FOR TEACHER FEEDBACK

Abstract

Peer feedback, as well as teacher feedback, can have a positive impact on student learning. This study compares both, addressing the question of whether peer feedback can serve as a substitute for teacher feedback. A pretest posttest control group design examined the long term learning effects of individual peer feedback and collective teacher feedback on writing assignments in secondary education. Moreover, it examined the added value of two measures to support the response of the assessee to peer feedback: an a priori question form and an a posteriori reply form. The study showed no significant difference in students' progress for essay marks between the condition with plain substitutional peer feedback and the control condition with teacher feedback. Both groups (plain peer feedback ànd teacher feedback) appeared, however, to make significantly less progress than the groups in the 'extended' feedback conditions with the question or the reply form.

Questionnaire data show that less than half of the students found the received peer feedback helpful, and less than a quarter found giving peer feedback helpful for their own learning process. Requiring the assessee to indicate personal feedback needs to the peer assessor beforehand, by means of the question form, had a positive influence on the perceived usefulness of the feedback, and also on the learning outcomes of the assessee. Although the requirement to demonstrate the use of the received feedback in a reply form also had a positive impact on performance, this procedure was disliked by the students.

PEER FEEDBACK AS A SUBSTITUTE FOR TEACHER FEEDBACK

Introduction

Why Use Peer Feedback as a Substitute for Teacher Feedback?

Peer feedback is a form of assessment that is performed by equal status learners. It does not contribute to the assessee's final grade and has a qualitative output, meaning that the assessor discusses strengths and weaknesses of a specific performance at length, and thereby also indicates suggestions for further improvement. It is the counterpart of feedback provided by a teacher. Both are outcomes of formative assessment, also called assessment for learning (Black et al., 1998).

This study addresses the question of whether, and in what form, peer feedback can be a substitute for teacher feedback.

A preliminary question to answer is, however: why should one even consider using peer feedback instead of teacher feedback? Sadler (1998) argues that it is important for students to become progressively independent of their teacher for lifelong learning and thus they have to acquire self-assessment skills in the long term. We therefore have to address these skills intentionally in our curriculum, and not leave them as an 'accidental or inconsequential adjunct' to the curriculum. Peer assessment is an excellent way to teach students the necessary assessment skills (see also 'peer assessment as a tool to learn-how-to-assess' in Gielen, Dochy, Onghena, Struyven, Smeets, & Decuyper, 2007). If peer feedback proves to be a worthy substitute for teacher feedback (peer assessment as a tool for learning), and in the meantime it teaches students to become self-regulated learners, two birds are killed with one stone. Another reason to introduce peer feedback is the more democratic concern of allowing students to be active participants in the assessment process (see also 'peer assessment as a tool for active participation' in Gielen et al., 2007). A third reason might be one of time saving, although this should not be the main goal of using peer assessment. Organising peer feedback also 'costs' considerable class and teacher time (Ballantyne et al., 2002), and time saving can only be considered a pleasant

side-effect if it does not come at the cost of reduced learning opportunities. Finally, if peer feedback should prove to be even more effective in the support of student learning than teacher feedback, then this is definitely a good reason to prefer it.

Comparison of Peer Feedback and Teacher Feedback

Sadler (1998) points out that good feedback lies at the heart of good pedagogy and that the source of the feedback (be it the teacher or, for instance, a peer) that facilitates learning is less important than its validity. Trust and personal interaction are important elements, he adds. This addition, however, emphasises two characteristics that might not always be present or feasible in both sources. Although peer assessors and teacher assessors might follow the same procedure in order to assess, they do this from different backgrounds. Since a teacher's background is more sophisticated, his feedback may be more trustworthy. On the other hand, teachers have to divide their time for personal interaction among a lot of students, giving peer feedback an advantage at this point.

A good teacher brings some baggage and skills to this process that might not be available for students (Sadler, 1998). A good teacher has superior knowledge, a set of attitudes and dispositions towards teaching as an activity and towards learners (e.g., empathy, desire to help), and a deep knowledge of criteria and standards or insights into the set of expectations for a specific assignment. Teachers also bring expertise in judgement by having made judgements about student efforts on similar tasks in the past, which also gives them insight into a variety of ways to solve the assignments, and the difficulties encountered by previous students. Finally, a good teacher has expertise in framing feedback statements for students in a way that maximises the learning effect for a particular student.

One might say that, for these reasons, students are not appropriate assessors and their feedback cannot function as a worthy substitute for teacher feedback. One might argue, however, that Sadler describes an 'ideal' teacher who does not always match with the characteristics of the 'average' teacher. Moreover, one might also train students' peer assessment skills so that their feedback becomes as effective as teacher feedback in the end (Sadler, 1998; Sluijsmans, 2002). Furthermore, peer feedback itself has also some advantages over teacher feedback that have a positive effect on student learning.

Firstly, peer feedback can increase the social pressure on students to do their best for an assignment (Cole, 1991; Pope, 2005). The potential embarrassment of colleagues – rather than of a teacher – seeing their poor quality work increases time and effort spent by students on assignments (Gibbs et al., 2004; Pope, 2001). In this light, the actual output of the peer feedback does not even matter; a mere announcement that it will take place might be enough to raise performance.

Secondly, research shows that peer feedback is often perceived as more understandable and more useful by students, because fellow students 'are on the same wavelength' (Topping, 2003). Teachers, being experts in the domain, often provide feedback that is based on a thorough insight in the complexities of the subject and the expectations of a domain. Although, as a teacher, they should be able to translate this for their students, research shows that they do often not succeed in this. Their feedback is often not understood or is misinterpreted (Gibbs et al., 2004; Yang et al., 2006). Higgins (2000) explains that feedback messages such as '*be more critical'* or '*your arguments need to be more academic'* do not have the same meaning for teachers and students, because they are associated with a discourse that is not directly accessible to students (Hounsell, 1987).

An additional, third, argument in favour of peer feedback is that the development of students' abilities to understand feedback (which is important for further learning) is expected to be one of the outcomes of the use of peer feedback. Making students participate in the assessment process (including feedback) enlarges their insight into it (Bloxham et al., 2004). Topping (1998) refers to this as an aspect of 'assessing for learning' (see also Gielen et al., 2007). An explanation for this is that by allowing a learner to see what happens behind the curtains of an assessment, and to participate in it, clarification and internalisation of these goals are supported (e.g., Rust et al., 2003). Having a clear view of the goals, criteria and standards is necessary to understand what feedback is aiming at, but it can even in itself (without feedback) raise the learning outcomes by generating appropriate learning activities (Gibbs et al., 2004).

Fourthly, peer feedback can realise a gain in speed of return of feedback. Teacher feedback is often provided with a considerable delay after the submission of an assignment or the administration of a test. Assessing a large group of students' work does take time, and assessment often receives a low priority in teachers' agendas. As a result, feedback sometimes is not

available until after the course has finished. In fact, this feedback is likely to be a waste of time. In that case, "imperfect feedback from a fellow student provided almost immediately may have much more impact than more perfect feedback from a tutor four weeks later" (Gibbs et al., 2004, p. 19).

Fifthly, the frequency or amount of feedback can also increase with peer feedback. Gibbs and Simpson emphasise that, for feedback to be useful, it should be provided regularly; at each step in a learning process. Waiting until the end and, for instance, only commenting on the final essay or report of a project, is not enough to support learning effectively and may provoke a lot of frustration on the part of the learner. The introduction of several 'intermediate' peer assessment sessions on draft versions of an essay or report could bring a solution, if staff are not able or willing to increase their frequency of providing feedback.

A sixth possible advantage lies in the level of individualisation of feedback. If staff try to provide more timely and more frequent feedback, they often choose to organise it collectively to make this feasible. Collective feedback cannot, however, address personal needs as much as individual feedback can. Moreover, the opportunity for personal interaction, identified as crucial by Sadler (1998), decreases: perhaps the opportunity to ask questions is offered, but a student has to share the teacher's time with several other students, and in his answer a teacher will try to address the collective interest in the question at the expense of personal interest. Additionally, students are not likely to show their ignorance or uncertainty during a collective session, so a lot of questions will not even be posed. Peer feedback can make it feasible to provide individual feedback and in the meantime, since the teacher does not have to provide general feedback in front of the class, the teacher may be available for personal interaction when assessors and assessees cannot find an answer to a specific question.

A final argument in favour of peer feedback lies in the association of feedback with power issues, emotions, and identity, which may launch an 'emotion-defence system' in students (Higgins, 2000). As a consequence, students may hide their weaknesses and doubts from the teacher, rendering teachers unaware of particular student difficulties or misconceptions. In that case, teacher feedback is less likely to connect to the learner, since it fails to address their problems or concerns. Peer feedback may bypass some of these difficulties since it is less power-sensitive.

To conclude, peer feedback can have an advantage for supporting learning in several ways. The social pressure stimulates students to try harder and perform better. Peer feedback might be more understandable, and the activity of giving feedback in itself may raise students' understanding of the learning goals. Furthermore, it can be more timely, more frequent, and more individualised. Finally, it may elicit fewer defensive reactions. These benefits of peer feedback should be weighed against the benefits of teacher feedback.

Previous Research

The comparison of peer feedback and teacher feedback and their learning effects has been addressed previously in several recent empirical studies. Yang, Badger, and Yu (2006) found, in their study of an English writing class at a Chinese university, that the impacts of teacher and peer feedback are different. More teacher feedback than peer feedback is incorporated in the revision of an essay, since students consider the teacher to be more 'professional', 'experienced' and 'trustworthy' than their peers. However, peer feedback appeared to bring about a higher percentage of meaning-changing revision while most teacher-influenced revisions happened at surface level. At the same time, teacher-initiated revisions were also less successful than peer-initiated revisions, due to more misinterpretations of the teacher feedback. Moreover, the study of Yang et al. (2006) also revealed that students who received peer feedback showed more initiative for self-correction. Having reservations about the feedback students received from peers stimulated them to look for confirmation in some way, by checking grammar books or asking the teacher, and to develop their own independent ideas for revision. In contrast, exposure to teacher feedback lowered their initiative for self-regulated learning, perhaps because students believed that the teacher had pointed out all their mistakes and there was no need for further correction.

The impact of peer and teacher feedback on the writing of secondary school students was studied by Tsui and Ng (2000). In their study, all students addressed a higher percentage of teacher feedback than of peer feedback in their revisions, but there was considerable individual variation. They also noted that some students reported that they benefited from reading other students' work as they prepared to give feedback. This last observation is related to the 'assessing for learning' principle, described above.

Kim (2005) explicitly separated the effects of giving and receiving peer feedback in her study. The assessee was also required to perform a "back-feedback activity" in which he wrote a reply to his assessor explaining his agreement or disagreement with the feedback. Kim studied the performance of students making a concept map, using two categories of performance criteria: factual 'objective' criteria (e.g., minimum 12 nodes present) and complex 'subjective' criteria (e.g., clarity of structure). Concerning the 'objective' criteria, Kim (2005) found a positive effect of the role of the assessee, compared to a control group, but not of the role of the assessor. Neither of the roles was found to be beneficial for the 'subjective' criteria, however. Surprisingly, playing either of the roles alone still promoted better performance than playing both roles together. Possible explanations are a lack of time for students to perform both roles appropriately, a lack of in-depth feedback (most peer assessors' feedback only indicated what was wrong, without providing suggestions for improvement), and a lack of peer interaction to discuss the different ideas (Kim, 2005).

Finally, Sadler and Good (2006) only focused on the perspective of 'assessing for learning', and on peer marking in particular. They compared the learning effects of students' experiences of peer grading, self grading, and no grading. All students took a test and then developed a scoring rubric for the test together with their teacher. One group then used this rubric to grade the tests of peers, another group graded their own tests, and a third group did no grading. The teacher graded the tests of all three groups, and these grades were used as a pretest measure of performance. The authors did not report on any feedback received by the students. An identical test was administered one week later as a posttest. Controlling for the pretest performance, the treatment had a significant effect on the posttest performance: the self grading group significantly outperformed the peer grading and control groups, which did not differ significantly from each other.

To summarise, the current empirical evidence seems to suggest that peer feedback is attended to with more reservations, leading to less impact, but on the other hand leaving room for self-correction. It also reduces the chance of misinterpretation, resulting in more successful revisions. Some researchers only found effects on simple performance criteria, while others also found them on more complex criteria. Finally, some found evidence for the assessing for learning principle, while others did not.

Aim of this Study

Since previous research has not yet found a conclusive answer to the question of the learning effects of peer feedback, this study will compare the effects of peer feedback versus teacher feedback on a complex performance, namely a writing assignment. The roles of assessor and assessee will not be separated, as in some previous studies, since it does not make sense to separate them in a realistic setting, and conclusions about the combined effect are most meaningful for practice.

The main research question is whether peer feedback can have an impact on learning equal to that of teacher feedback. Furthermore, we examine the impact of two specific design features of peer feedback (an a priori question form and an a posteriori reply form) that are introduced to attune feedback to the assessee's needs and to encourage the assessee to make use of the received feedback. These measures are based on Gibbs' and Simpson's description of conditions under which assessment and feedback support learning (Gibbs et al., 2004). The impact of the 'extended versions' of peer feedback that are supplemented with one of these measures is compared with a peer feedback design without these extra features.

Method

Participants

A total of 85 first grade students (12-13 years old) participated. They were divided into four different classes from the same secondary school, all taught by the same teacher. All were registered in the theoretically oriented *general secondary education* track (ASO). One class (N=24) functioned as a control group and the other three used peer assessment.

Peer Assessment Design and Procedure

The description of the particularities of the design and procedure of the peer assessment in this study will be structured by means of the inventory of peer assessment diversity (see Gielen, Dochy, & Onghena, 2007) in Table 1.

Table 1

Description of the current peer assessment (PA) design and procedure

Cluster	**Variable**	**Description of the current design and procedure**
Cluster I Decisions concerning the use of peer assessment	Setting	Educational use, Dutch writing curriculum (students' mother tongue), formal learning, class assignments during 2nd and 3rd trimester, 1st year general secondary education, 63% male, 12-13 years old, class sizes: 24, 19 and 22, all classes same teacher.
	Object	Artefact: several types of essays (a position paper, a story, a newspaper article, a reader's letter) Type of performance expected of learner: creative writing Information taken into account: writing performance (outcome) Draft version of a two-stage assignment
	Frequency & Experience	Frequently: PA of 3 successive assignments Novel to students
	Objectives	Tool for learning and learning-how-to-assess
	Function	Formative
Cluster II Link between peer assessment and other elements in the learning environment	Alignment	The learning goals of writing essays are central in the Dutch curriculum. Students assess the 'normal' writing assignments of the course, no 'extra' products. Normally, these assignments are not two-stage, so the intermediate feedback is considered extra. PA is also new to teachers, so the alignment with teaching practices is not perfect: teachers struggle with the time scheme, with the feedback forms and with the practical arrangements of assessors or assessees being absent or not prepared.
	Relationship to other assessments	Peers are the only assessors of the drafts. After revision the teacher assesses the final versions summatively (grade with a large delay in time).
	Scope of involvement	Aspects & extent of involvement: participation in development of assessment criteria, responsibility for formative judgements for each criterion, responsibility for providing knowledge of results (on 4-point-scale per criterion) and feedback.
Cluster III Interaction between peers	Output	Nature of information: quantitative (4 point scale with stars) and qualitative Extent of 'condensation': at level of single criteria Feedback stance: depending on student (prompts for evaluative remarks + collaborative suggestions)
	Directionality	Unidirectional
	Privacy	No anonymity of assessor/ee Output confidential (as regards third, not teacher & researcher)

	Contact	Writing feedback starts in class (e.g., last 20' of a lesson) and is finished at home Output provided in writing, read at home
	Role of assessee	Class 1: active by revision (=PA group) Class 2: active by revision and reply to teacher (=PA-REPLY group) Class 3: active by questions for assessor & revision (=QUEST-PA group)
Cluster IV Composition of the assessment groups	Matching	Principle for matching: same ability (based on pretest: writing exam December) Responsibility for matching: researcher Consistency of matching: fixed
	Constellation of assessors & assessees	Unit of assessor: individual Unit of assessee: individual Number of assessors per unit of assessee: 1 Number of assessees per unit of assessor: 1
Cluster V Management of the assessment procedure	Format	Fixed format (see Appendix): Class 1: form A, class 2: forms A and B, class 3: forms C and D. *Feedback form A:* paragraph per criterion (total=6), prompts for strengths + justification, for weaknesses + justification, for suggestions and for a quantitative judgement (colour 0-1-2-3 stars in). *Reply form B with following prompts:* By receiving/giving feedback I learned …, I revised my work on the following criteria …, My best piece is …, I paid special attention to … *Question form C with following prompts:* I paid attention to …, I doubt …, I found it difficult to …, I wish for feedback on the following criteria … *Feedback form D:* prompts for assessor's opinion on all aspects mentioned in the question form, plus one paragraph for each requested criterion (max=3) with prompts for strengths + justification, for weaknesses + justification, for suggestions and for a quantitative judgement (colour 0-1-2-3 stars in)
	Requirement	Compulsory for assessor/ee
	Reward	None
	Training/ Guidance	Explanation of the rationale of peer feedback and of the requirements of the feedback form that uses guiding prompts. Discussion of a worked out example of the peer assessment process. During PA, help if students do not know what to write.
	Quality control	None

Research Design

The present study is designed to focus on the effects of peer feedback on performance in the medium to long term, instead of measuring performance gains immediately after an assignment is submitted. The focus on longer term effects is chosen to allow students to become acquainted with the new feedback arrangements (Ballantyne et al., 2002) and to allow these feedback arrangements to develop their impact on learning, which may be delayed (Sadler, 1998).

A pretest posttest control group design is applied (see Figure 1). For the performance measures, the pretest consists of the writing assignment in the Dutch exam of the first trimester (December). The posttest is administered in the final exam of the third trimester (June). At the end of the third assignment, a short questionnaire was administered in the experimental conditions that collected information on students' perceptions regarding peer feedback.

The experimental group actually consists of three peer feedback arrangements, in which the roles of the assessees differ. The rationale for these arrangements and the differences between them will be addressed in the 'Variables' section below. The three experimental conditions will be compared to each other, as well as to a control condition, to answer the main question of this study: Can peer feedback function as a substitute for teacher feedback?

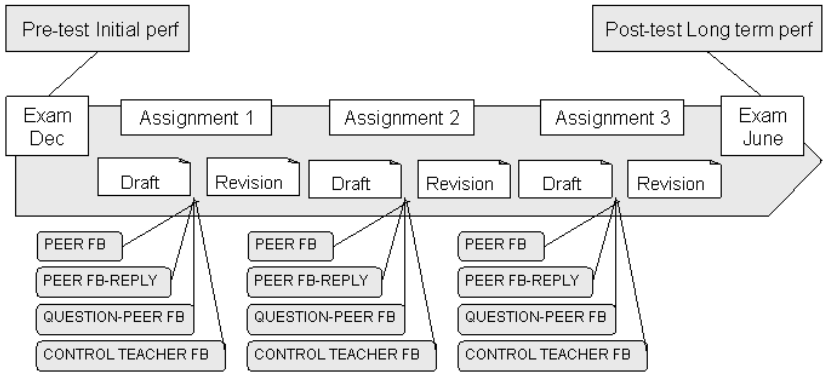


Figure 1. Representation of the research design.

Choice of Control Treatment

A difficult issue concerning peer assessment research is the choice of a control condition. What are the features of a 'zero treatment' (Kember, 2003)? A control condition should have the features of the 'normal, traditional, usual' way of doing things. But in our topic, it is not that easy to define one commonly shared design of a learning environment.

First of all, we restricted our study to a learning environment where the use of 'authentic assignments' (i.e., the essay assignments in our case) was already implemented as the mode of assessment. Thus these assignments were a shared feature for both the experimental and the control groups.

A second choice to make was whether or not teacher feedback was present in the learning environment and whether peer feedback was considered from a supplementary or a substitutional perspective. These choices resulted in three possible research contexts, defining different control conditions and different expectations.

Firstly, assessment can be regarded as an add-on, where peers give supplementary feedback to each other in addition to teacher feedback (or vice versa). The control condition would, in this case, receive teacher feedback. The question would be whether the experimental condition 'does better' (in terms of learning outcomes) than the control condition.

Secondly, peer assessment can be used to replace teacher feedback. In this case, it should also be compared with a control condition that receives teacher feedback. The central question in this case is whether the experimental condition does (as a minimum) equally as well as the control group (i.e. 'non-inferiority'), to find out whether peer feedback is a worthy substitute for teacher feedback.

A third possibility is that peer assessment compensates for the absence of teacher feedback. In that case, the control condition receives no feedback. Once more, the experimental condition should do better than the control condition to make peer assessment a better option than no feedback.

The focus of this study is on peer feedback as a substitute for teacher feedback. This eliminates the first choice. Comparing the second and third choices, the second appears to be the most sustainable from an ethical point of view. In an ecologically valid study, one has to take account of teachers' concerns to give all their students equal opportunities (Kember, 2003). A teacher would not allow participation in a long term study where some

students are not allowed to receive feedback and others receive extensive feedback. For these reasons, the second set-up was chosen. The experimental and the control conditions provide an equal amount of feedback, so their impact is expected to be similar too. One difference, however, is that peer feedback is provided on an individual basis, while teacher feedback is provided collectively. It is clear that the effects of the source of feedback are compounded with the effects of the level of individualisation of feedback. However, the interaction of both effects is precisely what happens in classrooms where it is not at all feasible for staff to provide frequent individual feedback to students on assignments (Ballantyne et al., 2002). This study aims to produce ecologically valid conclusions that are generalisable to classrooms where a teacher is confronted with the choice of providing collective teacher feedback or organising a peer feedback system to allow for individual feedback. Of course, a combination of both would also be possible, but this would require additional class time and an increased organisational burden.

Variables

Difference in performance. The exam essays, which were used to measure performance in the mid-long term and at the beginning of the study, were rated by the teacher. Since learning goals proceed during a semester, receiving the same mark in June as in December does not mean that a student's writing skills have not improved: the expectations are simply higher. Therefore it should be noted that the difference in performance measures cannot be interpreted as an absolute measure of 'progress'. This poses no problem for a relative comparison between groups, as is intended by this study, however.

In the first exam, students could choose between six topics (The final solution for traffic jams, Who wins Idol 2004?, What did Santa Claus bring this year?, Minister bans soft drinks in schools, Reading is good for your health, Winter dawns). At the second they could choose between four topics (Today's youth prefers telly above friends, Mother sells child, More busses and trams needed in cities, One out of two knows his neighbours). They wrote essays of 25 lines on average. Important criteria for the essays are: own opinion, good argumentation, clarity, variety in writing, readability, and spelling.

Condition. The study consists of three conditions with a version of peer feedback and one control condition with teacher feedback. Conditions were assigned at class level, randomly. Peer feedback conditions differed in the roles of the assessees. The different roles of the assessees that differentiate between the three experimental conditions refer to Gibbs' and Simpson's (2004) remark that it is important to address explicitly students' responses to feedback. Although all feedback arrangements in the study address this response to a certain extent by inserting the feedback in the middle of two-stage assignment, and in that way stimulated students to revise their work after the feedback, an extra measure was added in two peer feedback groups to encourage this response. In the first group (the 'PA-REPLY group') students were asked to report – in a written reply to the teacher – which feedback comments they took into account and how they did this, what they learned from giving peer feedback, and what they thought of their own accomplishments. This measure aims at closing the "feedback loop" (see Boud, 2000) and encouraging a "mindful reception" of the feedback (Bangert-Drowns et al., 1991). In the other group (referred to as the 'QUEST-PA group'), students were additionally required to complete a 'question form' in which they indicated to their peers the aspects or criteria on which they think they had some problems and for which they requested feedback. They could also add questions to their peers. It is hoped that this measure motivated and guided assessors to give useful feedback, and that assessees felt more personally addressed by the feedback.

The 'plain' peer feedback condition and the PA-REPLY condition shared the same feedback form (see Appendix, and its description in Table 1). The feedback form in the condition QUEST-PA condition is also slightly adapted from the PA-feedback form, in that it tried to direct peers to address the questions and comments of the assessee. They were expected to comment mainly on the requested criteria, not on the whole list (see Appendix and Table 1).

In the fourth class, the control condition, the teacher provided collective written feedback to the class, based on his assessment of a sample of draft essays. The teacher's feedback form was similar to the PA-feedback form, with similar guiding questions and the same criteria (see Appendix). For assignment three, this feedback was only discussed orally, based on a

model answer that the teacher distributed in class. This model, without the extended explanation, was also distributed in the experimental conditions.

Students' perceptions. A questionnaire was administered that asked students how they experienced giving peer feedback and receiving peer feedback and whether they would wish to continue using formative peer assessment for other courses or assignments in future. Answers to the first two questions were split into 'helpful' and 'not particularly helpful'. Answers to the third question were divided into 'wish for more peer feedback', 'wish for no more peer feedback'.

Research Questions

Based on the aforementioned prior research, we developed the following research questions:

1. Does any difference in scores between the pretest and the posttest differ between the three peer feedback conditions and the control condition?
2. Does any difference in scores between the pretest and the posttest differ between the two 'extended' peer feedback conditions and the 'plain' peer feedback condition?
3. Do students perceive giving and receiving peer feedback as helpful, and does this perception differ between the 'extended' peer feedback conditions and the 'plain' peer feedback condition?
4. Do students wish to use peer feedback in future, and are there any differences between the 'extended' peer feedback conditions and the 'plain' peer feedback condition?

Analyses

Data for research question one and two were analysed by means of descriptive statistics for the measures used in the study and by means of an Analysis of Variance (ANOVA), using the GLM-procedure with Dunnett's *t* tests for comparison of treatment groups against a control group in the SAS System (SAS Institute Inc., 2004). The questionnaire data for research questions three and four are studied using the FREQ-procedure from the SAS System (SAS Institute Inc., 2004).

Results

Descriptive Analyses

The summary statistics in Table 2 and the side-by-side box plots of the differences for each condition (Figure 2) show that the PA-group and the control group have a negative mean score, whereas the two other conditions have positive mean difference scores.

Table 2

Summary statistics

Condition	*N*	Mean	Std Dev
PA	21	-0.45	0.69
PA-REPLY	19	0.16	0.60
QUEST-PA	22	0.32	0.63
Control	23	-0.22	0.75

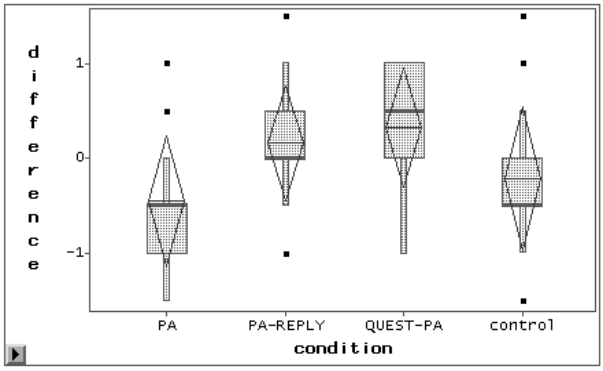


Figure 2. Box plot of differences in performance between the posttest and the pretest for each condition.

Comparison of All Experimental Groups against the Control Group

An ANOVA on the differences in performance in the posttest compared to the pretest shows a significant effect of condition overall, $F(3,81)=5.78$, $p=0.0012$. The comparison of the three experimental conditions against the control group (peer feedback conditions versus teacher feedback) by means of Dunnett's t test (Table 3) shows that only the QUEST-PA group differs significantly from the control group. The group that received and gave peer feedback, and additionally used the question

form, outperformed the group that received teacher feedback. The group with 'plain' peer feedback and the group with peer feedback and the reply form did not differ significantly from the control group.

Table 3

Dunnett's t tests for comparison of all experimental groups against the control group

Condition Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
QUEST-PA - control	0.5356	0.0543	1.0169	***
PA-REPLY - control	0.3753	-0.1251	0.8756	
PA - control	-0.2350	-0.7221	0.2521	

Comparisons significant at the 0.05 level are indicated by ***.

Comparison of the Two 'Extended' Peer Feedback Groups Against the 'Plain' Peer Feedback Group

For the second research question, we limited the dataset to the three experimental conditions. The ANOVA on the difference in performances in the posttest compared to the pretest again shows a significant effect of condition, $F(2,59)=8.51$, $p=0.0006$. We then compared the different designs of peer feedback by means of Dunnett's *t* tests, considering the plain peer assessment condition as a control group and comparing this with the conditions in which the question form or the reply form was added. The peer assessment condition is significantly different from both other conditions (see Table 4). Students who were asked to indicate their needs for feedback, or to reply after receiving feedback, made more progress between the pretest and the posttest than students who just gave and received peer feedback.

Table 4

Dunnett's t tests for comparison of the two 'extended' peer feedback groups against the 'plain' peer feedback group

Condition Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
QUEST-PA - PA	0.7706	0.3271	1.2140	***
PA-REPLY - PA	0.6103	0.1501	1.0705	***

Comparisons significant at the 0.05 level are indicated by ***.

Comparison of the Two 'Plain Feedback' Groups against the Two 'Extended Feedback' Groups

The descriptive analyses of the difference in scores for each condition show that two conditions realised a progress in students' mean scores and two conditions, in contrast, resulted in a decrease. Although it was not expected beforehand, it seems that the main feature that differentiates between the four classes might not be the source of the feedback, but the additional measures that were taken to stimulate a mindful reception of it. Clearly, as can be seen in Table 5, the two extended feedback groups had a significantly higher difference in scores than the two 'plain' feedback groups, $t(84)=3.92$, $p=0.0002$.

Table 5

Estimate of the contrast of the two 'plain feedback' groups against the two 'extended feedback' groups

Parameter	Estimate	Standard Error	*t* Value	Pr > |*t*|
Extended feedback vs. plain feedback	1.146	0.293	3.92	0.0002

Comparison of Students' Perceptions of Peer Feedback

Examination of the students' perceptions shows that students did not like the peer assessment procedure very much and did not find it very effective (see Table 6): on average only 44% considered the received feedback really useful, and only 23% found the experience of giving feedback really helpful. Moreover, up to 63% on average wished not to use peer feedback in future, for the reasons that it was boring and a waste of time.

Perceptions of the received feedback differ considerably between the groups. In the PA and PA-REPLY conditions, only 37-38% of the students reported the received feedback to be helpful, while 57% of the students did in the QUEST-PA condition. Also, with respect to the desire to continue using peer feedback, one group differs remarkably from the others. Concerning this variable, the PA-REPLY group had more negative answers than the other groups. Whether or not these observed differences are significant will be described in the next paragraph.

Table 6

Percentages of positive answers to the perception items on peer feedback, per condition

Condition	Receiving is helpful	Giving is helpful	Wish for more peer feedback
PA	38%	19%	47%
PA-REPLY	37%	26%	13%
QUEST-PA	57%	24%	47%
TOTAL	44%	23%	37%

We compared the results of both extended peer feedback conditions (with question or reply form) statistically to the condition with plain peer feedback. The odds ratio was calculated from each two-by-two table, indicating the ratio between the odds that students found giving (receiving) peer feedback helpful (desirable) against not helpful (undesirable) in the 'extended' peer feedback condition and the plain peer feedback condition (see Table 7).

Table 7

Overview of the odds ratios and confidence limits

Predictor	Outcome	Odds Ratio	95% Confid. Limits	
PA-REPLY vs. PA	Giving Helpful vs. Not helpful	1.5	0.3	6.8
PA-REPLY vs. PA	Receiving Helpful vs. Not helpful	0.9	0.3	3.4
PA-REPLY vs. PA	Wish More vs. No more	0.2*	0.03	0.9
PA vs. PA-REPLY	*Wish More vs. No more*	*6.3**	*1.1*	*35.7*
QUEST-PA vs. PA	Giving Helpful vs. Not helpful	1.3	0.3	5.8
QUEST-PA vs. PA	Receiving Helpful vs. Not helpful	2.2	0.6	7.4
QUEST-PA vs. PA	Wish More vs. No more	1.0	0.3	3.8

Only one odds ratio is significantly different from 1, indicating a significant relationship between the variables 'PA-REPLY or PA' and 'Desire to work with peer feedback in future'. To make the interpretation easier, we switched the order of the predictor values (italic row in Table 4), resulting in the following conclusion: the odds that students wish to work with peer feedback in future, against not wishing this, are 6.3 times higher when students experienced the plain PA-condition than the PA-REPLY

condition. The additional requirement of having to write a reply on the peer feedback thus made peer assessment significantly less attractive to students. Although the odds of experiencing the received peer feedback as helpful, against not being helpful, seem twice as high in the QUEST-PA condition than in the plain PA condition, this effect is not found to be significant with the given sample size.

Conclusion and Discussion

The study showed no significant difference in students' progress on essay marks between the condition with plain substitutional peer feedback and the control condition with teacher feedback.

The question of whether peer feedback can be used substitutionally, without a considerable loss in effectiveness, can thus be answered positively from our data. Our study does not confirm the findings of Yang and colleagues (2006), who found a larger impact and more improvement in the teacher feedback group when considering performance. Two differences from Yang's study should be noticed: it concerned short term effects (i.e. revision of the product that was the subject of feedback), and teacher feedback was individual.

Both groups (plain peer feedback ànd teacher feedback), however, appeared to make significantly less progress than the groups in the extended feedback conditions. The condition in which assessees first indicated their needs to the assessor (QUEST-PA) was shown to lead to more progress than both the control condition and the plain peer feedback condition. The PA-REPLY condition, in which students justified their use of the received peer feedback, appeared to be significantly more effective than the PA condition, but not better than the control condition with teacher feedback.

We can conclude that the extensions of the peer feedback had significant influences on its effectiveness for learning. A possible explanation for the effect of the question form is that assessors may provide more useful feedback when informed about the questions and doubts of the assessee beforehand. Moreover, this feedback may receive more attention from the assessee since it addresses personal questions and doubts. In the case of the

reply form, an explanation might be that it fostered reflection on the received feedback and the necessary revisions, realising a 'closed feedback loop' (see Boud, 2000). Since these conditions also outperform the control condition with teacher feedback, one might wonder whether these extensions would have had the same effect when added to the teacher feedback. Based on the arguments of Higgins (2000), described earlier, one might expect that the 'question form' and the 'reply form' will be perceived in different ways by assessees, when applied to teacher feedback instead of to peer feedback. The reason is that teacher feedback is associated with power issues, emotions and identity which may launch an 'emotion-defence system' in students and let students hide their weaknesses and doubts for the teacher (Higgins, 2000). Further research is needed to address these questions.

The questionnaire data confirm to a certain extent that the question form led to more effective feedback, which can be concluded from the learning outcomes. In the other peer feedback conditions, only 37-38% of the students reported the received feedback to be helpful, while in the QUEST-PA condition 57% of the students did so. Although this difference was not significant, this tendency in the perceptions, together with the significant difference in performance outcomes, suggests that the question form deserves further attention in research that tries to raise the quality of peer feedback. Until now, most studies have focused on two other methods to improve the quality of peer feedback: training and guidance of student's peer assessment skills on one hand (e.g., Sluijsmans, 2002), and external quality control on the other hand (e.g., Searby et al., 1997; Sitthiworachart et al., 2003; Topping, 1998). Letting the assessee indicate what he needs might be a more natural way of raising the quality of feedback. A replication study is needed to examine whether the feedback in a condition with a question form is indeed better than in a plain peer feedback condition. Moreover, future research might also directly compare the effectiveness of three methods to raise feedback quality: training, external control, and a question form.

Whereas the reply form was also showed to have a significant learning benefit when examining the outcome measures, this condition had no impact on the perceived helpfulness of giving and receiving feedback. This can be explained by the fact that the merit of the reply form is realised

later in the feedback loop, when giving and receiving peer feedback are already over.

The extra learning benefit from 'assessing for learning' was only recognised by 23% of the students overall. The latter finding might explain why students did not like the peer assessment experience very much. A considerable number of students (53-87%) did not wish to use peer feedback in future assignments or courses, because they considered it boring or a waste of time. The 'paper work' was especially disliked, and this is mostly applicable to the activity of giving feedback. If students did not experience the value of this, it is understandable that it became a boring activity. One would expect that the extra 'paper work' in the 'extended conditions' would have an even more detrimental effect on students' liking for the peer assessment experience. This is indeed the case in the PA-REPLY condition, where the students' desire to use peer feedback in future is significantly lower than in the PA condition. Although the outcome measures show that this extension certainly had an effect on student performance, the added value of the reply form was probably not recognised by the students themselves. However, the extra 'paper work' in the QUEST-PA condition did not have a negative impact on the preferences for peer assessment. An explanation might be that the value of the extra question form was clear to the students, since they experienced its positive influence on the helpfulness of the received feedback. These results show that students' likes or dislikes for a specific teaching method do not always match the learning benefit that is associated with it, but that their perception of the usefulness of a particular intervention appears to be more in line with the final learning effect.

References

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment & Evaluation in Higher Education, 27,* 427-441.

Bangert-Drowns, R., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61,* 213-238.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5,* 7-74.

Bloxham, S. & West, A. (2004). Understanding the rules of the game: Marking peer assessment as a medium for developing students' conceptions of assessment. *Assessment and Evaluation in Higher Education, 29,* 721-733.

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education, 22,* 151-167.

Cole, D. (1991). Change in self-perceived competence as a function of peer and teacher evaluation. *Developmental Psychology, 27,* 682-688.

Gibbs, G. & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education, 1,* 3-31.

Gielen, S., Dochy, F., & Onghena, P. (2007). An inventory of peer assessment diversity. In S. Gielen, *Peer assessment as a tool for learning* (pp. 67-94). Unpublished doctoral dissertation, Katholieke Universiteit Leuven, Leuven, Belgium.

Gielen, S., Dochy, F., Onghena, P., Struyven, K., Smeets, S., & Decuyper, S. (2007). Goals of peer assessment and their associated quality concepts. In S. Gielen, *Peer assessment as a tool for learning* (pp. 41-66). Unpublished doctoral dissertation, Katholieke Universiteit Leuven, Leuven, Belgium.

Higgins, R. (2000). "Be more critical!": Rethinking assessment feedback. In *Paper presented at the British Educational Research Association Conference.* Cardiff University.

Hounsell, D. (1987). Essay writing and the quality of feedback. In J. Richardson, M. W. Eysenck, & D. W. Piper (Eds.), *Student Learning: research in education and cognitive psychology* (Milton Keynes: Open University Press.

Kember, D. (2003). To control or not to control: The question of whether experimental designs are appropriate for evaluating teaching innovations in higher education. *Assessment & Evaluation in Higher Education, 28,* 89-101.

Kim, M. (2005). *The effects of the assessor and assessee's roles on preservice teachers' metacognitive awareness, performance, and attitude in a technology-related design task.* Unpublished doctoral dissertation. Florida State University.

Pope, N. (2001). An examination of the use of peer rating for formative assessment in the context of the theory of consumption values. *Assessment & Evaluation in Higher Education, 26,* 235-246.

Pope, N. K. L. (2005). The impact of stress in self- and peer assessment. *Assessment & Evaluation in Higher Education, 30,* 51-63.

Rust, C., Price, M., & O'Donovan, B. (2003). Improving students' learning by developing their understanding of assessment criteria and processes. *Assessment & Evaluation in Higher Education, 28,* 147-164.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in education, 5,* 77-84.

Sadler, P. & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11,* 1-31.

SAS Institute Inc. (2004). *SAS/STAT 9.1 User's Guide*. Cary: SAS Institute Inc.

Searby, M. & Ewers, T. (1997). An evaluation of the use of peer assessment in higher education: A case study in the School of Music, Kingston University. *Assessment & Evaluation in Higher Education, 22,* 371-383.

Sitthiworachart, J. & Joy, M. (2003). Deepening computer programming skills by using web-based peer assessment. In *Proceedings of the 4th Annual Conference of the LTSN Centre for Information and Computer Sciences.* NUI Galway (Ireland): LTSN-ICS.

Sluijsmans, D. (2002). *Student involvement in assessment. The training of peer assessment skills.* Unpublished doctoral dissertation. Open Universiteit Nederland, Heerlen.

Topping, K. J. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy, & E. Cascallar (Eds.), *Optimizing new modes of assessment: In search of qualities and standards* (pp. 55-87). Dordrecht: Kluwer Academic.

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

Tsui, A. B. M. & Ng, M. (2000). Do secondary L2 writers benefit from peer comments? *Journal of Second Language Writing, 9,* 147-170.

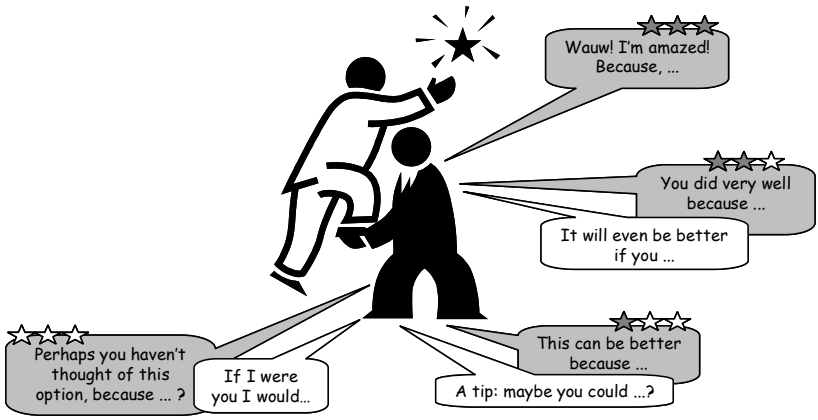
Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15,* 179-200.

Appendix

Feedback Form A

The 'critical friend'-assignment

Help each other
to make (even) beter assignments



Critical friend: Author of the essay:

For each criterion color the right amount of stars and explain what was right or could be better and why. Always suggest what your peer could do to improve his/her assignment. Work thoroughly and elaborately; be critical, honest and subtle!

CRITERION 1
What did he/she do well and why?

What didn't he/she do well and why?

If I were you I would ... , Maybe you could ... , It would even be better if you ...

CRITERION 2
What did he/she do well and why?

What didn't he/she do well and why?

If I were you I would ... , Maybe you could ... , It would even be better if you ...

CRITERION 3
What did he/she do well and why?

What didn't he/she do well and why?

If I were you I would ... , Maybe you could ... , It would even be better if you ...

CRITERION 4

What did he/she do well and why?

What didn't he/she do well and why?

If I were you I would ... , Maybe you could ... , It would even be better if you ...

CRITERION 5
What did he/she do well and why?

What didn't he/she do well and why?

If I were you I would ... , Maybe you could ... , It would even be better if you ...

CRITERION
What did he/she do well and why?

What didn't he/she do well and why?

If I were you I would ... , Maybe you could ... , It would even be better if you ...

Reply Form B

Letter to the reader of my essay

Dear reader,

I invite you to read my essay entitled

From the comments of my critical friend, I particularly remember that

By being a critical friend myself, and assessing the essay of somebody else, I learned that

After the 'critical friend-assignment' I revised my essay
1. with regard to (criterion)
   because
   and I tried to solve this by
2. with regard to (criterion)
   because
   and I tried to solve this by
3. with regard to (criterion)
   because
   and I tried to solve this by

My best piece is, in my opinion,
because

I paid this time special attention to
since

I hope you'll enjoy reading my essay

Kind regards,
(name)

Question Form C

Letter to my critical friend

Dear reader,

I invite you to have a look at my draft essay entitled

I tried to pay special attention to
by
Do you think it worked?

I still doubt on
1.
because
2.
because
3.
because

Most difficult, in my opinion, was
since
I'm curious what you think of it, and whether you can give me useful tips

I especially want to ask you to pay attention to the following criteria
1.
because
2.
because
3.
because

Finally, I want to remark that

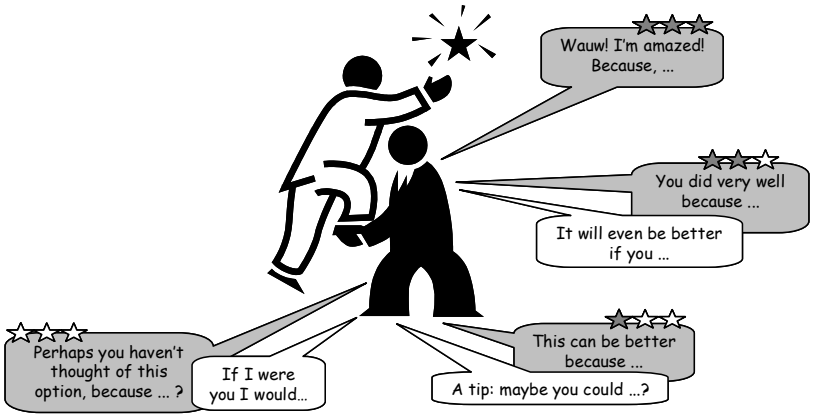
I hope you can give me useful tips, so I can improve my essay,

Kind regards,
(imaginary name)

Feedback Form D

The 'critical friend'-assignment

Help each other
to make (even) beter assignments



Critical friend: Author of the essay:

Work thoroughly and elaborately; be critical, honest and subtle!

After reading your essay, I first want to praise you because you

You said that you paid special attention to
Looking at your work, I think
since,

You said you had doubts on
1.
Looking at your work, I think
because,
2.
Looking at your work, I think
because,
3.
Looking at your work, I think
because,

You said that you found it difficult to
Looking at your work, I think
because,
In my opinion, it will even be better if you

Is it possible that you did not yet think of
since
Maybe you might solve this by

You asked feedback on the following criteria:
(1 star= can be better, 2 stars=well done, 3 stars=I'm amazed!)

CRITERION …

What did he/she do well and why?

What didn't he/she do well and why?

If I were you I would ... , Maybe you could ... , It would even be better if you ...

CRITERION …
What did he/she do well and why?

What didn't he/she do well and why?

If I were you I would ... , Maybe you could ... , It would even be better if you ...

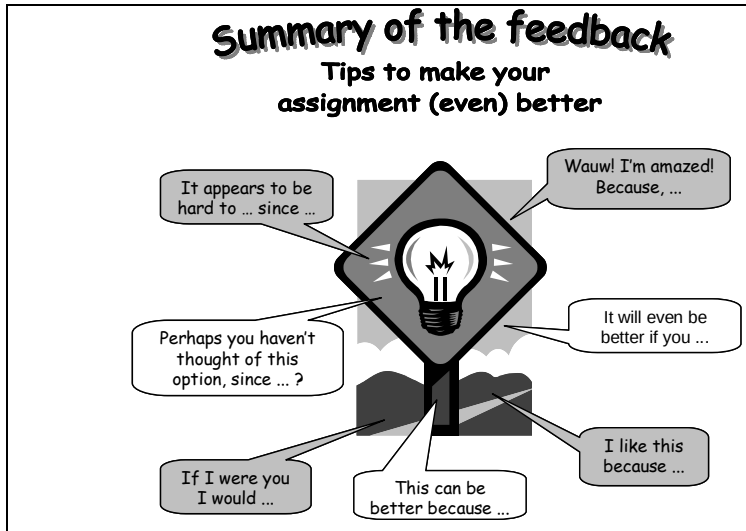
CRITERION …
What did he/she do well and why?

What didn't he/she do well and why?

If I were you I would ... , Maybe you could ... , It would even be better if you ...

Finally, I want to add that

Teacher Feedback Form



When reading your assignments, I noticed the following:

CRITERION 1
An example of how it should be, and why this is the right way to do it:

Frequent mistakes and why they are wrong:

Tips, points of special interest, suggestions:

CRITERION 2
An example of how it should be, and why this is the right way to do it:

Frequent mistakes and why they are wrong:

Tips, points of special interest, suggestions:

CRITERION 3
An example of how it should be, and why this is the right way to do it:

Frequent mistakes and why they are wrong:

Tips, points of special interest, suggestions:

CRITERION 4
An example of how it should be, and why this is the right way to do it:

Frequent mistakes and why they are wrong:

Tips, points of special interest, suggestions:

CRITERION 5
An example of how it should be, and why this is the right way to do it:

Frequent mistakes and why they are wrong:

Tips, points of special interest, suggestions:

CRITERION 6
An example of how it should be, and why this is the right way to do it:

Frequent mistakes and why they are wrong:

Tips, points of special interest, suggestions:

CHAPTER 6

THE EFFECTS OF CONSTRUCTIVENESS OF PEER FEEDBACK ON PERFORMANCE

Abstract

This study examines three assumptions about peer feedback. The first is based on the 'assessment for learning' principle and is the expectation that receiving 'rich, constructive' feedback will improve performance. The second is based on the 'assessing for learning' principle and is the expectation that writing 'rich, constructive' feedback is a learning experience too and will also improve performance. Finally, the third assumption is based on the concepts of 'mindful reception' and 'closed feedback loops' and is the expectation that formative peer assessment has a larger impact on performance when feedback addresses students' personal needs (condition 1) or when students are stimulated to reflect and act upon feedback (condition 2).

A group of 68 first year students in secondary education experienced formative peer assessment for three successive writing assignments. They were divided in two experimental conditions and a control condition with plain peer feedback. Their progress in writing performance is examined against the constructiveness of the peer feedback they gave and received, and against the condition in which they participated. The effect of the constructiveness of feedback is studied from two directions: from the point of view of the receiver of the peer feedback ('assessment for learning') and from the point of view of the assessor who gave peer feedback ('assessing for learning').

The results show a significant positive effect of the composition of the received peer feedback on student performance. The constructiveness of feedback that students provided themselves was not found to improve their learning. Nevertheless, the overall level of constructiveness of the feedback was low. Possible barriers preventing students from providing good feedback, and solutions to these, are discussed in the paper. Finally, the study could not replicate the effect of condition that was found in an earlier study.

THE EFFECTS OF CONSTRUCTIVENESS OF PEER FEEDBACK ON PERFORMANCE

Introduction

Formative Assessment to Support Learning

Assessment has an influence on learning. The best known influence of assessment on learning is due to the activity of looking back after the completion of an assessment task, referred to as "post-assessment effects" (Gielen et al., 2003). Post-assessment effects deal with how judgements about the quality of student performances shape, and hopefully improve, the students' competence by short-circuiting the randomness and inefficiency of trial-and-error learning (Sadler, 1989). Feedback is the most important trigger for these post-assessment effects to occur.

Mory (2003) discusses different perspectives on the use of feedback to support learning, associated with different learning paradigms. All of these perspectives consider assessment as a tool for learning (see Dochy et al., 1997; Gielen, Dochy, Onghena, Struyven, Smeets, and Decuyper, 2007). They differ, however, in their definitions of the desired learning effects, and in their views on how feedback stimulates them. Firstly, feedback can be considered as a motivator or an incentive for increasing a certain response rate and/or accuracy. Secondly, feedback may act to provide a reinforcing message that would automatically connect responses to prior stimuli - the focus being on correct responses. Thirdly, feedback can provide information that learners can use to validate or change a previous response - the focus falling on erroneous responses. Finally, feedback may focus on the provision of intellectual tools to help students to construct their own internal reality and to enable them to analyse their own learning processes. Within this last stance, associated with the constructivist paradigm, "feedback might also occur in the form of discussion among learners and through comparisons of internally structured knowledge" (Mory, 2003, p. 772).

Feedback comes in different shapes. In some cases it is just 'knowledge of results', whereas in other cases it is much more sophisticated. In her study, Narciss (1999) describes two important components of a feedback message; the 'evaluative component', indicating whether the answer or solution is right or wrong, and the 'informational component', providing additional information about the task or solution. Feedback messages largely differ in the volume of this informational component, and these differences appear to be related to their effectiveness in altering performance (Narciss, 1999). Wiliam (2006) shows different interpretations of this informational component: One might diagnose the problems identified in the evaluative component, but one might also go further and give the learner something to work with.

Mediated intentional feedback has, under certain conditions (see also Kluger et al., 1996), a strong positive effect on learning. This was shown convincingly in the meta-analysis of Bangert-Drowns, Kulik, Kulik, and Morgan (1991). When assessment is intentionally designed to provide rich feedback and support learning, it may be called 'assessment for learning' or 'formative assessment'. In their review study, Black and Wiliam (1998) confirmed the undeniable power of formative assessment in education. The benefits of 'assessment for learning' and feedback may be large. Wiliam (2006) contends that "supporting teachers in developing the use of assessment for learning has been shown to roughly double the speed of learning (…). In other words, students learned in six months what would have taken a year to learn in other classrooms" (p. 7).

A Rationale for Peer Feedback

The constructivist perspective on feedback ("feedback might occur in the form of discussion among learners") brings us to the rationale of formative peer assessment and peer feedback. The mere fact that feedback comes from a peer, and not from a "knowledge authority", alters its meaning and impact on learning (Topping, 2003). The uncertainty of a peer's authority stimulates a "mindful" reception of the feedback, identified by Bangert-Drowns, Kulik, Kulik and Morgan (1991) as crucial for the instructional benefit of feedback. This is also shown in the study of Yang, Badger, and Yu (2006), which found that revisions based on teacher feedback were less successful than peer-initiated revisions. The main reasons for unsuccessful

revision based on teacher feedback were misinterpretation and miscommunication. Peer feedback was received with many more reservations, leading to discussions amongst learners about its interpretation. As a result, students acquired a deeper understanding of the subject. Moreover, the study of Yang and colleagues also revealed that students who received peer feedback showed more initiative for self-correction. Reservations about the feedback students received from peers stimulated them to look for confirmation in some way, by checking grammar books or asking the teacher, and to develop their own independent ideas for revision. In contrast, exposure to teacher feedback lowered their initiative for self-regulated learning, perhaps because students believed that the teacher had pointed out all their mistakes and there was no need for further correction (Yang et al., 2006).

Moreover, the confrontation with different approaches and solution strategies of peers offers intellectual tools for further knowledge construction. Peer feedback can provoke discussion amongst learners, and reflection by learners about their own internal construction of knowledge. If this construction contains misconceptions, constructivists assume they cannot be corrected from outside. They can only be challenged by a confrontation with different perspectives from others, which may consequently lead towards an internal change. But even if the learner's own internal reality is successful in solving certain problems – and it might thus be considered 'correct' – it can also be enriched by becoming more flexible through the comparison with alternative realities of equal value held by peers.

A Definition of Constructive Peer Feedback

Based on the above described perspective on how peer feedback will function most effectively, constructivists expect a peer assessor to confront a learner's product or conduct with his own reality of understanding, to indicate what he thinks fits or does not fit in (instead of comparing it to an outside objective norm) and, because the norm is internal, to explain and justify why he considers things as correct or wrong. Moreover, he should suggest competing, or equivalent, alternatives to challenge the learner's reality. Finally, adding open questions is encouraged, to provoke reflection by the learner. These requirements deal with the composition of the 'informational component' (Narciss, 1999) of the feedback message.

The Bidirectional Impact of Peer Feedback on Performance

Previous research on the impact of the composition of peer feedback on performance is scarce. One earlier study, a study by Kim (2005), was found. She found no increase in performance of assessees who received higher quality peer feedback. Quality is defined in this study as providing marks and feedback for each content criterion with a rationale or a specific suggestion for revision. Two explanations for this lack of impact on performance are provided. The first reason why a direct effect was not observed may be that the difference in the quality of peer feedback across students, expressed by a score out of 10, was minimal (M=5.9, SD=.53). The second explanation is that students' sceptical attitudes toward their peers' abilities may have prevented assessees from internalizing peer feedback. It should be noted that this is exactly the argument that Yang et al. (cf. supra) used to explain a *higher* impact on performance. Although these arguments are seemingly contradictory, in fact they are not. The reservations which assessees had about peer feedback in the study of Yang et al. made them initiate discussions and self-corrections. These discussions and self-corrections in their turn led to successful revisions. These revisions, however, will not have been directly correlated to the composition of feedback (as Kim found indeed), but rather to the critical attitude of the assessee which was higher in a peer feedback situation compared to a teacher feedback situation. The impact of peer feedback on performance was thus indirect, mediated by a mindful approach to it. Whether or not the assessee will adopt a mindful approach, however, is not expected to be related to the quality of the feedback he receives.

So far, we have only discussed the possible impact of peer feedback on learning from the perspective of the assessee. This is the most obvious perspective when making comparisons with staff feedback. The underlying process is the well known 'assessment for learning' (Taras, 2002), also referred to as learning-oriented assessment (Carless et al., 2006) or formative assessment (Black et al., 1998). In this use of assessment, the learning of the assessee is central.

However, in peer feedback, the assessor is also a learner. This particularity of peer feedback gives rise to two extra ways in which learning can be augmented. A student may learn from assessing someone and providing extended feedback; and the social interaction that may follow

between peers as a consequence of the assessment may also enhance learning (see also Gielen et al., 2007).

Assessing for learning. The learning of the *assessor* might be strengthened through the peer assessment activity (e.g., Pryor & Lubisi, 2002; Topping, 1998). Reasons for these learning effects are twofold. Firstly, students discover interesting ideas or alternative approaches to the task when reading others' work or observing others' performances, and will incorporate these in their own work. In addition, they will probably also detect some weaknesses or mistakes by the others that will stimulate self-reflection and probably lead to a correction of similar flaws in their own work. Secondly, Pryor and Lubisi (2002) mention that the assessment activity engages students to cognitively operate at an evaluative level and to pose metacognitive questions. These are higher order learning activities that help the assessor acquire a deeper insight into the subject. Sluijsmans and Prins (2006) also found that training student assessors in assessment skills has positive effects on their development of content related skills. An explanation for this is provide by Stiggins (1991, p. 38): "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".

Peer learning. Finally, peer learning processes may arise from the interaction between peers that is elicited by the feedback process, provided that there is room for discussion to compare differences in opinions and to look for implications and solutions together. In his review of peer learning, Topping (2005) explicitly names peer assessment as an extension of the forms of peer learning from traditional peer tutoring and cooperative learning.

The Importance of Training, Guidance and Quality Control

Since students are no expert assessors, and since it is shown that feedback can have many faces but only some of these faces are expected to be effective, it is clear that peer feedback will not realise its full potential without some training, guidance or quality control to enhance the quality, or constructiveness, of peer feedback.

Nilson (2003) identifies three possible barriers that reduce the quality of peer feedback. The first is the intrusion of students' emotions into

the evaluative process. In her practice "students do not want to be responsible for lowering a fellow student's grade. In addition, they may fear that: 'If I do it to them, they'll do it to me' or they may be concerned that giving insightful critiques may raise the instructor's grading standards" (p. 35). Guidance of students into an understanding of the rationale of peer feedback is an important issue in the peer assessment procedure. Students should be convinced that their comments will, instead, give the peer a chance to increase his grade by reviewing the weaknesses.

A second barrier may be a lack of motivation. Nilson (2003) calls it 'laziness in studying the work and/or writing up the feedback'. Convincing students that they can help each other, as described above, might raise the feeling of 'individual accountability' and 'positive interdependency' (see Slavin, 1989; Sluijsmans, 2002) and function as an internal motivator to make the effort. Performing quality checks and rewarding engagement of the student by the teacher are other remedies, which can serve as extra external stimulators in the peer assessment procedure (e.g., Searby et al., 1997; Sitthiworachart et al., 2003).

A final barrier identified by Nilson (2003) is ability. Are students able to give good, constructive feedback? Sluijsmans (2002) emphasises that conducting a peer assessment is a complex skill, in which students have to be guided. Also Hanrahan and Isaacs (2001) identified in her qualitative analysis that students are uncertain about their ability to conduct peer assessment, and they feel discomfort because they do not trust the ability of their peers to assess them. Sluijsmans (2002) states: "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" (p. 21).

The Importance of Supporting Mindful Reception and Closing the Feedback Loop

Gibbs and Simpson (2004) emphasise that it is important to address students' response to feedback explicitly. Feedback that is not attended to or not acted upon can not be effective. In their description of conditions under which assessment and feedback support learning, they suggest several 'tactics' to address this issue. A first tactic is to provide only feedback to those aspects that students request. It is hoped that this measure motivates

and guides assessors to give useful feedback, and that assessees feel more personally addressed by the feedback, supporting a 'mindful reception' of it (Bangert-Drowns et al., 1991). A second tactic is to ask students to demonstrate how they used the feedback in their revisions. This tactic aims at closing the "feedback loop" (see Boud, 2000). The effectiveness of these tactics is shown in the study of Gielen, Dochy, Onghena, and Smeets (2007), in which both tactics proved to be able to raise the performance of secondary school students on their writing assignments, compared to a group of students that used peer feedback without these extra requirements.

Aim of this study

The present study is an extension of the research conducted by Sluijsmans (2002) and the research conducted by Kim (2005). In her experiments, Sluijsmans investigated both the composition of the peer feedback report (as an indicator of peer assessment skill) and student performance. Her focus was the comparison of a group which received peer assessment training with a control group, and the comparison between a pretest before and a posttest after the training. She did not, however, relate the composition of the peer feedback directly to the performance, to study the relationship between them. This relationship will be the main focus of this paper. Instead of comparing students' performances to a control group, or to a prior measure, we will try to relate it to individual differences in the composition of the feedback; that is the level of constructiveness of feedback. The study by Kim also addressed this relationship, but only from the point of view of the assessee, while the current study will also examine it from the perspective of the assessor.

Two hypotheses are examined. The first is based on the 'assessment for learning' principle and the expectation is that receiving 'rich, constructive' feedback will improve performance. The second is based on the 'assessing for learning' principle and the expectation is that writing 'rich, constructive' feedback is a learning experience too and will also improve performance. The third principle discussed in the literature as an explanation of the effectiveness of peer feedback, namely peer learning, will not be examined in this study since there will be no direct face-to-face interaction

between the peers concerning the feedback, and thus no chance for peer learning.

Two additional hypotheses address the measures to encourage the response of students to peer feedback. In the third hypothesis we will try to replicate the finding of Gielen, Dochy, Onghena, and Smeets (2007) that specific design features of peer feedback which are aiming at this goal (an a priori question form and a posteriori reply form), are able to increase student performance.

In a fourth hypothesis, finally, one of the explanations offered in the previous study (Gielen et al., 2007) for the success of the question form will be verified. It was assumed that the success of the question form was partly due to its motivating effect on the assessor. Answering to self-indicated needs of the assessee might appear more relevant to the assessor than just commenting on all aspects of an essay. As a consequence, it is expected that this feedback will be more extended in terms of suggestions and justifications.

Hypotheses

Based on the aforementioned literature, we formulate the following research hypotheses:

1. Students who received more constructive peer feedback will have a higher increase in performance (after revision of their essay) then students who received less constructive peer feedback.
2. Students who wrote more constructive peer feedback will have a higher increase in performance (after revision of their essay) then students who wrote less constructive peer feedback.
3. Students in conditions with an extra measure to encourage students to respond to the feedback will have a higher increase in performance (after revision of their essay) than students in a plain peer feedback condition without these extra measures.
4. Students in the condition that uses the question form will write more constructive peer feedback than students in the conditions without the question form.

Method

Participants

A total of 68 first grade students (12-13 years old) participated. They were divided in three different classes from the same secondary school, taught by the same teacher. All were enrolled in the theoretically oriented *general secondary education* track (ASO).

Peer Assessment Design and Procedure

The description of the particularities of the design and procedure of the peer assessment in this study will be structured by means of the inventory of peer assessment diversity (see Gielen, Dochy, & Onghena, 2007) in Table 1.

Table 1

Description of the current peer assessment (PA) design and procedure

Cluster	**Variable**	**Description of the current design and procedure**
Cluster I Decisions concerning the use of peer assessment	Setting	Educational use, Dutch writing curriculum (students' mother tongue), formal learning, class assignments during 2nd and 3rd trimester, 1st year general secondary education, 65% male, 12-13 years old, class sizes: 22, 21 and 25, all classes same teacher
	Object	Artefact: several types of essays (a story, a newspaper article, a reader's letter) Type of performance expected of learner: creative writing Information taken into account: writing performance (outcome) Draft version of a two-stage assignment
	Frequency & Experience	Frequently: PA of 3 successive assignments Novel to students
	Objectives	Tool for learning and learning-how-to-assess
	Function	Formative
Cluster II Link between peer assessment and other elements in the learning environment	Alignment	The learning goals of writing essays are central to the Dutch curriculum. Students assess the 'normal' writing assignments of the course, no 'extra' products. Normally, these assignments are not two-stage, so the intermediate feedback is considered extra. PA is also new to teachers, so the alignment with teaching practices is not perfect: teachers

		struggle with the time scheme, with the feedback forms and with the practical arrangements of assessors or assessees being absent or not prepared
	Relationship to other assessments	Peers are the only assessors of the drafts. After revision the teacher assesses the final versions: mainly summatively, with limited and delayed feedback
	Scope of involvement	Aspects & extent of involvement: participation in development of assessment criteria, responsibility for formative judgements for each criterion, responsibility for providing of knowledge of results (on 4-point-scale per criterion) and feedback
Cluster III Interaction between peers	Output	Nature of information: quantitative (4 point scale) and qualitative Extent of 'condensation': at level of single criteria Feedback stance: depending on student (prompts for evaluative remarks + collaborative suggestions)
	Directionality	Unidirectional
	Privacy	No anonymity of assessor/assessee Output confidential (as regards third, not teacher & researcher)
	Contact	Writing feedback starts in class (e.g., last 20 minutes of a lesson) and is finished at home Output provided in writing, read at home
	Role of assessee	Class 1: active by revision (=PA group) Class 2: active by revision and reply to teacher (=PA-REPLY group) Class 3: active by questions for assessor & revision (=QUEST-PA group)
Cluster IV Composition of the assessment groups	Matching	Principle for matching: same ability (based on pretest: writing exam December) Responsibility for matching: researcher Consistency of matching: fixed
	Constellation of assessors & assessees	Unit of assessor: individual Unit of assessee: individual Number of assessors per unit of assessee: 1 Number of assessees per unit of assessor: 1
Cluster V Management of the assessment procedure	Format	Fixed format (see Appendix): Class 1: form A, class 2: forms A and B, class 3: forms C and D. *Feedback form A:* paragraph per criterion (total=6), prompts for strengths + justification, for weaknesses + justification, for suggestions and for a quantitative judgement (colour 0-1-2-3 stars in). *Reply form B with following prompts:* By receiving/giving feedback I learned …, I revised my work on the following criteria …,

		My best piece is …, I paid special attention to … *Question form C with following prompts:* I paid attention to …, I doubt …, I found it difficult to …, I wish for feedback on the following criteria … *Feedback form D:* prompts for assessor's opinion on all aspects mentioned in the question form, plus one paragraph for each requested criterion (max=3) with prompts for strengths + justification, for weaknesses + justification, for suggestions and for a quantitative judgement (colour 0-1-2-3 stars in)
	Requirement	Compulsory for assessor/assessee
	Reward	None
	Training/ Guidance	Explanation of the rationale of peer feedback and of the requirements of the feedback form that uses guiding prompts. Discussion of a worked out example of the peer assessment process. During PA, help if students do not know what to write
	Quality control	None

Research Design

The present study adopts a repeated measures pretest posttest design (see Figure 1). The performance in the draft version of an essay is the pretest measure, and the performance in the revised version is the corresponding posttest measure. This process of data collection is repeated over three successive assignments.

The composition of the peer feedback is represented by its Feedback Constructiveness Index score (see below). These scores are obtained for each of the three assignments. Each feedback form (and associated FCI score) is used in two directions: firstly as a measure of the composition of the received feedback (from the point of view of the assessee, cf. 'assessment for learning'); secondly as a measure of the composition of the feedback given to a peer (from the point of view of the assessor, cf. 'assessing for learning').

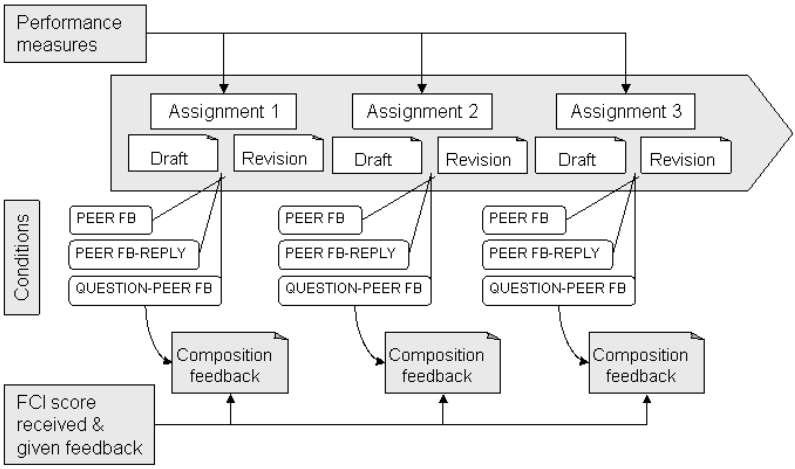


Figure 1. Representation of the research design.

Variables

Performance. For the measures of performance, research assistants rated the quality of the draft and final essays. They used a scoring protocol to make a valid assessment of performance, based on the class-defined criteria for the essay assignments. Every criterion was judged on certain required aspects, and scores were added to obtain a general appreciation of the draft and the final essay. The maximum score is 12. Inter-rater reliability is not expressed as a score, but it is assured by double-checking of every assessment by two researchers. In the analyses, a difference score is used, that is obtained by subtracting the pretest score from the posttest score.

Composition of the peer feedback. The level of constructiveness of the peer feedback was measured by an adaptation of the Feedback Quality Index by Prins, Sluijsmans and Kirschner (2006), based on various versions of earlier 'rating forms' by Sluijsmans (2002). The original instrument is formulated at the level of a feedback report, whereas our adaptation is formulated at the level of the feedback paragraph that comments on one particular content criterion, and is averaged afterwards. The latter approach is similar to that adopted by Kim (2005). It is the result of the difference

between the peer assessment procedure used by Prins et al. (2006) and Sluijsmans (2002) and the procedure in this study, which used a feedback form with pre-structured feedback paragraphs. To avoid confusion, we refer to our instrument as the Feedback Constructiveness Index.

The written feedback is evaluated against a list of 'peer assessment criteria' defined by Sluijsmans and colleagues regarding the use of adequate criteria, giving feedback and the style of a written assessment report. These criteria represent the level of constructiveness of a feedback message. As a result of the different structure of our feedback form, we did not count the number of specific aspects or items in the feedback, but checked for the presence of each in every feedback paragraph. A feedback form contained comments on three to six content criteria – depending on the condition – , each of which was scored independently. The comments of a peer in such a feedback paragraph are scored against the 'peer assessment criteria' presented in Table 2, resulting in a maximum score of 14 per paragraph. Again, inter-rater reliability is not expressed as a score, but it is assured by double-checking of every FCI score by two researchers.

Table 2

Peer assessment criteria of the Feedback Constructiveness Index

PA criterion	**Instruction for scoring**	**Points**
Readable	Is the feedback readable and understandable? Is the text structured logically and is it formulated in comprehensible sentences (no text message language, symbols, private abbreviations, etc.)?	1
Specificity	Is it indicated to which text fragment a comment is related? Does the assessor mention concrete examples from, or references to, the piece of work?	1
Appropriate	Is the feedback connected to the content criterion of the feedback paragraph?	2
Positive & Negative	Does the assessor incorporate both positive and negative aspects in the feedback? The numbers of positive and negative comments do not need to be in balance, but both need to be present. A negative comment is not required only if the researcher agrees that there is nothing negative to remark on. Feedback does not have to be complete; a peer assessor may miss an aspect without consequences. Comments need to be related to the content criterion (e.g., a spelling remark does not count in a paragraph on structure).	2
Justification	Does the assessor explain and justify a reason for which something can remain the way it is (positive remark), or should be changed (negative remark)? At	4/2

	least one meaningful and clear justification of the feedback, or a part of it, should be present to get the maximum score. If only a short explanation (written without much effort) is present, half of the maximum score can still be awarded. Simply repeating the content criterion is not sufficient. The explanation does not have to be correct. Although this is preferable, it was decided not to punish students who try and fail. For a similar reason, it was decided not to require every comment to be justified to comply with this 'peer assessment criterion'. Again this would be preferred, but is omitted since this would be to the advantage of students with short feedback (they do not have to justify much), and to the disadvantage of students who made an effort to comment extensively and will probably omit an explanation somewhere.	
Suggestion	Does the assessor suggest possibilities for improvement (when a negative comment is made) to help the assessee? At least one useful suggestion to improve the piece of work is expected.	2
Reflective question	Does the assessor formulate questions in the feedback to invite the other to think more deeply about their own piece of work? These should not be rhetorical questions, but ones to which the assessor does not know an immediate answer. At least one thought-provoking question is expected.	2

Condition. The study consists of three conditions using a version of peer feedback. Conditions are assigned at class level, randomly. The difference between the conditions lies in the role of the assessee (see Table 1), which is in two of the three conditions extended with an extra requirement to encourage the assessee to respond to the feedback. In the first 'extended condition' (the 'PA-REPLY group') students are asked to report – in a written reply to the teacher – which feedback comments they took into account and how, what they learned from giving peer feedback, and what they think of their own accomplishments. In the other 'extended condition' (referred to as the 'QUEST-PA group'), students are additionally required to complete a 'question form' in which they indicate to their peers on which aspects or criteria they think to have some problems and for which they request feedback. They can also add questions to their peers. The third condition is a 'plain' peer feedback condition (referred to as the 'PA group'), as described in Table 1.

The 'plain' peer feedback condition and the PA-REPLY condition share the same feedback form (see Appendix and its description in Table 1). The feedback form in the condition QUEST-PA condition is also slightly adapted from the PA-feedback form, in that it is tries to direct peers to address the questions and comments of the assessee. They are expected to comment mainly on the requested criteria, not on the whole list (see Appendix and Table 1).

Analyses

The dataset contains different assignments by the same student. Therefore, a multilevel approach will be used. Students' essays are located at the first level. The second level is the student. For the descriptive data analysis of feedback constructiveness, one extra level has been identified, located below the essay level; namely the level of the feedback paragraph that discusses one particular content criterion of the assignment. The presence or absence of certain features of constructive feedback was originally measured at this level. All statistical analyses were executed by means of the SAS System (SAS Institute Inc., 2004).

Results

Descriptive Analysis

Performance. The summary statistics for the performance measurements, a subtraction of the pretest score from the posttest score, are presented in Table 3.

Table 3

Summary statistics for the variable performance (difference score between draft and final)

Measurement	Minimum	Mean	Maximum	N	Std Dev
1	-1.50	0.93	3.00	63	0.94
2	-1.00	0.70	4.00	61	0.85
3	-1.50	0.90	4.00	67	1.06

Composition of feedback. The Feedback Constructiveness Index, representing the richness of the composition of the peer feedback, shows that

the overall quality of the peer feedback is rather low. The overall mean score is 5.9 with a standard deviation of 1.68, while the scale of the FCI goes from 0 to 14. The summary statistics of the FCI score per measurement are presented in Table 4.

Table 4

Summary statistics for the variable composition of feedback

Measurement	Minimum	Mean	Maximum	N	Std Dev
1	3.50	6.32	10.50	61	1.54
2	1.50	5.79	9.00	62	1.67
3	1.60	5.63	10.00	67	1.75

The fourth hypothesis addresses the differences in composition of feedback per condition. Therefore, the summary statistics per condition are included in Table 5. The mean value on the FCI scale in the condition that uses the question form is indeed higher than the mean values in the other two conditions.

Table 5

Summary statistics for the variable composition of feedback per condition

Condition	Minimum	Mean	Maximum	N	Std Dev
PA	1.60	5.78	10.50	63	1.66
PA-REPLY	1.50	5.74	9.00	57	1.56
QUEST-PA	1.67	6.14	10.00	70	1.77

Examining the feedback composition in detail reveals that the components going beyond a mere evaluative stance (giving suggestions for improvement, justifying a judgement or posing an open thought-provoking question) are rarely present in the peer feedback. This is surprising since the feedback form gave explicit prompts, in each feedback paragraph, to add a justification as well as a constructive suggestion.

Table 6 presents the frequency and percentages of suggestions and justifications within feedback paragraphs. Feedback paragraphs are sections on the feedback form that invite the assessor to comment on one specific content criterion.

In theory, we should have had 774 feedback paragraphs for the PA and PA-REPLY conditions (43 students * 3 assignments * 6 criteria), and we could have had an additional 225 feedback paragraphs for the QUEST-PA

condition (25 students * 3 assignments * maximum 3 criteria). In total 84 are missing, as a consequence of 14 assignments or feedback forms that were not handed in. Of the 720 feedback paragraphs that were analysed in the first two conditions, only 16 feedback paragraphs did not contain any feedback at all. This means that in these conditions, most students indeed commented on every criterion that was included in the feedback form. In the third condition, the number of paragraphs containing feedback depended on the requested criteria by the assessee. In total, 57 paragraphs did not contain any feedback.

However, of the remaining 872 feedback paragraphs in the three conditions, 58% contained no suggestion for improvement and 53% contained no justification for the feedback given. In 39% a short explanation was present, and in 8% students really provided an extended explanation why something was good or bad. A reflective question was found in just one paragraph out of these 872.

Table 6

Frequency table (including percentages) for the presence of a suggestion and/or justification in a feedback paragraph.

	Suggestion		
Justification	0	2	Total
0	355	108	463
	40.7 %	12.4 %	53.1 %
2	156	181	337
	17.9 %	20.8 %	38.6 %
4	15	57	72
	2 %	7 %	8 %
Total	506	346	872
	58.0 %	39.7 %	100 %

Effect on Performance

Receiving feedback. The question firstly addressed is whether the composition, more specifically the level of constructiveness, of the received feedback explains some of the variation in the increase in performance on the final version of a paper, compared to the performance on the draft version.

The analysis shows that there is indeed a significant effect of the composition of the received feedback on the performance ($F(1,118)=6.8$, $p=0.010$).

Giving feedback. Restructuring the dataset enables us to relate the FCI score of the assessor to his own performance scores for the assignment. Examination of the relationship between the two, by means of a hierarchical linear model, shows no significant association with the composition of the given feedback ($F(1,111)=0.30$, $p=0.588$).

Condition. Finally condition also yields no significant effect on performance scores ($F(2,123)=0.84$, $p=0.435$). Comparing the extended peer feedback groups against the baseline group with plain peer feedback (PA-REPLY versus PA and QUEST-PA versus PA) also results in non-significant outcomes ($t(123)=1.29$, $p=0.198$ and $t(123)=0.69$, $p=0.490$).

Effect of Condition on the Level of Constructiveness of Feedback

To investigate the fourth hypothesis, the FCI score is taken as a dependent variable, and the condition as a predictor. More specifically, the contrast of the QUEST-PA group versus the PA AND PA-REPLY groups is tested. Although the estimated parameter indeed shows that the average FCI score of the QUEST-PA group is 0.7 points higher, this effect is not significant, $t(122)=1.04$, $p=0.298$.

Conclusion and discussion

Based on the available literature, a positive effect of the constructiveness of the received feedback was expected on performance. This study indeed showed a significant effect. If the received feedback was more specific, more appropriate to the assessment criteria, contained positive as well as negative comments, and in addition included some justifications, suggestions and thought-provoking questions, the assessee made better revisions, resulting in a higher progress between the draft and the final version of the essay. These results are in line with earlier reports on the importance of the type of information that is included in feedback (e.g., Bangert-Drowns et al., 1991; Flower, Hayes, Carey, Schriver, & Stratman,

1986; Narciss, 1999), and also provide support for the importance of the characteristics of constructive feedback that were part of the FCI.

It should be noted that the correctness or the completeness of the feedback was not part of the FCI. This shows that the impact of peer feedback can be enhanced by addressing other issues than only its validity and reliability, which has been the main focus in many previous studies (Falchikov & Goldfinch, 2000; Magin & Helmore, 2001).

This finding that constructiveness of feedback is important, together with the observation that there was still considerable room for improvement in the level of constructiveness of the peer feedback in our study, emphasise the need to pay attention to peer assessment training, guidance and quality control in a peer assessment procedure, in order to raise its learning benefits. The effectiveness of peer assessment training to raise the quality of feedback reports was already shown by Sluijsmans and Prins (2006) in the context of teacher education. They found considerable effects on the use of criteria, the presence of constructive comments and on the structure and language of the feedback reports. Although there was a training session included in the current peer assessment procedure, this was only limited in scope and might not have been sufficient to realise the full potential of peer feedback.

To compensate for this limited training, some of the issues that Sluijsmans and Prins addressed in their training were taken over by the prestructured feedback form in our setting: the need for structure was incorporated into the feedback paragraphs, which in the meantime also steered students to comment on every assessment criterion. Moreover, prompts for constructive comments (praise, critics, suggestions and justifications) were present. Although Miller (2003) found that giving explicit prompts to comment on different criteria stimulated students to do so – as was confirmed by our study –, the same trick did not work for justification and suggestion. The descriptive data analysis revealed that in almost two-thirds of the feedback paragraphs there was no suggestion or no justification present; in more than 40% neither was present. Reflective, open questions were absent throughout, except for one student in one assignment.

Thinking of a suggestion or a challenging question, or explaining and justifying your judgment are activities that require some effort by the assessor. Nilson (2003) suggests that a lack of motivation – he calls it "laziness in studying the work and/or writing up the feedback" – is an important barrier to good peer feedback. Two solutions to this problem are

possible. Firstly, taking account of the principles of 'social interaction', 'individual accountability' and 'positive interdependency' in the design of peer assessment may motivate students to do their best for each other (see Slavin, 1989; Sluijsmans, 2002). In this study, the principles of 'individual accountability' and 'positive interdependency' were built into the peer assessment procedure. It is possible, however, that not all students perceived them as such. An in-depth study of students' perceptions of this type of peer assessment design should bring clarity to this question. Introducing real social interaction by means of an oral face-to-face discussion of the feedback between peers may probably also help to intensify the feeling of individual accountability and positive interdependency. Van den Berg (2003) found that the proportion of explanatory comments (justification) or suggestions for revision increased substantially between written and oral feedback.

A second remedy to address a motivational problem is external quality control by the teacher. Teachers might reward students for good feedback or punish them for weak feedback (e.g., Searby et al., 1997; Sitthiworachart et al., 2003). This second remedy, however, is less preferable from the point of view of self determination theory, since it might be considered less supportive of autonomy (Deci et al., 1985).

The present study did not find support for the second hypothesis. Although the study of Sluijsmans and Prins (2006) found that their experimental group, after receiving peer assessment training, showed an improvement both in the quality of the peer feedback they wrote as in their content-related task performance, we could not identify a relationship between the latter two skills. Students do not perform better becàuse they have been more intensively working on their feedback to a peer, measured by means of the level of constructiveness of this feedback. Although there might have been a general 'assessing for learning' (Gielen, Dochy, Onghena, Struyven, et al., 2007; Topping, 1998) effect, it was not mediated by the extensiveness of the written feedback, as it was measured by the FCI. Further research is necessary to examine how the 'assessing for learning' effect of peer feedback can be maximised, in order to allow teachers to address these issues in their peer assessment procedure, training and guidance.

Finally, the differences in the conditions also yielded no significant effect on performance. This finding is in contrast with the earlier reported

finding in Gielen, Dochy, Onghena, and Smeets (2007), although the current study had more power (due to the repeated measures design) and would thus be able to detect even a smaller effect than detected by the previous study). Based on the available literature, we assumed that the question form would motivate assessors to provide better feedback and that it would raise assessees attention for the feedback since it would address personal needs. Furthermore, also the reply form was expected to improve performance, by stimulating the assessee to reflect and act upon the received feedback. None of these expected implications of the extended peer feedback conditions, however, were large enough to result in an increased performance compared to the plain peer feedback condition in the present study. The test of the fourth hypothesis showed that the provided feedback in the condition using the question form was indeed more constructive than in the other conditions, but this effect was not significant either.

A possible explanation for the absence of an effect from the QUEST-PA condition on performance might be that the effect of a more effective response is compensated with the effect of a broader scope of the feedback in the other conditions. In the QUEST-PA condition, students could only request feedback for maximum three criteria, while in the other conditions students almost always received feedback on all six criteria. This however was also the case in the previous study, and in that case the question form did realise an added value.

A possible explanation for the absence of an effect from the PA-REPLY condition on the other hand, might be that in the current peer assessment procedure all conditions where expected to revise their essay after the feedback. The reply form might thus have been redundant in stimulating students to act upon the feedback. However, again, this does not provide a sufficient explanation for the difference between the current and the previous study, since the feedback procedures, the setting and the characteristics of the participants were very similar between both studies.

In both studies, teachers indicated that they had experienced difficulties with including the peer assessment exercises into their time scheme, and that they struggled with the different feedback forms and practical arrangements for the different conditions. Although both teachers managed to find solutions to these difficulties, differences in the details of these solutions might be responsible for the different impact of the conditions

in both studies. We do not have sufficient information on these details, however, to get a clear view of the interaction that might have taken place.

To reach more insight in the causes of these contradictory findings, a new replication study of the impact of the extended peer feedback procedures is needed, which should also address a more in-depth investigation of how teachers and students use these forms and how they perceive them.

References

Bangert-Drowns, R., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61,* 213-238.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5,* 7-74.

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education, 22,* 151-167.

Carless, D., Joughin, G., & Mok, M. M. C. (2006). Learning-oriented assessment: Principles and practice. *Assessment & Evaluation in Higher Education, 31,* 395-398.

Deci, E. L. & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*.

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

Falchikov, N. & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70,* 287-322.

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

Gibbs, G. & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education, 1,* 3-31.

Gielen, S., Dochy, F., & Dierick, S. (2003). Evaluating the consequential validity of new modes of assessment: The influence of assessment on learning, including pre-, post-, and true assessment effects. In M. Segers, F. Dochy, & E. Cascallar (Eds.), *Optimizing new modes of assessment: In search of qualities and standards* (pp. 37-54). Dordrecht: Kluwer Academic.

Gielen, S., Dochy, F., & Onghena, P. (2007). An inventory of peer assessment diversity. In S. Gielen, *Peer assessment as a tool for learning* (pp. 67-94). Unpublished doctoral dissertation, Katholieke Universiteit Leuven, Leuven, Belgium.

Gielen, S., Dochy, F., Onghena, P., Struyven, K., Smeets, S., & Decuyper, S. (2007). Goals of peer assessment and their associated quality concepts. In S. Gielen, *Peer assessment as a tool for learning* (pp. 41-66). Unpublished doctoral dissertation, Katholieke Universiteit Leuven, Leuven, Belgium.

Gielen, S., Tops, L., Dochy, F., Onghena, P., & Smeets, S. (2007). Peer feedback as a substitute for teacher feedback. In S. Gielen, *Peer assessment as a tool for learning* (pp. 95-124). Unpublished doctoral dissertation, Katholieke Universiteit Leuven, Leuven, Belgium.

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

Kim, M. (2005). *The effects of the assessor and assessee's roles on preservice teachers' metacognitive awareness, performance, and attitude in a technology-related design task.* Unpublished doctoral dissertation. Florida State University.

Kluger, A. N. & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin, 119,* 254-284.

Magin, D. & Helmore, P. (2001). Peer and teacher assessments of oral presentation skills: how reliable are they? *Studies in Higher Education, 26,* 287-298.

Miller, P. (2003). The effect of scoring criteria specificity on peer and self-assessment. *Assessment and Evaluation in Higher Education, 28,* 383-394.

Mory, E. H. (2003). Feedback research revisited. In D. H. Jonassen (Ed.), *Handbook of Research for Educational Communications and Technology* (pp. 745-783). New York: Macmillan.

Narciss, S. (1999). Motivational effects of the informativeness of feedback. In *Annual Meeting of the American Educational Research Association* Montreal.

Nilson, L. B. (2003). Improving student peer feedback. *College Teaching, 51,* 34-38.

Prins, F., Sluijsmans, D., & Kirschner, P. A. (2006). Feedback for general practitioners in training: Quality, styles, and preferences. *Advances in Health Sciences Education, 11,* 289-303.

Pryor, J. & Lubisi, C. (2002). Reconceptualising educational assessment in South Africa - testing times for teachers. *International Journal of Educational Development, 22,* 673-686.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18,* 119-144.

SAS Institute Inc. (2004). *SAS/STAT 9.1 User's Guide*. Cary: SAS Institute Inc.

Searby, M. & Ewers, T. (1997). An evaluation of the use of peer assessment in higher education: A case study in the School of Music, Kingston University. *Assessment & Evaluation in Higher Education, 22,* 371-383.

Sitthiworachart, J. & Joy, M. (2003). Deepening computer programming skills by using web-based peer assessment. In *Proceedings of the 4th Annual Conference of the LTSN Centre for Information and Computer Sciences.* NUI Galway (Ireland): LTSN-ICS.

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

Sluijsmans, D. (2002). *Student involvement in assessment. The training of peer assessment skills.* Unpublished doctoral dissertation. Open Universiteit Nederland, Heerlen.

Sluijsmans, D. & Prins, F. (2006). A conceptual framework for integrating peer assessment in teacher education. *Studies in Educational Evaluation, 32,* 6-22.

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

Taras, M. (2002). Using assessment for learning and learning from assessment. *Assessment and Evaluation in Higher Education, 27,* 501-510.

Topping, K. J. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy, & E. Cascallar (Eds.), *Optimizing new modes of assessment: In search of qualities and standards* (pp. 55-87). Dordrecht: Kluwer Academic.

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

Topping, K. J. (2005). Trends in peer learning. *Educational Psychology, 25,* 631-645.

van den Berg, I. (2003). *Peer assessment in universitair onderwijs.* Unpublished doctoral dissertation. Universiteit Utrecht, Utrecht.

Wiliam, D. (2006). Does assessment hinder learning? In *Speech delivered at the ETS Europe breakfast salon (11th July 2006).*

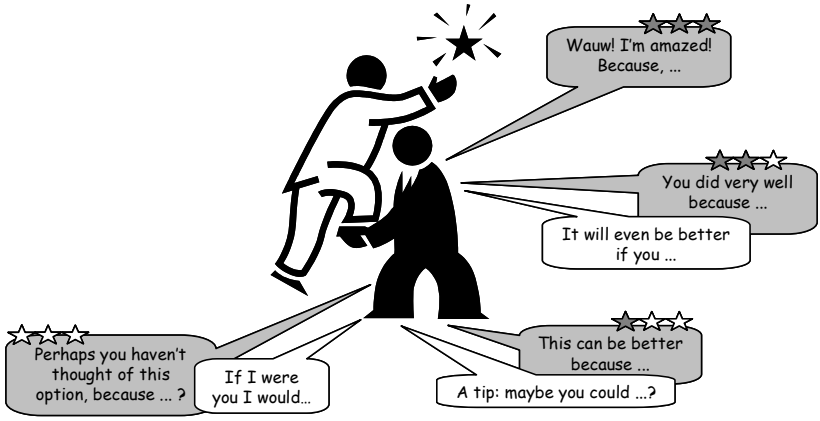
Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15,* 179-200.

Appendix

Feedback Form A

The 'critical friend'-assignment

Help each other
to make (even) beter assignments



Critical friend: Author of the essay:

For each criterion color the right amount of stars and explain what was right or could be better and why. Always suggest what your peer could do to improve his/her assignment. Work thoroughly and elaborately; be critical, honest and subtle!

CRITERION 1
What did he/she do well and why?

What didn't he/she do well and why?

If I were you I would ... , Maybe you could ... , It would even be better if you ...

CRITERION 2
What did he/she do well and why?

What didn't he/she do well and why?

If I were you I would ... , Maybe you could ... , It would even be better if you ...

CRITERION 3
What did he/she do well and why?

What didn't he/she do well and why?

If I were you I would ... , Maybe you could ... , It would even be better if you ...

CRITERION 4
What did he/she do well and why?

What didn't he/she do well and why?

If I were you I would ... , Maybe you could ... , It would even be better if you ...

CRITERION 5
What did he/she do well and why?

What didn't he/she do well and why?

If I were you I would ... , Maybe you could ... , It would even be better if you ...

CRITERION
What did he/she do well and why?

What didn't he/she do well and why?

If I were you I would ... , Maybe you could ... , It would even be better if you ...

Reply Form B

Letter to the reader of my essay

Dear reader,

I invite you to read my essay entitled

From the comments of my critical friend, I particularly remember that

By being a critical friend myself, and assessing the essay of somebody else, I learned that

After the 'critical friend-assignment' I revised my essay
1. with regard to (criterion)
   because
   and I tried to solve this by
2. with regard to (criterion)
   because
   and I tried to solve this by
3. with regard to (criterion)
   because
   and I tried to solve this by

My best piece is, in my opinion,
because

I paid this time special attention to
since

I hope you'll enjoy reading my essay

Kind regards,
(name)

Question Form C

Letter to my critical friend

Dear reader,

I invite you to have a look at my draft essay entitled

I tried to pay special attention to
by
Do you think it worked?

I still doubt on
1.
because
2.
because
3.
because

Most difficult, in my opinion, was
since
I'm curious what you think of it, and whether you can give me useful tips

I especially want to ask you to pay attention to the following criteria
1.
because
2.
because
3.
because

Finally, I want to remark that

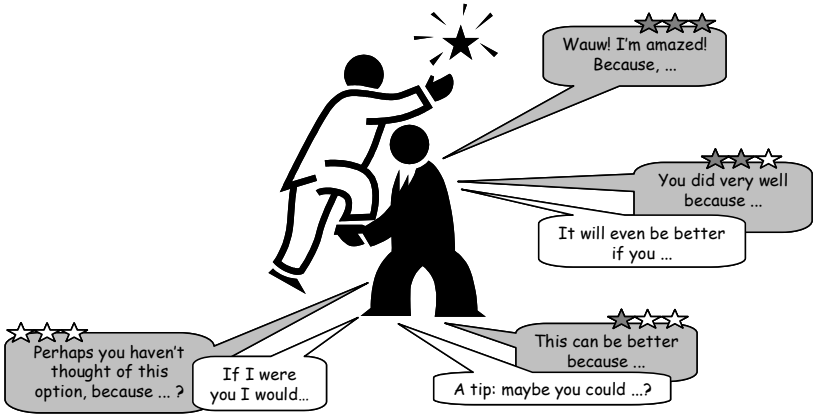
I hope you can give me useful tips, so I can improve my essay,

Kind regards,
(imaginary name)

Feedback Form D

The 'critical friend'-assignment

Help each other
to make (even) beter assignments



Critical friend: Author of the essay:

Work thoroughly and elaborately; be critical, honest and subtle!

After reading your essay, I first want to praise you because you

You said that you paid special attention to
Looking at your work, I think
since,

You said you had doubts on
1.
Looking at your work, I think
because,
2.
Looking at your work, I think
because,
3.
Looking at your work, I think
because,

You said that you found it difficult to
Looking at your work, I think
because,
In my opinion, it will even be better if you

Is it possible that you did not yet think of
since
Maybe you might solve this by

You asked feedback on the following criteria:
(1 star= can be better, 2 stars=well done, 3 stars=I'm amazed!)

CRITERION …

What did he/she do well and why?

What didn't he/she do well and why?

If I were you I would ... , Maybe you could ... , It would even be better if you ...

CRITERION …
What did he/she do well and why?

What didn't he/she do well and why?

If I were you I would ... , Maybe you could ... , It would even be better if you ...

CRITERION …
What did he/she do well and why?

What didn't he/she do well and why?

If I were you I would ... , Maybe you could ... , It would even be better if you ...

Finally, I want to add that

CHAPTER 7

A COMPLEMENTARY ROLE FOR PEER FEEDBACK AND STAFF FEEDBACK IN POWERFUL LEARNING ENVIRONMENTS

Abstract

This study aims to compare the strengths and weaknesses of two sources of feedback, namely peer feedback and staff feedback, from the student's perspective. The study is situated in a university course with a large number of students enrolled (N=192), where staff were only able to provide feedback on the draft versions of a series of cumulative assignments collectively to a whole class at once. This feedback was complemented by a formative peer assessment system. Starting from a hypothetical forced choice, students' preferences for one of the two sources of feedback are examined. A further in-depth study of the advantages and disadvantages addresses the perceived characteristics of the two sources of feedback, and their perceived contributions to a learning environment that attends to the learner's needs. These perspectives are complemented with reasons reported by students for their preferences for one of the two sources of feedback. Closed-ended questionnaire items are triangulated with qualitative data from open-ended questions.

Results show that approximately half of the students were willing to trade in the credibility of staff feedback for the specificity of peer feedback, if they had to choose. However, both sources of feedback showed their own strengths and weaknesses from the student's perspective. They were complementary and each even provided the conditions under which the complementary source became better. Peer feedback took care of the need for specific individual feedback, and thereby allowed staff feedback to concentrate more on an in-depth focus and a broader overview of all expectations in a collective session. Moreover, by providing staff feedback collectively, teacher resources were saved, allowing them to organise and guide a strong qualitative peer feedback system which offered opportunities for personal coaching, cooperative learning and metacognitive growth. Finally, by withholding immediate and complete help from staff until after the peer feedback sessions, students were stimulated to engage in lively discussions in the peer feedback sessions, eventually resulting in deeper learning.

A COMPLEMENTARY ROLE FOR PEER FEEDBACK AND STAFF FEEDBACK IN POWERFUL LEARNING ENVIRONMENTS

Introduction

The importance of feedback in education is undeniable (Black et al., 1998). However, providing feedback to students is often a major burden for teachers of large classes (Ballantyne et al., 2002). Because of this, some consider the use of supplementary or substitutional peer feedback, to share the work. In order to make an informed decision about how to combine peer and staff feedback (e.g., supplementary or substitutional use), it is necessary to have a view on the particularities, strengths and weaknesses of each source of feedback. Several perspectives can be taken in this enquiry. Firstly, the outcomes of both sources of feedback can be studied: which source leads to the most revisions; which source results in best student performance? Examples of this type of research are Yang, Badger and Yu (2006), Sadler and Good (2006) and Gielen, Dochy, Onghena, and Smeets (2007). Secondly, the content of both sources of feedback can be objectively studied and compared: which source is the most correct; which source contains the most encouraging comments; which source makes the most suggestions? Examples of this perspective are Falchikov (Falchikov, 1995), Magin (2001) and Topping, Smith, Swanson and Elliot (2000). Finally, one can take the student perspective and ask: which source do you prefer; which source do you experience as most helpful or encouraging? This study will take the perspective of the assessee, instead of investigating the composition of feedback in an 'objective' way. A good reason to investigate feedback from the perspective of the student is that the meaning of feedback messages is constructed by the receiver. Content analysis by researchers cannot take this 'transformation by the reader' into account. As Topping et al. (2000) mention: "The difficulty of conducting the qualitative analysis of similarity of semantic content raises questions about what students are likely to read into written feedback, even when of relatively high quality, well structured,

and substantial in quality. The assessed student might be less likely to extract the sense intended by the writer than researchers striving for objectivity" (p. 163).

The purpose of this study is to investigate whether peer feedback and staff feedback are comparable in the eye of the student, or whether each source has its own specific merits, and can thus be complementary in some way (see also Gielen, Dochy, & Onghena, 2007). It should be noted that the focus of this paper lies on the feedback that is received, abstracting the fact that peer assessment has the additional feature that learners also experience the assessment from the other side, by being an assessor (Topping, 2003). The latter has no counterpart in a teacher feedback system at all.

Framework

Previous Research

Students' preferences with regard to feedback sources have already been addressed by several empirical studies in the context of writing classes. Zhang (1995) suggests that a major distinction in the results lies between the studies that examine the impact of feedback on the writing of native speakers (referred to as L1) and those which consider it on the writing of non-native speakers of a language (referred to as L2). In her summary of previous research, Zhang describes that peer feedback is often found to be preferred by students above teacher feedback in L1 studies. Teachers are perceived as 'nit pickers', peers provide a more compelling impetus for revision, students see more social support in peer responses than in teacher feedback, teachers' comments are unhelpful and confusing, etcetera. All these studies converge on one theme, Zhang argues: teacher-controlled feedback is inherently lacking in affective appeal to L1 students when compared with peer feedback. This is in contrast with findings in L2 studies. Nine studies in L2 writing are described by Zhang. In all these studies students clearly preferred teacher feedback to peer feedback: it was more helpful and it was more attended to in revisions. However, when not asked to make a choice, learners showed reserved but generally positive attitudes toward peer feedback (Leki, 1990). It should be noted that these studies were mostly small case studies, usually with less than 20 participants. In her own study, Zhang (1995) asked 81 L2 learners about their preferences for peer feedback or staff feedback.

Again, 94% of the students chose staff feedback. On the other hand, when formulating their question from the supplementary perspective instead of the substitutional, Jacobs, Curtis, Braine and Huang (1998) found that 93% of their students would like to receive peer feedback as one kind of feedback.

The distinction between L1 and L2 students, however, does not seem to be as clear as described by Zhang. Some of our own studies in L1 writing classes have results that are more similar to the L2 results described by Zhang (Gielen, Dochy, Onghena, & Smeets, 2007). Less than 50% of the students found the feedback from their peers helpful, and less than a quarter found the experience of giving feedback useful to their own learning process. Only 37% wished to use peer feedback, as a substitute for staff feedback, for future assignments.

Writing classes are one of the first subjects in which peer review methods or peer feedback were introduced, since they are quite a natural way to bring the audience to the writer. More recently, however, peer feedback has also been introduced in other subjects in education as well, where writing is just a means to show what you have learned on a specific topic. Studies on the preference for peer feedback or staff feedback in more knowledge oriented subjects are sparse.

This study will address the characteristics of peer and staff feedback in a first year introductory course in educational sciences. Based on the available literature, it is not clear what preference of students should be expected. It is clear that more insight is needed into the relationship between peer and staff feedback when it is applied to more knowledge oriented curricula instead of those focusing on writing skills. Two other differences from the available body of previous research are that this study will use a larger sample than most of the studies reported in Zhang (1995), and it will go beyond the mere preference or the anecdotal information about the strengths and weaknesses of the two sources of feedback, by addressing these differences directly in our data collection. This more in-depth study of the student perspective on peer feedback compared to staff feedback, going beyond the expressed preference, will take place in two ways. Firstly, a deductive approach will be adopted to see whether students perceive peer and staff feedback to be different concerning some aspects that are identified in the literature. Secondly, an inductive approach will lead to an investigation of students' own experiences of the differences between the sources. The

deductive approach will use two starting points, which will now be described in turn.

First Starting Point for the Deductive Approach: Content Analysis Research

The content analysis research on feedback will be an initial source of information for the deductive approach. Based on this research, the type of information which is important to be present in a feedback message (such as suggestions or a balance of positive and negative statements) is defined. Several researchers have studied the content of peer and staff feedback, all at the level of higher education. Yang, Badger and Yu (2006), for instance, found that peer feedback contained more content-related comments, leading to meaning-changing revisions, whereas staff feedback held more surface remarks. Falchikov (1996), on the other hand, found in her 'peer feedback marking II' study that peers' comments were of a more practical nature, while teachers' comments were deeper. Topping et al. (2000) also identified a different focus in peer feedback compared with staff feedback. Rather than disagreement about aspects on which each had focused, they found that staff comments showed a tendency to be global, while peer comments were more particular and detailed, mentioning more specific examples.

The focus and the level of feedback are not the only important aspects of the content: the 'feedback sign' and the motivational framing of the statements are too. Students in Weaver's (2006) qualitative study mention a lack of positive comments in tutor feedback having a discouraging effect. According to Falchikov (1996), the presence of positive feedback is very desirable, particularly for students who lack confidence; but a lack of negative feedback is a problem too, since it is necessary to stimulate reflection and change. In Falchikov's (1996) studies I and II, more strengths than weaknesses were identified by both peers and staff, especially for stronger students. In her study III however, peers provided both more, and more positive, feedback than staff, together with more prompts and suggestions. Moreover, peers also provided more balanced feedback, containing both positive and negative statements for the same criterion. It is important that these negative statements are framed in a motivational way, for instance by adding prompts or suggestions, so that the learner knows how to proceed (Wiliam, 2006). Sadler (1998) indeed stresses the importance not

only of the technical quality of feedback, but also its catalytic and coaching value and its ability to inspire confidence and hope.

Finally, content analysts point to the issue of discourse in feedback messages. One of the 'technical qualities' of feedback is its comprehensibility (Sadler, 1998). Research that compares peer and staff feedback shows that they differ on this issue. In the study of Yang et al. (2006), peer feedback appeared to be more comprehensible for students leading to fewer misunderstandings and, consequently, more successful revisions. Higgins (2000) also describes this gap between teacher feedback and student understanding. Feedback messages such as "be more critical" or "your arguments need to be more academic" do not have the same meaning for teachers and students, because they are associated with a specific discourse that is not directly accessible to students. More explanation and different language are needed. In their feedback, peers use a language that is closer to the learners' language; it is the discourse of the learner in a domain, not the discourse of the expert.

Taking the student's perspective in our study, we do not investigate whether these components are actually in the messages provided by peers and staff, but whether they are received by the assessee. For instance, we do not count objectively which source of feedback provides most suggestions, but ask students which source is most inspiring in terms of suggestions for improvement.

Second Starting Point for the Deductive Approach: Research on Powerful Learning Environments

The second starting point is research on powerful learning environments that support the learner's needs. This type of research does not focus on feedback in particular, but describes the functions that a learning environment as a whole should fulfil in order to stimulate and maintain meaningful learning. Feedback, being an important component in such a learning environment, can take up several of these roles (Gielen, Dochy, Onghena, Struyven, et al., 2007).

In their description of the 'equilibrium model' for the construction of powerful learning environments, Schelfhout, Dochy, Janssens, Struyven, and Gielen (2006) try to summarize the main dimensions that a learning environment should take care of. Depending on the learning goals, students' and teachers' characteristics, and available resources, one has to search for an

effective and manageable balance between possibly conflicting teaching activities. They therefore call it an 'equilibrium model'.

The first main group of educational activities is those which foster the motivation of students to exert effort to engage in learning activities and to sustain these efforts. The second dimension incorporates all teacher activities which aim at engaging students in learning tasks in which they have to link their prior knowledge with new information to be able to construct meaning. Furthermore, students should be coached before, during and after these tasks. Here lies an important role for feedback and opportunities for revision: to activate the students to take the next step in their learning processes. Part of this feedback will be on the learning content, but another important part will have to support the metacognitive learning processes which the learners have to go through. The final dimension described by Schelfhout et al. (2006) incorporates all teacher activities which aim at structuring and steering the learning processes.

Although the most obvious contribution of peer and staff feedback is to the 'coaching and feedback' dimension of the equilibrium model, previous research has already pointed at their impact on the other dimensions too. The cooperative nature of learning from peers, which can be found in peer feedback, is shown to be an important motivator in learning environments (Slavin, 1989). The introduction of a social control element into the learning environment by means of peer feedback might also raise student motivation (Gibbs et al., 2004; Pope, 2001). Moreover, by letting students give feedback themselves, cognitive processes at the level of evaluation are activated (Pryor et al., 2002). These processes may lead towards a critical reflection on the subject, its criteria and the appropriateness of several solutions or approaches; those of the peer as well as the student's own. By receiving feedback from a non-expert, students also feel the necessity to reflect critically on the validity of these arguments, giving rise to further in-depth study of doubts, consultation of additional resources, and self-correction (Yang et al., 2006). Finally, raising the awareness of learning goals, criteria, and standards of a given course can be considered as a structuring and steering element in the learning environment.

On the other hand, peer feedback also brings some risks into the learning environment when not managed properly. For instance, group dynamics problems in a feedback group (e.g., resentment about unequal investment of effort), misconceptions or questions concerning the learning

content that remain undetected or unanswered by the peers, can lower motivation and distract attention from the learning goals. The provided feedback itself may also be discouraging if not formulated in a constructive way. Finally, students with a strong dependence upon approval by the teacher might become frustrated when staff feedback is completely substituted by peer feedback.

The impact of peer and staff feedback on these dimensions of the learning environment may differ from student to student, since students have different perceptions of the 'objective' feedback, and have different needs that have to be addressed. Again, therefore, we will not decide as researchers whether peer and staff feedback did a good job on these dimensions, but instead ask the learners whether they felt both sources of feedback contributed to supporting their needs.

An Inductive Approach to Complement the Deductive Approach

The third way of investigating the student perspective of peer feedback in relation to staff feedback does not make use of theory-based closed-ended items. As will be described further in the method section, we plan to conduct a qualitative analysis on students' own reported reasons for choosing one of the two sources of feedback. Previous studies have already addressed the perceived advantages and disadvantages of peer assessment in rather general terms (e.g., Hanrahan & Isaacs, 2001), or the perceived validity and reliability of peer assessment (e.g., Robinson, 2002). A comparison between staff and peer feedback, however, might elicit a more finely graded view of the strengths and weaknesses of peer feedback, as opposed to staff feedback. It is also an opportunity to validate the information retrieved from our theory-driven instruments based on the first two starting points. Do students spontaneously mention our predefined characteristics or dimensions, or do they have completely different reasons for choosing one or the other source of feedback?

Individual Peer Feedback versus Collective Staff Feedback: Making an Ecologically Valid Comparison

A characteristic of the present study is that it compares individual peer feedback with collective staff feedback. Most studies that performed a content analysis dealt with individual peer and individual staff feedback, as did several studies which discuss their impact on the learning environment.

However, this is not a realistic comparison in large classrooms. The provision of individual staff feedback, especially on the drafts of two-step assignments where staff will have to read a revised version again for summative marking, is not common in higher education (Ballantyne et al., 2002). Even the provision of this early type of feedback, which is formative in a true sense, in a collective way is more of an exception than common practice. If feedback is provided by staff, it is mostly done after summative assessment. 'Intermediate' feedback can realise the most learning benefits, however, and so it is taken as the focus of this study (Black et al., 1998; Crooks, 1988). The potential of peer feedback is compared to its most realistic counterpart at this stage, namely collective staff feedback. It is clear that the effects of the source of feedback are compounded with the effects of the specificity of the feedback. The interaction of both effects is, however, precisely what happens in large classrooms where it is not feasible for staff to provide frequent individual feedback to students on intermediate assignments.

Research Questions

Based on the aforementioned approaches and starting points, and aiming to unravel the student perspective on individual peer feedback compared with collective staff feedback, we formulate the following research questions:

1. Do students prefer peer feedback or staff feedback?
2. What are peer feedback's and staff feedback's strengths and weaknesses concerning the characteristics of effective feedback?
3. What are peer feedback's and staff feedback's strengths and weaknesses with regard to the impact that both sources of feedback have on learning, by means of contributing to a powerful learning environment?
4. What are students' self-reported reasons for their preferences for peer or staff feedback?

Method

Participants

A total of 192 first-year university students (93% female, aged 18-20) enrolled in the 'Educational Sciences' program participated in this study. All students of the 1st year of the Bachelor program were expected to take part, and although there was no external reward for participation, the response rate to the (electronic) questionnaires was approximately 95%. The total group was divided into three subgroups, each studying a different case.

Peer Assessment Design and Procedure

Formative peer assessment is applied to three successive assignments (the first being a training case). Typical of the assignments is that they are rather open, so there is no black and white answer to whether a solution is correct and no two assignments are the same. Students can still revise their assignment after feedback, before the summative assessment. Students are grouped in threes, each providing written and oral feedback to the two other group members. The thoroughness of the feedback is part of the summative mark for the feedback giver, to stimulate effort and to justify the investment of time. A more in-depth description of the particularities of the design and procedure of the peer assessment in this study, structured by means of the inventory of peer assessment diversity (see Gielen, Dochy, & Onghena, 2007), is added in Appendix A.

Staff Feedback Procedure

Each staff member and a teaching assistant read a sample of the draft assignments that were posted on Blackboard to get an idea of the common misconceptions and difficulties. The three staff members provided feedback, each for one case. Most of the feedback was identical between the three subgroups, however, since it was jointly prepared at a team meeting. Only the examples were sometimes case specific. Staff feedback was organized collectively for students and always came on Mondays, after the peer feedback session of the preceding Friday, to stimulate students to think for themselves first before relying on help from staff. Teaching assistants were present at the peer feedback sessions, but did not answer questions about the assignments, only ones about the peer feedback procedure. Staff

feedback was a one-hour lecture, in which staff gave examples of good and bad work, clarified and illustrated the criteria and rationale of the assignment and answered questions.

Research Design

Our first questionnaire was administered in the revision phase of an assignment and was repeated twice, once after the second assignment (the first 'independent' experience with peer feedback) and once after the third assignment (see Figure 1). This questionnaire's data are used to answer the first, second and fourth research question. The second questionnaire deals with research question 3, and was administered at the end of the course. Both instruments will be described in detail in the next section.

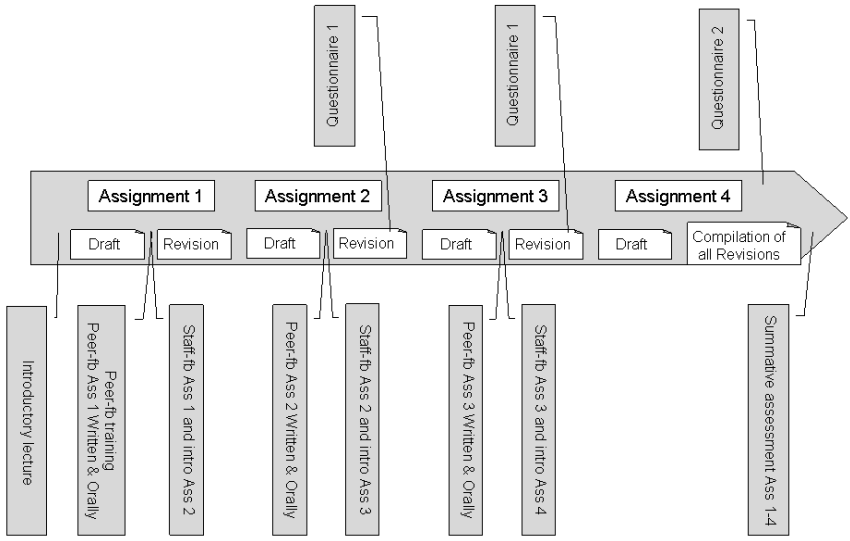


Figure 1. Diagram of the peer feedback procedure and data collection

Instruments

Questionnaire 1 – Forced choice. By means of the first questionnaire, students were placed in a (hypothetical) forced choice situation after the revision of assignment 2 and 3, and were asked to choose between peer or staff feedback for the next assignment (research question 1). In an

open-ended subquestion, they were asked to substantiate their choice (research question 4).

"If you could only get one of both sources of feedback for the next assignment, whose feedback would you choose? Why?"

Students could only choose one or the other source of feedback in the multiple choice part of the question. In the open question, students were free to add their wishes to combine both sources. These answers are categorised in the qualitative section of the analyses.

Questionnaire 1 – Characteristics of feedback. Content analyses studied differences in level, focus, encouragement, criticism, and comprehensibility of peer and staff feedback. Reformulation of these differences, in terms of the perceptions of students, gave us the following list of six characteristics, on which students were asked to compare the received peer and staff feedback: specificity, being comprehensible, being inspiring, informativeness, being encouraging and being thought-provoking.

The items to measure students' perceptions of these characteristics can be found in Appendix B. Students had four response options with which to answer these comparative items: 'peer feedback is better', 'staff feedback is better', 'both are equal', and 'not applicable'. The questionnaire also contained two control items concerning the presence of peer feedback and the attendance at staff feedback sessions, and only the answers of students who received both sources of feedback were retained.

Questionnaire 2 – Functions of a powerful learning environment. Based on the dimensions of the equilibrium model of Schelfhout et al. (2006), six functions of a powerful learning environment were identified: motivating, activating, coaching, steering, structuring, and support of metacognitive knowledge and skills. We first asked students whether they felt sufficiently supported in their needs associated with the dimensions (e.g., whether they felt activated to study the subject thoroughly). This is a base measure to give weight to the second measure that asked students to indicate to what extent the two feedback systems contributed to this function of the learning environment. The items of questionnaire 2 can be found in Appendix C. All items were presented on a 5-point Likert scale.

Analyses

All statistical analyses were executed with the SAS System (SAS Institute Inc., 2004). Details of the analyses are discussed for each research question.

Research question 1. Chi-square goodness-of-fit tests for equal proportions of students choosing one of the two sources of feedback were performed, to determine whether or not the options were equally chosen. These tests were performed separately for each measurement occasion. Then a McNemar test with a continuity correction was performed to compare the choices made after the second and the third assignment, and to determine whether the measurement occasion was associated with the choice. An odds ratio provided a measure of the magnitude of the effect.

Research question 2. Chi-square goodness-of-fit tests were performed for equal proportions of students choosing each of the response options. These tests were performed separately for each measurement occasion. For those proportions that were above the level of true majority (more than 50% of the students), a one-sided exact binomial test, with the null hypothesis that the proportion is equal to 0.5, was executed.

Research question 3. Summary statistics and 95% confidence intervals were calculated for the individual Likert scale items. To investigate the relative contributions of peer feedback and staff feedback to the support of each of the learner's needs, paired *t*-tests were performed.

Research question 4. A qualitative analysis of the answers to the open question "Why did you choose this source of feedback?" was carried out. Students' spontaneous arguments were coded in the Atlas.ti software (Atlas.ti Scientific Software Development, 2006). Each answer was considered as a unit of analysis (a 'quotation' in Atlas.ti). In total 146 quotations of the first measurement occasion and 159 quotations of the second were coded.

The analysis started from 12 theory-driven codes based on the six characteristics of effective feedback (cf questionnaire 1) and the six functions of a powerful learning environment (cf questionnaire 2). These 12 codes were structured into two higher order code families ('characteristics of feedback'

and ‘functions of the learning environment’). Additional codes were constructed inductively and iteratively (Neuman, 1994; Patton, 1987). Finally, the family structure was revised, by extending some families with additional codes from the inductive process, and by creating new families within the remaining additional codes.

The coding system was not exclusive. Each quotation could receive several codes, both from the same code family or from several families. The absence of a code for a quotation only indicates that the student did not mention this reason spontaneously. This is especially true for the codes that are not theory-driven. For the codes that were theory-driven, students had already encountered a closed question earlier on in the questionnaire that asked them to compare peer feedback and staff feedback with regard to each characteristic or function (see also research questions 2 and 3). As a consequence, we gave students a certain ‘vocabulary’ to compare the two sources of feedback. It can be expected that these descriptions would often be used in the open questions too. But when students added other reasons, and their reasons entered the coding scheme inductively, one could not expect that all students had considered all these reasons. Therefore, we should be cautious in our interpretation of low percentages. For instance: if 20% of the students who chose staff feedback mentioned that one of their reasons was that they considered the assessor to be an expert, one cannot conclude that the other 80% of the students thought the assessor was not an expert. We can only say that the level of expertise of the assessor was not mentioned spontaneously as a reason for their choice. High percentages indicate important reasons; reasons with low frequencies help us to broaden our view of other possible reasons for a choice.

Results

Research Question 1: Students’ Preferences

We asked students which source of feedback they would choose if they were given the choice for the next assignment. After the second assignment (and the first real peer feedback experience), 61% of the students chose peer feedback above staff feedback (see Table 1). This is a significant majority, χ^2 (1, N=177)= 8.593, p=0.0034.

After the third assignment 48% of the students still preferred peer feedback and the other half; 52% chose staff feedback (see Table 2). These percentages do not differ significantly from a 50-50 distribution, χ^2 (1, N=169)= 0.290, p=0.590.

Table 1

Frequency table (including row percentages) of students' choices between the sources of feedback at the two measurement occasions

	Choice next assignment		
Frequency Row pct	Peer feedback	Staff feedback	Total
Assignment 2	108	69	177
	61.02 %	38.98 %	100 %
Assignment 3	81	88	169
	47.93 %	52.07 %	100 %

If there was no association between measurement occasion and choice, one would expect that the number of students that chose peer feedback in assignment 2 and staff feedback in assignment 3 would be equal to the number of students that chose staff feedback in assignment 2 and peer feedback in assignment 3. In this study, there were 39 discordant pairs (students who changed their choices). There were 10 (25.6 %) who first chose staff feedback and then chose peer feedback, and 29 (74.6 %) who first chose peer feedback and then switched to staff feedback (see Table 2).

Table 2

Frequency of students' choice evolutions between assignment 2 & 3

Choice evolution	Frequency
Peer-peer	69
Staff-staff	52
Peer-staff	29
Staff-peer	10

McNemar's test shows that there is a significant difference between the two assignments in the proportion of students that chose peer feedback, χ^2 (1, N=160)= 8.308, p=0.0039. The odds ratio is 2.9, with a confidence interval of (1.374 ; 6.670), indicating that the odds of choosing staff feedback rather than peer feedback in assignment 3 is almost three times higher than

after assignment 2. Although half of the students still chose peer feedback at assignment 3, the number had decreased significantly compared to at assignment 2.

Research Question 2: Characteristics of Feedback

Students were asked to compare peer feedback to staff feedback with regard to six characteristics: specificity, being comprehensible, informativeness, being inspiring, being encouraging and being thought-provoking. Results are presented in Table 3 and Figures 2 and 3. Each axis in the graphs represents one characteristic of effective feedback and each hexagon represents one response option. The wide dashed line stands for 'both are equal', the solid line for 'peer feedback is better', the thin dashed line for 'staff-feedback is better'. The height of the crossing of the hexagon at the axes indicates approximately1 the percentage of students that chose one of the alternatives. At each axis, the sum of the three crossings is 100%.

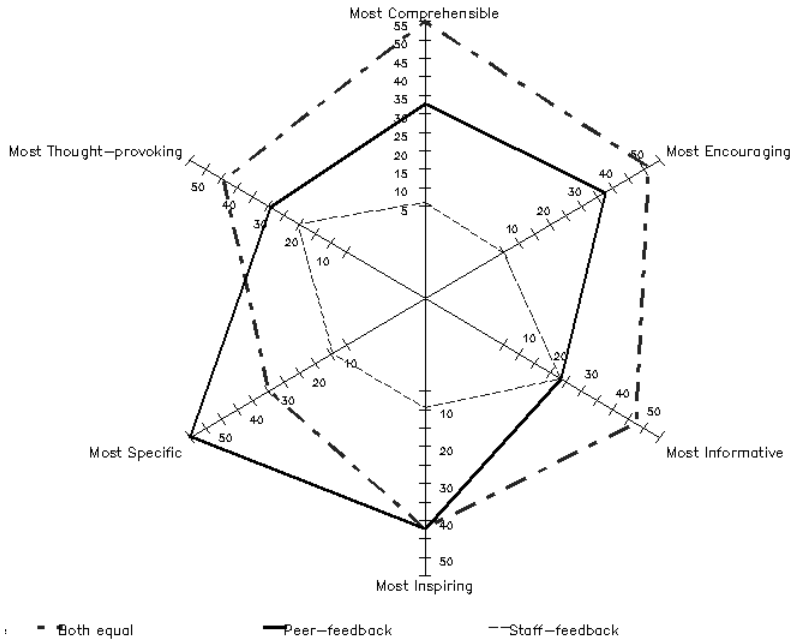


Figure 2. Radar Plot indicating the percentage of students choosing one of the three response options for the characteristics items at assignment 2.

[1] The values indicated on the axes are only an approximation, due to technical limitations in SAS Graph to determine freely the distance between the ticks. Correct percentages can be found in Table 3.

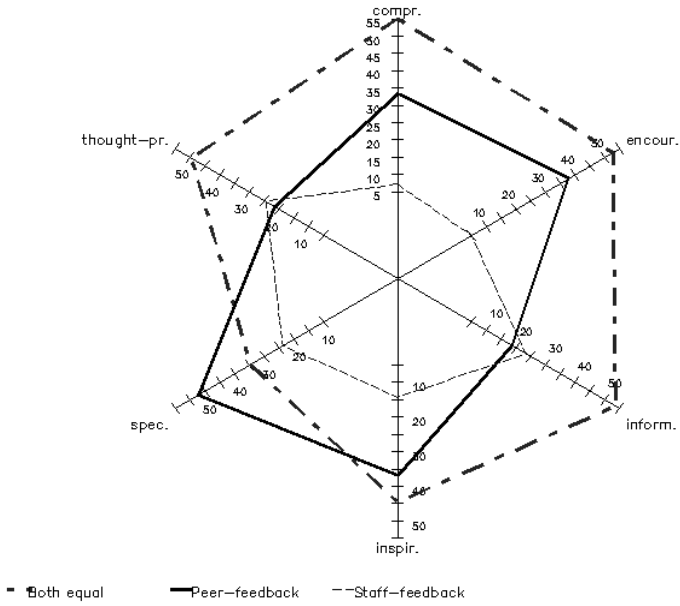


Figure 3. Radar Plot indicating the percentage of students choosing one of the three response options for the characteristics items at assignment 3.

Table 3

Summary statistics for students' comparisons of peer and staff feedback concerning six characteristics of effective feedback

		Assignment 2		Assignment 3	
Aspect	Source	*N*	*P*	*N*	*P*
Most Specific	Peer	173	0.57	161	0.49
	Staff		0.11		0.20
	Equal		0.32		0.31
Most Comprehensible	Peer	172	0.35	163	0.35
	Staff		0.08		0.08
	Equal		0.58		0.57
Most Inspiring	Peer	173	0.45	164	0.38
	Staff		0.11		0.15
	Equal		0.45		0.46
Most Informative	Peer	171	0.25	167	0.20
	Staff		0.25		0.25
	Equal		0.50		0.56

Most Encouraging	Peer	171	0.40	157	0.39
	Staff		0.06		0.06
	Equal		0.54		0.55
Most Thought-provoking	Peer	170	0.31	163	0.23
	Staff		0.22		0.26
	Equal		0.46		0.52

Table 4

Summary statistics for students' comparison of peer and staff feedback concerning six characteristics of effective feedback

			Chi-Square test for equal proportions			Exact Binomial test for proportion* = 0.5
Aspect	Assign-ment	*N*	*Chi-Square*	*df*	Pr>*ChiSq*	One-sided Pr>*P*
Most	2	172	64.686	2	<0.001	0.028
Comprehensible	3	163	59.092	2	<0.001	0.042
Most	2	171	60.737	2	<0.001	0.179
Encouraging	3	157	59.325	2	<0.001	0.132
Most	2	171	20.632	2	<0.001	
Informative	3	167	38.132	2	<0.001	0.082
Most	2	173	38.890	2	<0.001	
Inspiring	3	164	25.695	2	<0.001	
Most	2	173	55.676	2	<0.001	0.034
Specific	3	161	20.957	2	<0.001	
Most	2	170	15.188	2	0.001	
Thought-provoking	3	163	24.528	2	<0.001	0.377

* For 'Most Specific' the proportion of 'Peer feedback' is tested. For all other aspects, the proportion of 'Both were equal' is tested.

Although small variations exist between the assignments, the overall patterns are similar. We therefore discuss both assignments together. For all characteristics, the chi-square tests of equal proportions reject the null hypothesis significantly (see Table 4). For all but two of the characteristics, the group of students that perceived both sources of feedback as equal is the

largest. With regard to being specific, the largest group is those that perceived peer feedback as more specific. With regard to being inspiring the group that considered peer feedback as better and the group that considered both sources as equal are more or less of the same size. With exception of two characteristics, the group that considered staff feedback as better was always the smallest group. This was not the case for being informative and being thought-provoking.

Seven values are above the 'true majority' level of more than 50% of the students. Only three of these are significantly different from $P = 0.5$ (see Table 4), indicating that a true majority of students perceived peer and staff feedback as equally comprehensible after the two assignments, and that a true majority of the students considered peer feedback as more specific after the first assignment.

Research Question 3: Functions of a Powerful Learning Environment

Table 5 contains the results of the first subquestion (Is the learning environment powerful?). Figure 4 and Table 6 summarize the results of the second subquestion, which relates the feedback systems to the powerfulness of the learning environment (What is the contribution of the two feedback systems to the learning environment?). The contribution of peer feedback is indicated by a wide solid line and the contribution of staff feedback by a thin dashed line. The dotted lines are reference lines at value 3 (midpoint) and 3.5.

Table 5

Summary statistics for students' feelings of being supported in their needs associated with the six functions of a powerful learning environment

Variable	Min	Lower 95% CL	Mean	Upper 95% CL	Max	*N*	Std Dev
Activated	2	3.97	4.06	4.15	5	144	0.54
Coached	1	2.58	2.74	2.89	4	145	0.94
Metacogn. supported	1	3.48	3.60	3.72	5	143	0.71
Motivated	1	3.82	3.93	4.04	5	145	0.66
Steered	1	2.52	2.66	2.81	4	145	0.89
Structured	3	4.08	4.16	4.24	5	144	0.50

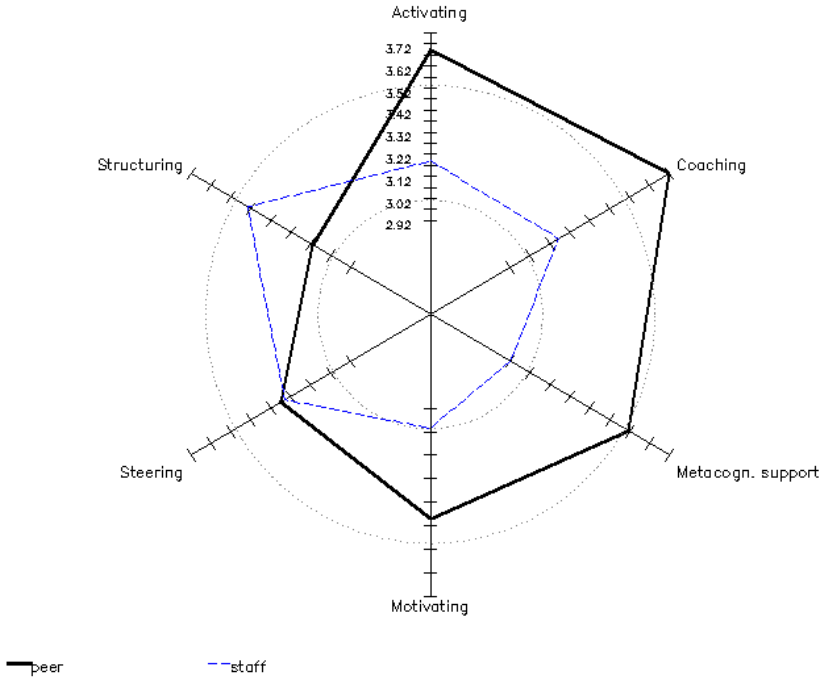


Figure 4. Radar Plot indicating the mean values on a 5-point Likert scale concerning the contribution of both feedback systems to the functions of a powerful learning environment.

Students felt activated and motivated on this course, and the structure of the subject was clear to them (results are significantly different from the scale centre, see Table 5). In their opinion, the first two needs were mainly supported by their peers (the wide solid hexagon in Figure 4), the last one by the staff. Differences between sources of feedback are significant, as indicated in Table 6.

To a smaller extent, but still significantly positively, they felt they had gained insight into their own learning during this course (metacognition), and peer feedback had contributed to the support of their metacognitive needs while staff feedback had not, as indicated in the plot (Figure 4) and Table 6.

On the other hand, students felt they were not sufficiently coached (the CI of the score mean lies completely below the midpoint of the scale, see Table 5). Although they thought their peers really did a great job in the coaching (this item even has the highest mean score of all, see Table 6), it

appears not to be sufficient. The overall perception of coaching was probably negative (below the midpoint) because they felt the staff's input into the coaching was too little. The staff's contribution to coaching was rated significantly lower than that of the peers (Table 6).

Finally, the need for steering by clear expectations in this course was also not fulfilled. The mean score and confidence interval lies entirely below the centre of the scale. The answer to the question of who helped them in getting a clear view of the expectations was mixed. Both mean values lie above the midpoint of the scale, but are less pronounced (Table 6). They do not differ significantly. Investigating the distribution of the raw data reveals that, for half of the students (51%), staff succeeded in clarifying the expectations during their feedback sessions, but a quarter disagreed and another quarter was undecided.

Table 6

Paired t-tests for the differences between the contributions of peer and staff feedback to the functions of a powerful learning environment

	Mean		Paired Differences				*t*	*df*	*p*
	Peer fb	Staff fb	Mean	St.Dev.	Lower 95% CL	Upper 95% CL			
Activating	3.69	3.19	0.507	0.902	0.356	0.658	6.655	139	<.001
Coaching	3.77	3.18	0.592	1.012	0.424	0.759	6.968	141	<.001
Metacogn. support	3.55	2.92	0.630	0.920	0.473	0.786	7.950	134	<.001
Motivating	3.42	3.01	0.414	0.982	0.250	0.578	4.993	139	<.001
Steering	3.29	3.27	0.022	1.066	-0.157	0.200	0.239	138	.812
Structuring	3.12	3.47	-0.343	0.844	-0.486	-0.200	-4.758	136	<.001

Research Question 4: Reasons for Students' Preferences

The initial code system, based on the six characteristics of effective feedback and the six functions of a powerful learning environment, appeared to be insufficient to catch all the reasons students had to justify their choices between staff and peer feedback. Additional codes were created inductively and iteratively. Some of these additional codes appeared to be an important supplement for the initial, theory-driven, code family of the 'characteristics of feedback'. For instance, it soon became clear that this family lacked a code

for ‘being safe’ or trustworthy - and its subcodes ‘being correct’ and ‘being complete’. Other codes that were added to this family are ‘deep level of feedback’ and its counterpart ‘poor quality of the non-chosen option’, ‘also positive feedback’, ‘more interesting focus of feedback’ and finally ‘redundancy of the non-chosen option’ which can be considered the opposite of usefulness. The second theory-driven family (functions of a powerful learning environment) remained as it was.

The other additional codes that were created could be grouped into two new higher order families and one ‘rest category’. The first new family is related to the characteristics of the feedback but does not focus on the feedback itself; instead, it relates to the features of the assessor or the feedback setting. It contains codes such as ‘assessor is expert’ or ‘personal interaction with assessor’ (see Table 9 for the other codes). The second new family contains several codes referring to a comparison of the two sources of feedback. For instance, a student explained that he chose one source of feedback because the non-chosen source lacked certain qualities, or a student gave additional discussion of the advantages of the non-chosen source of feedback or the disadvantages of the chosen source. Finally, some utterances about ‘ideal’ situations that go beyond the ‘forced choice’ are situated in this fourth code family.

Each code family will now be discussed and illustrated by means of quotations. At the end of each section, a table is provided with the frequencies of each code in the data, grouped within the choices that the students made.

Characteristics of feedback. Most students mentioned at least one characteristic of the feedback itself, to justify their choice, and a reference to the lack of that characteristic in the other type of feedback was also often added (“opposite to non-chosen option”). The two main reasons for choosing peer feedback or staff feedback are clearly specificity on one hand and safety (reliability, credibility) on the other hand. One of these two reasons is mentioned in approximately 60% of all answers to the open question (see Table 7).

Table 7

Percentage of students within a category of choice for each assignment who mention a certain reason of the code family 'characteristics of feedback'

	Assignment			
	2		3	
	Choice		Choice	
CHARACTERISTICS OF FEEDBACK	Peer *N*=83	Staff *N*=60	Peer *N*=76	Staff *N*=79
Specificity	**77.1 %**	.	**78.9 %**	2.5 %
Safe	.	**51.6 %**	1.3 %	**35.4 %**
Correct	.	**40.0 %**	1.3 %	**27.8 %**
Informative	**44.5 %**	**18.3 %**	**48.6 %**	**15.1 %**
Inspiring	**21.6 %**	1.6 %	**22.3 %**	.
Comprehensible	9.6 %	5.0 %	**13.1 %**	5.0 %
Non-chosen option redundant	8.4 %	.	6.5 %	1.2 %
Encouraging	6.0 %	.	5.2 %	.
Deep level	4.8 %	1.6 %	3.9 %	3.7 %
Non-chosen option possibly poor	.	1.6 %	.	3.7 %
Complete	1.2 %	8.3 %	.	5.0 %
Also positive fb	2.4 %	.	2.6 %	.
Thought-provoking	1.2 %	.	1.3 %	1.2 %
More interesting focus of fb	1.2 %	1.6 %	2.6 %	.

In the case of peer feedback, specificity was indeed a major reason to choose it. Three students out of four mentioned this characteristic.

> *I chose peer feedback: (…) Peers comment on the content of my paper and on my specific way of working. Teachers give good info and tips, but their feedback is rather general and applies to everyone. You have to figure out yourself what you can do with that info and what specifically applies to you.*

Beyond being specific, peer feedback was also liked because it was informative and inspiring.

> *I chose peer feedback: I found the feedback from my group members very inspiring. It contained a lot of useful tips that helped me proceed and that I could pay attention to in order to improve my assignment further. I experienced the feedback as very encouraging too.*

Other reasons that were mentioned were, from a more cognitive perspective, that it was more comprehensible, thought-provoking, provided more meaningful, deep-level information and had a more interesting focus.

From an affective perspective, reasons included that it was more encouraging and also addressed positive aspects of a performance.

> *I chose peer feedback: (…) Although you can doubt about how correct peer feedback actually is, it encourages you in a more specific way to start correcting your assignment. You can interpret it [peer feedback] as a number of questions that are asked about your work on which you decide yourself whether to take them into account when correcting it [your paper] or not. (…)*

Finally, from an eliminating perspective, some chose peer feedback because staff feedback was redundant.

> *I chose peer feedback, because it is a form of individual feedback and because it is about your own personal work. (…) To get a good idea of what teachers eventually expect of you, the evaluation checklist already helps a lot. [You don't need staff feedback for this.]*

The value of staff feedback on the other hand, apart from that it could be trusted to be correct and/or complete, was that it could also be very informative, students reported. Informativeness is thus a characteristic that was mentioned in both 'camps', although it is clear that it was mentioned more often as a characteristic of peer feedback.

> *I chose staff feedback: In the end, I find that this feedback is somewhat more reliable. I received good peer feedback on my previous assignment, and there were also some suggestions to revise something. But also a lot of things that came to the surface in the staff feedback were overlooked [by the peers].*

Furthermore, the comprehensibility, the deeper level, the thought-provoking nature and the more interesting focus were also mentioned by some students as their reasons for choosing staff feedback.

> *Staff feedback: I find this a very difficult question, because peer feedback surely is useful and valuable. But I would nevertheless choose staff feedback because it offers you comments that neither you yourself, nor your peers, had thought about already. These comments go deeper into the content of the assignment.*

Finally, the possibility that peer feedback could be poor, or the belief that peer feedback was redundant, made some students choose staff feedback.

> *I chose staff feedback: The teachers are the ones that actually have to correct your work so they can point out to you whether you are*

already thinking in the right direction. You can, if necessary, ask your fellow students for feedback but there is no guarantee that they've got it right.

Functions of the learning environment. Although the characteristics of feedback were mentioned quite clearly by students, the functions it fulfilled in the learning environment were more hidden in their answers. Students did not reason in terms of the abstract functions, but referred to them by means of concrete examples or implications. This code family therefore required more interpretation by the researcher. The presence of a 'function' code almost always co-occurs with 'more concrete' codes of the 'characteristics' or the 'features of assessor/setting' family.

Table 8

Percentage of students within a category of choice for each assignment who mention a certain reason of the code family 'functions of the learning environment'

	Assignment			
	2		3	
	Choice		Choice	
FUNCTIONS OF LEARNING ENVIRONM.	Peer *N*=83	Staff *N*=60	Peer *N*=76	Staff *N*=79
Coaching	**83.1 %**	**31.6 %**	**76.3 %**	**35.4 %**
Steering	.	**50.0 %**	2.6 %	**37.9 %**
Activating	4.8 %	1.6 %	3.9 %	1.2 %
Motivating	4.8 %	.	3.9 %	.
Metacognitive support	1.2 %	1.6 %	1.3 %	.
Structuring	.	.	.	.

We see in Table 8 that most students who chose peer feedback referred to its strengths in the coaching function.

I chose peer feedback: Peers honestly tell you when you've written something that is not completely clear. They focus on my paper and are able to compare it to previous assignments. They understand my topic well because they are more or less co-writers of my paper.

Furthermore, all other functions apart from the structuring function were mentioned, although not often.

I chose peer feedback: You spend more time on this feedback and discuss it more deeply.

Staff feedback was often chosen for its capability to steer students in the right direction. Teachers knew best what the requirements of the assignments were, and in the end, it was them who would assess their work, these students argued.

> *I chose staff feedback because it is the staff feedback that makes me understand better what is expected of me, and I try to incorporate this into my assignment.*

Finally, around one third of the students choosing staff feedback mentioned its coaching function. This percentage is much lower than in case of peer feedback, however.

> *Staff feedback: It is difficult to choose between both; they are both equally valuable... Well, [if I must choose] then staff feedback, as a way to know how to proceed and if we are on the right track: yes or no.*

Motivating and structuring were two functions of the learning environment that were never referred to as reasons for having chosen staff feedback.

Features of assessor/setting. Thirdly, there is a group of reasons that were related to the person who provided the feedback or the setting in which feedback was provided (see Table 9). These aspects were typically not applicable to both sources of feedback, as were the more general characteristics and functions.

Peer feedback provided the opportunity to interact personally with the assessor.

> *I chose peer feedback: (...) It is a much more personal way of working that enables you to ask questions more directly. To me that was the most useful aspect throughout the assignment, (maybe because the staff feedback didn't really apply to my work). (...)*

This feedback setting also had the potential to become a cooperative learning environment, and added a learning opportunity by having to read each other's work.

> *I chose peer feedback: I think you learn more from peer feedback and you get insight in what your colleagues (= fellow students) are doing. They can really make you think, not just about your own piece of work but also about theirs. (...)*

Furthermore, the affective safeness or enjoyment of peer feedback was larger: critique from peers was less threatening and more agreeable. Finally, some students appreciated the fact that the assessor was at the same level as themselves, or that they could also consult their peer, outside the formal feedback session.

Table 9

Percentage of students within a category of choice for each assignment who mention a certain reason of the code family 'features of the assessor or assessment setting'

	Assignment			
	2		3	
	Choice		Choice	
FEATURES OF ASSESSOR / SETTING	Peer *N*=83	Staff *N*=60	Peer *N*=76	Staff *N*=79
Assessor is expert, experienced	.	**20.0 %**	.	**27.8 %**
Assessor knows expectations	.	**21.6 %**	.	**16.4 %**
Cooperative learning environment	**13.2 %**	.	3.9 %	.
Personal interaction with assessor	4.8 %	1.6 %	10.5 %	.
Affective preference for assessor / setting	6.0 %	.	2.6 %	.
Assessor on same level	4.8 %	.	3.9 %	.
Learning from others' work	1.2 %	.	1.3 %	.
Assessor larger commitment in time	1.2 %	.	.	.
Less time or effort	.	.	.	5.0 %
Assessor less accessible, nonchoice still possible	.	3.3 %	.	2.5 %

The strength of staff feedback, on the other hand, was that it gave students the opportunity to listen to an expert in the subject who was experienced in guiding students learning processes and who had the best view of the expectations for the summative assessment.

> *I chose staff feedback: They pointed out better whether you were thinking in the right direction or whether you had to reconsider your selection of (pedagogical) support services (for the problem formulated in the case). The teachers also know a lot more about the different pedagogical support services. Your team members know a lot about their own selected services, but they only have a restricted (superficial) knowledge of your topic (through lack of time). They form their opinion mainly on the information I give them about a*

certain service, which makes them less able to judge its relevance (to the problem in the case).

An additional advantage for some students, related to the setting, was that it required less time and effort in times of high workload. A final reason is that, given the forced choice, if they chose peer feedback, the teacher assessor would not be available for feedback anymore. If they chose to receive formal feedback from the staff, however, they could still ask their peers to give them feedback informally.

Comparison between sources. A different group of codes is those referring to comparisons made between both sources of feedback.

Table 10

Percentage of students within a category of choice for each assignment who mention a certain reason of the code family 'comparison between sources of feedback'

	Assignment			
	2		3	
	Choice		Choice	
COMPARISONS BETWEEN SOURCES	Peer *N*=83	Staff *N*=60	Peer *N*=76	Staff *N*=79
Opposite to non-chosen option	**34.9 %**	**48.3 %**	**19.7 %**	**35.4 %**
Non-chosen option also has advantage	13.2 %	18.3 %	11.8 %	11.3 %
Chosen option also has disadvantage	7.2 %	.	7.8 %	2.5 %
Both complementary	10.8 %	10.0 %	7.8 %	5.0 %
Wishing for individual staff feedback	2.4 %	.	.	2.5 %

The relatively high percentages of 'opposite to non-chosen option' (see Table 10) tells us that students definitely saw some differences between the two sources of feedback, so they are not completely substitutional. About a quarter of the students (31.5% after assignment 2 and 23.1% after assignment 3) also mentioned advantages of the non-chosen option, and 7 to 10% mentioned disadvantages of their choice. This indicates that the forced choice was not always easy, and students had to balance the pros and cons of both sources.

I would prefer both, they are complementary. The peer feedback provides feedback on your own paper, staff feedback does not. Staff feedback provides feedback in general, so this can contain feedback which your peers did not mention because they did not yet understand this themselves.

Finally, a few students' ideal was not a combination of both peer and staff feedback, but an integration of the characteristics 'individual' and 'professional' in personal feedback to each student by staff.

I choose peer feedback: It is much more personal and you can ask for clarification immediately. If it were possible, however, that staff feedback was more personal, I would opt for that. Then you feel safe that the feedback is correct, which is not the case with the students…

Conclusions and Discussion

This study aimed to compare the strengths and weaknesses of two sources of feedback, namely peers and staff, from the student's perspective. The study was situated in a university course with a large number of students enrolled, where staff were only able to provide collective feedback on the draft versions of a series of cumulative assignments, and where this feedback was complemented with a formative peer assessment system. This study investigated students' preferences for one of the two sources, and searched for explanations in the perceived characteristics of the sources of feedback, their perceived contributions to a powerful learning environment, and other self-reported reasons for their choices.

Given a forced choice, approximately 60% of the students chose peer feedback above staff feedback after their first real feedback experience; after the second about half of the students preferred feedback from peers. Large numbers of students seemed to prefer individual feedback - even if it was delivered by peers - to collectively delivered staff feedback. Although this does not mean that 50% to 60% of the students thought staff feedback was redundant, it tells us that peer feedback had some value to about half of the students, as does staff feedback.

Comparing our findings to the existing literature on student preferences for peer versus staff feedback shows that our study is situated between the two opposing positions described by Zhang (1995). While the L1

literature reported students to have an absolute preference for peer feedback, and the L2 literature showed students who had an absolute preference for staff feedback (unless a supplementary perspective was taken), our study found that students were divided in two groups. These groups were those that preferred peer feedback, and those that preferred staff feedback. It should be noted that the specific procedures for peer and staff feedback will probably have had an impact on the results. Two features might have made staff feedback less popular and peer feedback more acceptable, in comparison with earlier studies: staff feedback was organised collectively and peer feedback was strictly guided by staff by means of a training session, a prestructured form, and a reward for good feedback. However, the choice for collective staff feedback was, as explained earlier, the only ecologically valid option. Although this is just an initial study on preferences outside the field of writing education, it indicates that there is a path between the two extremes reported in the literature before now.

The current study took a step further, by also addressing more systematically the reasons for students' preferences for peer and staff feedback. The preference of students only served as a starting point in this study, to elicit the perceived advantages and disadvantages of both sources. It became immediately clear in this study that both sources of feedback were equal opponents. Half of the students chose one and the other half chose the alternative. In the open-ended questions, students also often indicated that it was a difficult choice because they preferred both as they each had their strengths.

When comparing six characteristics of effective feedback, we saw that most students perceived both sources of feedback as equal with regard to being comprehensible, inspiring, encouraging, thought-provoking and with regard to their informativeness. This is not in line with the findings from Higgins (2000), that students do not understand the discourse of the expert. A possible explanation could be that when staff organise their feedback collectively, they can take the time to provide some generic examples and clearly explain their expectations. In studies where students complained about the comprehensibility of teacher feedback, teachers often did make an attempt to provide individual written feedback to all students (Weaver, 2006). This happened, however, at the expense of the volume of clarifications: they simply wrote some quick words in the margin, students reported.

The same reasoning goes for being encouraging. If staff have to comment on a large number of individual papers, the risk is that they will try to safe time by only writing down the most vital comments, and therefore only indicate the mistakes and weaknesses (Weaver, 2006). When providing more general feedback, they can provide some examples of good practices in which students can recognise their work. Getting an overview of all the requirements of a task, and not just those that you did not accomplish, is probably more encouraging for learners.

The other side of the coin, nevertheless, is a decrease in specificity. The qualitative analysis clearly showed that students appreciated the personal and concrete nature of peer feedback, and did not like that staff feedback was general. Specificity was even so important to students that 77% of the students who chose peer feedback in the forced choice mentioned its specificity as a main reason. This might also be the reason for the low score that staff feedback received on the coaching dimension. Being first year students at university, they did not feel ready for the task they were facing and they had hoped to get more support from the staff team. Although the teaching assistant was always available via e-mail and a consultation hour for individual questions and problems after the feedback sessions had taken place, students indicated clearly that they felt insufficiently coached during the course.

A few students explicitly mentioned that their ideal was an integration of the characteristics 'individual' and 'professional', by means of personal feedback to each student by staff. A majority of students would probably welcome this since students think, in terms of feedback, that 'more is always better'. The complementarity of peer feedback and staff feedback, however, might lie precisely in this combination of specific peer feedback and a gain in thorough, though collective, staff feedback, as explained above. Moreover, one might question whether 'more is always better' after all, even if it were feasible to provide individual staff feedback. Prins, Sluijsmans, and Kirschner (2006) argue in their discussion of feedback preferences that "What we like is not always what we need" (p. 300). One of the purposes of peer feedback was to make students less dependent on staff in their learning. This is a process that needs time, and it can be expected that it will not happen without any discomfort on the part of some students (Sluijsmans, 2002). University staff explicitly wanted to create a degree of tension to stimulate independent self-directed deep-level learning. Too much

scaffolding is not desirable; a powerful learning environment searches for a balance between scaffolding and fading (De Corte, 1996; Schelfhout, Dochy, Janssens, Struyven, & Gielen, 2006). Yang et al. (2006) provide an example of self correction, which was considerably reduced by exposure to staff feedback: "The over-dependence on teacher feedback is likely to lower students' initiative and lead to fewer self-initiated corrections" (p. 192).

The possible pitfall of peer feedback, that peers do not take the process seriously or only provide surface suggestions, as reported by Hanrahan and Isaacs (2001) and Nilson (2003) was countered by our study. This is shown by the fact that peer feedback was also considered informative, inspiring and thought-provoking, and that peer feedback was perceived to have made an important contribution to the coaching function of the learning environment. Four possible reasons are identified. Firstly, students were matched in teams that remained fixed during several feedback sessions, so the social pressure as well as the give and take mechanism probably motivated students to make an effort for each other. Secondly, students received feedback from two peers instead of just one, so forces were combined and weaknesses could be compensated for. Thirdly, the written feedback was extended with an oral discussion, and assessees were trained to get the best out of that discussion by taking an active role themselves and asking for the feedback they needed. Fourthly, assessors had an external stimulus to do their best, since a quarter of their mark depended on the quality of the feedback they provided. Another problem of peer feedback reported in other studies is that students had misgivings about peer feedback due to insincere, uncritical responses or - in contrast - an overly critical tone (Leki, 1990). The truly formative nature of the feedback in our study, and the fact that it was only a draft that peers had to comment on, probably prevented some of the hesitations or discomfort students often have about commenting on their peers (Hanrahan et al., 2001). Moreover, the training emphasised that just being critical was not good feedback either. Assessors were expected to provide a positive, as well as a critical, comment on each criterion. This might have enhanced the encouraging and motivating nature of the peer feedback.

Finally, in our study, the lack of trustworthiness and doubts about credibility and accuracy of peer feedback was also a major reason to choose staff feedback. This is an issue that is repeatedly reported in the literature as a disadvantage of peer feedback (Yang et al., 2006; Zhang, 1995). Staff

feedback is valued because it can be trusted to be correct and complete. The teacher is an expert in the domain, while the peer is only a novice. This finding also appeared in the study of Yang et al. (2006), who found that although students recognised the value of peer feedback, staff feedback had more impact on their revision because it was considered trustworthy. A second reason related to this issue is that the teacher is also the assessor of the summative assessment. Therefore, he is the best source of information about what is expected from the final paper. It is obvious that peers cannot easily substitute for the teacher on this issue (Sadler, 1998). As a consequence, this value of staff feedback can be seen as being test driven.

The finding that students in our study were particularly in need of more steering could explain why a large group of students preferred staff feedback if they had to make a choice, even if they also recognised the value added by peer feedback. This might actually be the explanation for the significant increase in the number of students that chose staff feedback after assignment three. At that time there was only one assignment to go, so the fact that it was the last opportunity for receiving feedback before the final report had to be handed in might have been a reason for some students to change their minds and to prefer staff feedback above peer feedback. The test driven nature of the staff feedback seems to become a stronger argument when the test becomes closer. Assessment steers learning, or "the tail wags the dog" (Dochy & McDowell, 1997, p. 219), and this seems not only true for students' approaches to learning, but also for their preferences for a certain type of feedback.

A second reason that can explain this shift in preference was revealed by the qualitative analysis: the time pressure students experienced at the end of the semester. Peer feedback requires a considerable investment of time and, when facing a high workload, some students might have decided to take the easiest and safest way out.

We can conclude that both sources of feedback were shown to have their own strengths and weaknesses from the perspective of the student. The forced choice question was only a starting point to study the details of both sources of feedback in more depth. Peer and staff feedback were shown to be complementary, and they each even provided the conditions under which the complementary source became better. Peer feedback took care of the need for specific individual feedback, and thus allowed staff feedback to concentrate more on correcting general misconceptions, explaining the difficult concepts

and providing a broader overview of all expectations in a collective session. Moreover, by providing staff feedback collectively, teacher resources were saved to organise and guide a strong qualitative peer feedback system, which offered opportunities for personal coaching, cooperative learning and metacognitive growth. Finally, given the provision of a backup of individual consultation opportunities with staff members in case of serious doubts or problems, the perceived lack of coaching might possibly be attributed to a phase of habituation of first-year students to an independent learning approach that is expected in higher education. Creating this kind of gap in immediate and complete help by suppressing staff feedback until after the peer feedback session might even have been a stimulus for true discussions among peers during the peer feedback sessions, eventually resulting in deeper learning than receiving an easy answer from the teacher.

References

Atlas.ti Scientific Software Development (2006). Atlas.ti (Version 2.0) [Computer software]. Berlin.

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment & Evaluation in Higher Education, 27,* 427-441.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5,* 7-74.

Blackboard Inc. (2005). *Blackboard Academic Suite. Instructor Manual*. Washington, DC: Blackboard Inc.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58,* 438-481.

De Corte, E. (1996). Instructional psychology: Overview. In E. De Corte & F. E. Weinert (Eds.), *International encyclopedia of developmental and instructional psychology* (pp. 33-43). Oxford: Elsevier Science.

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

Falchikov, N. (1996). Improving learning through critical peer feedback and reflection. In *Different approaches: Theory and practice in Higher Education. Proceedings HERDSA Conference 1996.* Perth. (available at http://www.herdsa.org.au/confs/1996/falchikov.html).

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

Gibbs, G. & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education, 1,* 3-31.

Gielen, S., Dochy, F., & Onghena, P. (2007). An inventory of peer assessment diversity. In S. Gielen, *Peer assessment as a tool for learning* (pp. 67-94). Unpublished doctoral dissertation, Katholieke Universiteit Leuven, Leuven, Belgium.

Gielen, S., Dochy, F., Onghena, P., Struyven, K., Smeets, S., & Decuyper, S. (2007). Goals of peer assessment and their associated quality concepts. In S. Gielen, *Peer assessment as a tool for learning* (pp. 41-66). Unpublished doctoral dissertation, Katholieke Universiteit Leuven, Leuven, Belgium.

Gielen, S., Tops, L., Dochy, F., Onghena, P., & Smeets, S. (2007). Peer feedback as a substitute for teacher feedback. In S. Gielen, *Peer assessment as a tool for learning* (pp. 95-124). Unpublished doctoral dissertation, Katholieke Universiteit Leuven, Leuven, Belgium.

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

Higgins, R. (2000). "Be more critical!": Rethinking assessment feedback. In *Paper presented at the British Educational Research Association Conference.* Cardiff University.

Jacobs, G., Curtis, A., Braine, G., & Huang, S.-Y. (1998). Feedback on student writing: taking the middle path. *Journal of Second Language Writing, 7,* 307-317.

Leki, I. (1990). Potential problems with peer responding in ESL writing classes. *CATESOL Journal, 3,* 5-19.

Magin, D. (2001). Reciprocity as a source of bias in multiple peer assessment of group work. *Studies in Higher Education, 26,* 53.

Neuman, W. (1994). *Social research methods: Qualitative and quantitative approaches*. Boston: Simon & Schuster.

Nilson, L. B. (2003). Improving student peer feedback. *College Teaching, 51,* 34-38.

Patton, M. Q. (1987). *How to use qualitative methods in evaluation*. Newbury Park: Sage.

Pope, N. (2001). An examination of the use of peer rating for formative assessment in the context of the theory of consumption values. *Assessment & Evaluation in Higher Education, 26,* 235-246.

Prins, F., Sluijsmans, D., & Kirschner, P. A. (2006). Feedback for general practitioners in training: Quality, styles, and preferences. *Advances in Health Sciences Education, 11,* 289-303.

Pryor, J. & Lubisi, C. (2002). Reconceptualising educational assessment in South Africa - testing times for teachers. *International Journal of Educational Development, 22,* 673-686.

Robinson, J. M. (2002). In search of fairness: An application of multi-reviewer anonymous peer review in a large class. *Journal of Further and Higher Education, 26,* 183-192.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in education, 5,* 77-84.

Sadler, P. & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11,* 1-31.

SAS Institute Inc. (2004). *SAS/STAT 9.1 User's Guide*. Cary: SAS Institute Inc.

Schelfhout, W., Dochy, F., Janssens, S., Struyven, K., & Gielen, S. (2006). Towards an equilibrium model for creating powerful learning environments. Validation of a questionnaire on creating powerful learning environments during teacher training internships. *European Journal of Teacher Education, 29,* 471-505.

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

Sluijsmans, D. (2002). *Student involvement in assessment. The training of peer assessment skills.* Unpublished doctoral dissertation. Open Universiteit Nederland, Heerlen.

Topping, K. J. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy, & E.

Cascallar (Eds.), *Optimizing new modes of assessment: In search of qualities and standards* (pp. 55-87). Dordrecht: Kluwer Academic.

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment & Evaluation in Higher Education, 25,* 149-169.

Weaver, M. R. (2006). Do students value feedback? Student perceptions of tutors' written responses. *Assessment & Evaluation in Higher Education, 31,* 379-394.

Wiliam, D. (2006). Does assessment hinder learning? In *Speech delivered at the ETS Europe breakfast salon (11th July 2006).*

Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15,* 179-200.

Zhang, S. (1995). Reexamining the affective advantage of peer feedback in the ESL writing class. *Journal of Second Language Writing, 4,* 209-222.

Appendix A

Description of the current peer assessment (PA) design and procedure

Cluster	Variable	Dimensions & range of Variation
Cluster I Decisions concerning the use of PA	Setting	PA is used in an introductory course in educational science (formal learning, 1st year of university program in Educational Sciences). It is a case-based course with individual assignments and a minimum of lectures. The course goal is to let students become acquainted with the field of pedagogical work in three domains (family, school and adult learning). The course is taught to 192 students (93% female, aged 18-20), and they all participate in peer assessment. The total group is divided in three subgroups, each studying a different case and receiving feedback from a different staff member. Each case contains pedagogical issues related to the three domains.
	Object	The course consists of four cumulative assignments, in which students analyse a pedagogical demand (derived from their case) in one domain and the available supply for it. Each assignment is one chapter (artefact) of their final report. Typical of the assignments is that they are rather open, so there is no black or white answer to whether a solution is correct. The assessment must focus on the reasoning a student has used to arrive at a certain solution, and the justifications that are made for the choices. Each student also chooses a different starting point within the case, so no two papers are the same. PA is applied to the drafts of the assignments, and only pays attention to the 'outcome' in writing.
	Frequency & Experience	PA is applied to the first three assignments (the first application is integrated in the training session, see below). PA is novel to most students.
	Objectives	PA is mainly applied as a tool for learning, in the sense described by Gielen et al. (2007f). A second goal is to function as a teaching method to also let students learn about the domains that their peers are studying, because the learning goals of the course cover three domains, and each peer only focuses on one in his own assignments. A third goal is social control, encouraging students to start working on the cumulative assignments in time. Fourthly, staff hope to equip students with some important learning strategies like asking for feedback from peers and collaborative learning that will be useful in the rest of their academic studies. And finally, a social goal is aimed at for these freshmen at university: by linking them to two peers from the beginning of the 1st semester, it is hoped that those who do not know yet any peers in their class will find a safe base for their first weeks and months.
	Function	Formative, and therefore explicitly called 'peer feedback'

		instead of peer assessment or peer evaluation.
Cluster II Link between PA and other elements in the learning environment	Alignment	PA is very central in the teaching. The teaching method for this course is presented to students as having two main properties: being case based and leaning on peer feedback. Considerable energy of the staff team is awarded to organising PA (training session, development of suitable electronic environment, presence at oral feedback sessions, and quality control of peer feedback). Some of the learning goals (i.e. those concerning the two 'other' domains) are only 'taught' through PA.
	Relationship to other assessments	PA is supplementary to collective staff feedback, which is organised after PA. At the PA sessions, students can not ask questions of staff: they first have to attempt to solve them with peers, and only ask staff for help a few days later at staff feedback. Summative assessment is postponed until the end of the course, when all four final assignments are merged.
	Scope of involvement	Aspects of involvement: providing of feedback, monitoring/guiding of a peer's progress Extent of involvement: assessment criteria are provided by staff together with the description of the assignment (on Blackboard[2]); individual feedback is the main responsibility of the student.
Cluster III Interaction between peers	Output	Nature of information: qualitative plus an indicative rating on a four-point scale for each part of the assignment (colour in no-1-2-3 stars), see Appendix D. Extent of 'condensation': each assignment had three main parts, each part is discussed separately, but different criteria for each part are discussed holistically. Feedback stance: students were trained to take a probing or collaborative stance, but that does not mean they did.
	Directionality	PA is organized in groups of three, and is provided mutually.
	Privacy	No anonymity of assessor/ee Teacher can read written feedback on Blackboard, and is present at oral feedback, but 20 groups give feedback simultaneously, so teacher does not take part in, nor actively follows oral discussions (some groups are filmed for research purposes). Feedback is not public (confidential within groups of 3 + staff team).
	Contact	Feedback is provided in writing (prepared at home and exchanged through Blackboard, which is also used to exchange the draft assignments) as well as discussed orally in their groups (during a two-hour session) in a pc lab, to provide access to all documents on Blackboard plus the Internet. Written feedback has to be available the night before the oral discussion.

[2] an electronic learning platform (Blackboard Inc., 2005)

	Role of assessee	Assessees are expected to participate actively in the discussion, and afterwards to revise their work.
Cluster IV Composition of the assessment groups	Matching	Students are matched within the same subgroup (these were randomly assigned initially): they work on the same case, but each chooses a different domain, so their papers are parallel but not identical. Groups are self-chosen and remain fixed for the course as a whole.
	Constellation of assessors & assessees	Unit of assessor: individual Unit of assessee: individual Number of assessors per unit of assessee: 2 Number of assessees per unit of assessor: 2
Cluster V Management of the assessment procedure	Format	Written feedback is provided on a prestructured feedback form (see Appendix D) addressing the three parts of the assignment, and a section for additional remarks. The form mentions the appropriate criteria for each part and provides the 3 open stars for quantitative feedback. Guidelines for oral feedback are discussed in the training session, and repeated at the beginning of the first oral session.
	Requirement	Compulsory
	Reward	Conditional reward: see quality control
	Training/ Guidance	Students are carefully prepared for the requirements of PA by means of a training session (integrated with the peer feedback on assignment 1*) in which they learn and practice to give feedback in a constructive way (always giving concrete positive & negative remarks, including an explanation, and preferably also a suggestion for improvement or a critical open question). They also learn and practice being an active feedback receiver (listen, summarize, ask questions) during the oral feedback sessions, to ensure that these discussions can add value to the written feedback. Assessment criteria for the first assignment are constructed by the students under the guidance of a tutor. * Assignment 1 started as group work, to allow peers to get acquainted with each other. Peer feedback on the first assignment was provided by another student from the class who was present at the same training session (6 parallel sessions were organised).
	Quality control	The thoroughness of the feedback is part of the summative mark for the feedback giver (a quarter of the total mark is assigned to it), to stimulate effort and to justify the investment of time.

Appendix B

If you compare the peer feedback on assignment x with the staff feedback on assignment x, whose feedback was in your opinion…

1. *most specific (directed toward specific mistakes and strengths)?*
2. *most comprehensible (well-argued)?*
3. *most inspiring to revise your assignment (with suggestions how to improve)?*
4. *most informative (who gave you most new insights into the subject)?*
5. *most encouraging to continue?*
6. *most thought-provoking?*

Appendix C

1. Function: motivating

1a. I felt motivated to engage in this course.

1b. The peer feedback system contributed to my motivation to engage in this course.

1c. The staff feedback sessions contributed to my motivation to engage in this course.

2. Function: activating

2a. I have studied the contents of this course (pedagogical demands, supplies and key concepts) thoroughly.

2b. The peer feedback system contributed to my activation to study the contents of this course thoroughly.

2c. The staff feedback sessions contributed to my activation to study the contents of this course thoroughly.

3. Function: coaching

3a. I felt supported (coached) on this course.

3b. The peer feedback system contributed to my coaching on this course.

3c. The staff feedback sessions contributed to my coaching on this course.

4. Function: steering

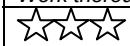
4a. I had a clear view of the expectations for the various assignments on this course.

4b. The peer feedback system contributed to a clear view of the expectations for the various assignments on this course.
4c. The staff feedback sessions contributed to a clear view of the expectations for the various assignments on this course.
5. Function: structuring
5a. I gained a clear view of the main thread (cohesion) in the successive assignments on this course.
5b. The peer feedback system contributed to a clear view of the main thread in the successive assignments on this course.
5c. The staff feedback sessions contributed to a clear view of the main thread in the successive assignments on this course.
6. Function: metacognitive support
6a. When working on the assignments for this course, I gained an insight into my own way of learning and solving assignments.
6b. The peer feedback system contributed to the fact that I gained an insight into my own way of learning and solving assignments.
6c. The staff feedback sessions contributed to the fact that I gained an insight into my own way of learning and solving assignments.

Appendix D

Peer feedback

Help each other to make an (even) better assignment

ASSIGNMENT … : ANALYSIS OF …	
Provider of feedback (name and student number): **Author of the assignment (name and student number):** *For each criterion color the right amount of stars and explain what was right or could be better and why. Always suggest what your peer could do to improve his/her assignment. Always add a question for further reflection on the assignment.* *Work thoroughly and elaborately; be critical, honest and subtle!*	
 Perhaps you haven't thought of this yet, because ... ? If I were you I would … .  This can be better because ... . A tip: maybe you could … ?  You did very well because ... . It can even be better if you … .  Wauw! I'm amazed! Because	
– **Criterion 1: ...** (this includes: ... .)	
– **Criterion 2: ...** (this includes: ... .)	
– **Criterion 3: ...** (this includes: ... .)	
– **Other comments, suggestions or questions, …**	

CHAPTER 8

FINAL REFLECTIONS

FINAL REFLECTIONS

Conclusions and Discussion

Just like peer tutoring or cooperative learning, peer assessment as a tool for learning is not new in itself. It has probably existed from the beginning of schooling, and is also present in many informal learning contexts. One of the most important changes over the last 25 years, as Topping (2005) describes, however, has been a greater focus upon 'implementation integrity'. The conditions for good implementation and the effects of various organisational variables are questioned and described. Consequently, there is a growing body of research trying to find answers to questions such as: 'What can be the role of peer assessment?', 'Does it work?' and 'How can it be optimised?'. Hundreds of studies have revealed some pieces of the puzzle already. Most report positive reactions of students and teachers, but studies on reliability, validity or performance outcomes are not always in line with each other. Even if positive results are found, it is not sure whether they are generalisable. This can be explained by the large diversity in peer assessment practices that exist. As a consequence it is often unclear which feature of a specific practice is crucial to provoke the desired effect. To be able to give a clear answer about the potential of peer assessment, the different single studies have to be combined and exceeded.

Several reviews and a meta-analysis have attempted to fit all the pieces of the puzzle together in order to get a clear view on peer assessment effectiveness (Dochy et al., 1999; Falchikov et al., 2000; Kane et al., 1978; Lewin & Zwany, 1976; Mouton, Blake, & Fruchter, 1955b; Mouton, Blake, & Fruchter, 1955a; Sluijsmans, Dochy, & Moerkerke, 1998; Topping, 2003; Topping, 1998). All these meta-studies have their merits for the domain, but some lacunas still prevent us from finding a clear answer to some of the questions.

The reviews on peer assessment by Mouton and colleagues, dating back to 1955 (Mouton et al., 1955a; Mouton et al., 1955b), gave a first systematic overview of findings on peer assessment, thereby referring to studies as old as from the years 1920. However, the authors only discussed peer assessment from a pure sociometrical perspective. As a consequence,

this review gave very few information on the educational applications of peer assessment. In 1976 a second review appeared (Lewin et al., 1976), this time with a more narrow focus on peer rating or peer nomination (also called 'buddy rating' in the article). This review summarized in a qualitative way the available studies on validity and reliability, factors affecting the peer evaluation process, the theoretical status of the peer evaluation process and finally, on feasibility, acceptability and cost effectiveness speculation. Lewin and Zwany, however, did not distinguish different methods of peer evaluation. As a result, they compared apples or oranges to some extent. The different methods of peer assessment were taken as the focus of the review by Kane and Lawler (1978). Their review discussed the pros and cons of the three at that time distinguished methods of peer assessment: peer nomination, peer rating and peer ranking. The authors discussed their 'effectiveness', defined as 'practicality of design, administration and scoring (i.e., efficiency), reliability, concurrent or predictive validity, freedom of bias and acceptability by users'. A limitation of this review was that it dealt with applications outside education, mostly within human resources management environments. Their findings might not be generalisable to educational applications of peer assessment. The reviews of Dochy and colleagues (1999) and Sluijsmans and colleagues (1998) discussed the validity, fairness, accuracy and effects of peer assessment within education. They distinguished between applications where peer assessment was used on its own, where it was combined with self-assessment, and finally where peer and self-assessment were complemented with staff assessment in co-assessment. However, they did not consider other differences in a peer assessment design, which can again be considered a limitation of these reviews. The review study of Topping (1998) recognised this need to pay more attention to the differences within peer assessment. Based on different studies, Topping (1998) developed a typology of variables on which peer assessment designs differ. Topping was the first one to recognise the problem that many individual studies lack detailed information about their peer assessment practice, which made it difficult to compare them. In his review, Topping used different ordering principles to summarise the results, such as type of effects (cognitive, metacognitive, affective, social and transferable skills, systemic, and disadvantages), psychometric requirements, object (writing, oral presentation skills, group work and projects, and professional skills), setting (subject area of professional skills), or type of contact (computer-

assisted). However, this list of ordering principles was incomplete, leaving several of the variables in the typology unaddressed. Furthermore, the ordering principles that Topping (1998) used to categorise the studies were not applied in a systematic way (e.g., the computer-assisted approach of peer assessment is not 'crossed' with the different objects or settings but was discussed at the same level as a new category next to for instance peer assessment of writing). Due to these two lacunas, some crucial questions regarding peer assessment remained unanswered. In his review of 2003, Topping again discussed the reliability, validity, utility and effects of peer assessment but added an analysis of the processes through which peer assessment realises learning benefits, and an analysis of the social and communication issues that are a prerequisite or a possible threat to peer assessment success. In his discussion of the effects of peer assessment, Topping (2003) used similar ordering principles as in his previous review, thereby retaining also the same limitations. Finally, only one meta-analysis has been performed in the domain of peer assessment (Falchikov et al., 2000). Unfortunately, it focuses on a single subset of peer assessment applications available, namely those involving peer marking or peer grading. Falchikov and Goldfinch studied the effects of seven variables (subject area, object of assessment, level of the course, format of assessment, nature of criteria, number of peers and faculty, and quality of research design) on the comparability between peer and teacher marks. This study of Falchikov and Goldfinch is the first review in the domain of peer assessment that systematically addresses the impact of differences between peer assessment practices. Unfortunately, however, their meta-analysis is not very informative to the designer of peer assessment environments, because most independent variables that are taken into account in their study are not design variables: variables such as subject area or level of the course cannot be changed by a teacher. The meta-analysis primarily examines in which context peer assessment has the highest chance of success, where success is defined as a high consistency between peer and teacher marks. This review gives few information on how a peer assessment design can be optimised if the setting is a given. Moreover, the focus of optimisation is restricted to reaching a higher agreement between peer and staff marks, which in itself is only a restricted view on 'quality of peer assessment'.

With the expansion of peer assessment research in the last decade, peer assessment started to take a multitude of different faces, and to serve

divergent goals. In the available research syntheses described above, effects of different types of peer assessment were thrown together, and apples and oranges were being compared. When positive outcomes of peer assessment were reported, it was not clear under what circumstances (what type of peer assessment) these outcomes might be expected to occur. Moreover, also different categories of outcomes (such as affective effects, cognitive benefits, gains of social skills, or validity of an assessment) were not always clearly distinguished, making it difficult to reach an overview of the domains in which peer assessment can be expected to play a beneficial role.

This dissertation showed that the main problem underlying the lacunas of previous reviews was that some basic elements for a comprehensive research synthesis such as a meta-analysis on peer assessment were missing. There was no framework to categorise the dependent variables of the individual studies, and no comprehensive overview of independent variables that should be taken into account to represent the specificity of each individual study. This dissertation aimed at the clarification and covering of this theoretical deficit. Chapter 2, 3 and 4, reported on the contributions to achieve this goal.

Chapter 2 delineated the role that peer assessment can play in raising the consequential validity of an assessment system. Although 'consequential validity' has been accepted widely as a 'new' quality criterion for educational assessment, it is a very open concept. In a strict sense, it only suggests that assessment should meet the goals it intends and it should not have unexpected effects that are undesirable. In the context of classroom assessment, this definition has been narrowed to having a positive, or at least not a negative, influence on student learning, also referred to as 'assessment as a tool for learning' (Dochy et al., 1997). Although peer assessment was 'believed' to be able to play a role in this type of consequential validity (Dochy et al., 1999), a clear theoretical framework to study this topic was lacking. This dissertation clarified the type of effects that assessment in general can have on learning, and formulated design principles for how to increase the consequential validity of an assessment system. These design principles appeared to be the missing link to understand how peer assessment is related to consequential validity. Peer assessment should not be considered as an assessment method in se, which has a consequential validity of its own, since it is always attached to a 'parent'-assessment method that defines the content, the modalities and the criteria of the assessment. Peer assessment

only adds one extra feature to the original features of the 'parent'-assessment method, namely that peers function as assessors. However, by adding this extra feature, peer assessment has the potential to increase the consequential validity of the new 'assessment system'. It helps to meet the identified design principles that enhance consequential validity. More specifically, this dissertation showed that peer assessment can make it more feasible to include challenging and authentic tasks in one's assessment system; it can help making the assessment demands more clear to the students; it can provide a supplement or a substitute for formative staff assessment; and finally, it can support the response to teacher feedback.

The contribution of peer assessment to the consequential validity of the assessment system, however, is not the only quality concept that is used in the literature to evaluate the success of peer assessment. Chapter 3 described the problem that is faced in the literature of a multitude of sometimes competing quality conceptualizations and associated quality criteria for peer assessment, providing a cluttered picture of the quality question regarding peer assessment. A thorough investigation of the link between all these different quality concepts and the underlying goals of peer assessment they are aiming at, appeared to be useful to clarify the picture. Chapter 3 offered an outline of five main goals that peer assessment may serve. The most obvious goal is its use as assessment tool. Also the learning goal of peer assessment has been well-established. Investigating the literature more closely yielded three additional goals: installation of social control in the learning environment; preparation of students for self-monitoring and self-regulation in lifelong learning; and active participation of students in the classroom. Our review of the literature showed that each of these goals results in different quality criteria. Only the criteria that are congruent with the goal that one is trying to achieve should be considered when evaluating the quality of peer assessment.

The final conceptual contribution in this dissertation addressed the problem of a practitioner or a researcher facing a design task or an analytical challenge with regard to peer assessment. Beyond the decision or description of the goal and the associated quality criteria for peer assessment, a pile of other characteristics need to be addressed. In 1998, Topping already provided a first overview of variables on which peer assessment designs may differ, which he called a typology of peer assessment. His typology proved useful for the practitioner to get an overview of important decisions to take, or

possible alternatives for an existing practice. Also for applied research, such a typology is useful for a comparison of different peer assessment settings. Since Topping's literature search on peer assessment, that collected all relevant papers between 1980 and 1996, the number of studies of peer assessment research has increased fast. Depending on the exact 'search string' and database used for the literature search, between 39% and 92% of the studies on peer assessment are found to have been published after the original literature search of Topping. This finding clearly indicated that there was a need to check Topping's typology against the recent literature, and to update it if necessary.

In Chapter 4 of this dissertation, a new inventory of peer assessment diversity has been developed. Based on a review of the recent literature, eight new variables were added to the typology of Topping and another eight variables were extended with extra subdimensions. Five original variables of Topping were absorbed in larger entities, and also the implementation factors of Topping were given a place within the variables of the inventory. Finally, the 20 resulting variables were grouped into five clusters, building on an earlier clustering by van den Berg, Admiraal, and Pilot (2006b). To researchers, this inventory may serve as a guideline for providing the necessary information on the particularities of their peer assessment design. Moreover, the framework developed in Chapter 4 may help to clarify the conceptual confusion that originates from the use of a single term to cover a multitude of sometimes incompatible practices.

This dissertation provided the basic elements needed for a synthesis of peer assessment research, namely an overview of possible quality criteria as dependent variables in Chapter 3 and an inventory of possible independent variables in Chapter 4. It offered a sketch of the model that may guide the puzzler. Further research is needed to complete the puzzle of peer assessment effectiveness, and attempt to formulate an answer to the questions raised in the beginning of this chapter, such as 'Does peer assessment work?' and 'How can we optimise it?'. One difficulty that this further enquiry will have to face, although, is that several existing studies will be pieces of the puzzle that lie upside down on the table, with only their contours visible, due to their incomplete description of the particularities of their peer assessment practice. Hence, we conclude the discussion of the conceptual part of the dissertation with a plea for goal clarification and detailed descriptions of peer assessment practices in the research literature.

Beyond its contribution to the general theory on peer assessment, this dissertation also made a contribution to the empirical investigation of the effectiveness of some options in a peer assessment design. In these empirical studies, a focus had to be determined concerning the goal of peer assessment. This focus immediately indicated one of the limitations of this dissertation: by focusing on one goal, the other goals of peer assessment would not be addressed in the empirical part of this work. Peer assessment as a tool for learning was taken as the main focus. The framework that was developed in Chapter 3 made it clear that by choosing this goal, the main quality criterion that would have to be addressed was the consequential validity of peer assessment, meaning its contribution to the learning environment and its effects on students learning processes and outcomes.

The studies reported in Chapters 5, 6 and 7 examined the impact of different values on three of the variables of the inventory of peer assessment diversity described in Chapter 4. Chapter 5 addressed the relationship with other assessments and the role of the assessee; Chapter 6 dealt with the role of the assessee and the format of peer assessment; and Chapter 7, finally, looked into the relationship with other assessments once again. Nevertheless, it should be noted that the choices that were made regarding the other variables of the inventory are extremely important to frame the results obtained in these studies. For instance, all studies implemented formative peer assessment in the middle of a two stage assignment, so that students had time for revision after receiving feedback; and all studies used written assignments as the object of peer feedback. Therefore, each chapter included a description of all specific arrangements within all 20 variables of the inventory. The reader should keep these descriptions in mind when reading the summary of the results reported here.

Chapter 5 showed that, within the given circumstances, formative assessment (or feedback) provided by staff compared to formative peer assessment did not result in better learning outcomes after a semester. Peer feedback might thus be considered as a worthy substitute of staff feedback. Both, however, proved to be surpassed when the role of the assessee was extended with the requirement to use a question form or a reply form. The condition where assessees first indicated their needs to the assessor by means of a question form showed to lead to a higher progress than the control condition of teacher feedback and the plain peer feedback condition. The

condition with the reply form in which students justified their use of the received peer feedback appeared to be significantly more effective than the plain peer feedback condition, but not better than the control condition with teacher feedback. A possible explanation for the effect of the question form is that assessors may provide more useful feedback when they are informed about the questions and doubts of the assessee beforehand. Moreover this feedback may receive more attention from the assessee, since it addresses personal questions and doubts. This is in line with earlier findings of Bangert-Drowns, Kulik, Kulik, and Morgan (1991) on the importance of a 'mindful reception' of feedback. In the case of the reply form, an explanation might be that it fostered reflection on the received feedback and the necessary revisions, realising a 'closed feedback loop' (see Boud, 2000).

The finding that the extended roles of the assessee increased performance, however, could not be replicated by the study in Chapter 6, which used exactly the same extensions within the role of the assessee. Since no clear explanations for these contradictory findings could be identified, a new replication study is needed that examines the impact of the role of the assessee again. Another suggestion for further research is to explore the effect of these different roles of the assessee, when applied to staff feedback. When assessees use a question form or a reply form prior to or at the end of receiving staff feedback, does this elicit the same processes as when they are used with peer feedback?

Although the role of the assessee did not influence the learning outcomes in the second study (Chapter 6), this study showed that the impact of peer feedback on the learning outcomes increased when student-assessors incorporated more 'constructive' elements into their feedback. If the received feedback was more specific, more appropriate to the assessment criteria, contained positive as well as negative comments, and in addition included some justifications, suggestions and thought-provoking questions, the assessee made better revisions, resulting in a higher progress between the draft and the final version of the essay. These results are in line with earlier reports on the importance of the type of information that is included in feedback (e.g., Bangert-Drowns et al., 1991; Flower et al., 1986; Narciss, 1999). An interesting question to be addressed in further research is which components of the 'Feedback Constructiveness Index' are most important in realising this effect.

A consequence of the finding that the composition of feedback is important for its impact, is that measures that stimulate or support students in providing more constructive feedback, should be able to raise the effectiveness of peer feedback. In the third study, reported in Chapter 7, several of these measures were taken: the social interaction between peers was increased in the design of peer assessment to emphasise the combination of individual accountability and positive interdependency in order to motivate students to do their best for each other (see Slavin, 1989; Sluijsmans, 2002); students received a more intensive training in providing constructive feedback to make them more competent assessors; this training also taught assessees how to make sure themselves that they received the feedback they needed; and finally a quality control system was set up in which student-assessors would be rewarded for good feedback and punished for poor feedback. Although the comparison may be compounded with the influence of the higher level of education in which the study was situated, it was clear that the level of constructiveness of the feedback provided in the third study was much higher than in the second study (the FCI score increased from an average of 5.9 out of 14 to 9.1).

Whereas the study in Chapter 6 showed that the effectiveness of peer feedback rose when it had more constructive characteristics, even without taking account of the correctness or completeness of the feedback, the study in Chapter 7 made clear that this correctness or completeness is however an important feature of good feedback from the student perspective. The results of this study confirmed that there is still a complementary role for staff feedback, even if peer feedback definitely has an added value on its own. Even though peer feedback showed to be a worthy substitute of staff feedback in the first study, the third study – taking a more in-depth look – displayed that both sources did not provide identical feedback. Going beyond a mere outcome measure in terms of improvement in marks, this study clearly identified differences in the characteristics and the functions of both sources of feedback. When students were asked to state their preference for one of both, they hesitated because each source had it own strengths and weaknesses. Peer feedback is more specific, and is better for activating, motivating and coaching students; staff feedback is more trustworthy, and it helps to understand the assessment requirements and the structure of the course. Moreover, it can correct peer feedback if peers make inappropriate suggestions.

We argue that by combining individual peer feedback with collective staff feedback, the result of 1 and 1 is not 2 but 3, since the presence of each source provides the conditions under which the complementary source became better. Peer feedback met the need for specific individual feedback, and thereby allowed staff feedback to concentrate more on correcting general misconceptions, explaining the difficult concepts and providing a broader overview of all expectations in a collective session. By providing collective staff feedback on the other hand, teacher resources were saved to organise and guide a strong qualitative peer feedback system, which offered opportunities for personal coaching, cooperative learning and metacognitive growth.

Since half of the students in the third study chose peer feedback in a forced choice, it is clear that peer feedback was no second-class feedback. On the other hand, trying to save time by substituting staff feedback by peer feedback would also be a deterioration of the learning environment when there is no backup for those functions that peer feedback cannot fulfil, such as structuring. Moreover, trying to save time by not investing staff resources in the organisation and guidance of the peer feedback system also is a threat to the generalisation of the success of peer feedback as illustrated by the third study in this dissertation.

Finally, some questions were left unexplored by this dissertation. The three empirical studies did not address the question of individual differences between students. Is there a relationship between for instance students' performance level, learning approach or motivation and students' perceptions of, preferences for and learning benefits from peer assessment? Is there an impact of for instance students' verbal abilities or metacognitive skills on the quality of the feedback they provide?

Moreover, this dissertation did only provide limited information on the long term effects of peer assessment on performance, perceptions, and preferences. In the third study, we found a significant decrease in preferences for peer feedback after the second experience with it. Although we expected that the reason for this was time pressure as well as a stronger influence of the upcoming summative assessment, one might also think of a disappearing halo-effect (see methodological issues). Further research might address this question of the impact of a longer experience of peer assessment.

Implications for Educational Practice

This section summarises the major implications for educational practice, which have already been discussed more extensively. The research reported in this dissertation has broadened and deepened the view on possible uses of peer assessment. Peer assessment should not only be considered as an assessment tool, it can also function as a tool for social control, a tool for learning, a tool for learning-how-to-assess and a tool for raising active participation of students. One of the merits of implementing peer assessment as a tool for learning, is that it can actually increase the consequential validity of the 'parent'-assessment method to which it is attached, by making it more feasible to include challenging and authentic tasks in one's assessment system; by helping to make the assessment demands more clear to the students; by providing a supplement or a substitute for formative staff assessment; and finally, by supporting the response to teacher feedback.

It is important that practitioners reflect on their intended goals, when they use peer assessment, since the choice of a goal (or a combination of goals) also determines which quality criteria are appropriate to evaluate the effectiveness of one's peer assessment practice. Today, there is much confusion among practitioners about how they should design their peer assessment practice, and doubts about validity and reliability issues are often impediments to its use. This dissertation pointed to the fact that these design and quality questions need to be preceded by the decision on the intended goal. Only the criteria that are congruent with the goal that one is trying to achieve should be considered when evaluating the quality of peer assessment. These quality criteria on their turn can be used to evaluate the value of several design alternatives, for instance whether or not an oral discussion of the assessment can deliver an added value to its effectiveness in the light of the intended goal.

In this design phase the practitioner can use the inventory of peer assessment diversity that is developed in this dissertation as a checklist for important decisions to be taken. When designing a peer assessment practice for a particular setting, having a certain (combination of) goal(s) in mind, one should consider several elements: the object of the peer assessment, its frequency and students' prior experience, its function, its alignment to other components of the learning environment, its relationship to other

assessments, the scope of student involvement, its output, its directionality, arrangements concerning privacy and contact between assessors and assessees, the role of the assessee, the matching and constellation of assessors and assessees, the format of the assessment, whether or not it is compulsory, its reward, the training and guidance of students and finally its quality control system. Due to this multitude of variables on which one has to take a decision, it is clear that the term 'peer assessment' covers many practices. If practitioners want to compare their practices to other peer assessment applications, they should pay attention to all these different variables. Practitioners who are not satisfied with the effectiveness of their current approach, can use the inventory as a source of inspiration for possible alternatives to a specific practice.

The empirical studies in this dissertation showed that qualitative formative peer assessment (referred to as peer feedback), applied as a tool for learning, is no inferior form of feedback. Peer feedback might be considered as a worthy substitute of staff feedback. It might even lead to higher performance than staff feedback when it is extended with measures to enhance the 'mindful reception' of feedback by means of an a priori question form or an a posteriori reply form administered to the assessee. However, it is not yet completely clear under what circumstances these measures are most effective.

Moreover, this dissertation showed that it is important to stimulate or support students in providing more constructive feedback in order to raise the effectiveness of peer feedback. In order to be considered 'constructive', feedback should be specific, appropriate to the assessment criteria, contain positive as well as negative comments, and in addition include some justifications, suggestions and thought-provoking questions. Assessees who receive this type of feedback make better revisions, resulting in a higher progress between the draft and the final version of the essay. Examples of measures that can be taken to enhance the constructiveness of peer feedback are increasing the social interaction between peers; training peer assessors in providing constructive feedback; training assessees how to make sure themselves that they receive the feedback they need; or installing a quality control system in which student-assessors are rewarded for good feedback or punished for clearly poor feedback.

Feedback can reach a high level of 'constructiveness' without necessarily being correct or complete. Nevertheless, this dissertation showed

that the correctness or completeness is also an important feature of good feedback from the student perspective, and it is typically a characteristic of staff feedback. In a powerful learning environment, peer feedback and staff feedback can play a complementary role. Even when peer feedback showed to be a worthy substitute of staff feedback when considering their separate impact on progress in performance, taking a more in-depth look revealed that the feedback of both sources is not identical. Each source has it own strengths and weaknesses, without one being more preferable than the other. Peer feedback can be more specific, and is better in activating, motivating and coaching students; staff feedback is more trustworthy, and it helps to understand the assessment requirements and the structure of the course. Moreover, it can correct peer feedback if peers make inappropriate suggestions.

The presence of one source can even make the complementary source better. When peer feedback takes care of the need for specific individual feedback, it allows staff to organise their feedback collectively. As a result, teachers can concentrate on correcting general misconceptions, explaining the difficult concepts and providing a broader overview of all expectations, instead of dividing their attention over all individual students whereby time often forces them to restrict their individual feedback to the absolute minimum. Moreover, by providing staff feedback collectively, the saved teacher resources can be used to organise and guide a strong qualitative peer feedback system, which offers opportunities for personal coaching, cooperative learning and metacognitive growth.

Finally, practitioners should notice that implementing peer assessment as a tool for learning does not necessarily result in a saving of time. Trying to save time by substituting staff feedback by peer feedback would be a deterioration of the learning environment when there is no backup for those functions that peer feedback cannot fulfil, such as structuring in case of a complex assignment. Moreover, trying to save time by not investing staff resources in the organisation and guidance of the peer feedback system also is a threat to the success of peer feedback in a learning environment. Peer assessment, or more specifically peer feedback, can not replace the teacher in all his facets, nor is it likely to function on its own without a proper organisation behind the scene.

Final Reflection of Some Methodological Issues of This Research

This final section discusses some methodological issues related to this dissertation, and assumes that the reader is familiar with the different studies reported in this dissertation, and with the basic information regarding their methodology as provided earlier. The dissertation includes three types of research designs, each having their own merits and shortcomings, we will discuss the pros and cons of our literature reviews, our experimental designs of chapter 5 and 6 and finally our case study of chapter 7.

The first chapters are based on a literature review. The strength of this type of research is that it enables a researcher to build on the previous research in an area, and to transcend it in order to construct new insights. A limitation of our literature review is that we did not provide a synthesis of previous empirical findings. However, we argued that we would not yet be able to provide an added value to the existing reviews in the field, without first constructing a rather comprehensive framework in which the various types of peer assessment studies would fit. We therefore opted to address our attention to the construction of an overview of different foci in peer assessment research with regard to the goal and its associated quality concepts, and to make an inventory of the discerning variables for different peer assessment practices.

The differences between the research designs of the empirical studies in Chapters 5 and 6 on the one hand and 7 on the other hand, can be described by means of four contrasts. The first contrast (referred to as contrast A) is that of a (quasi-)experimental study comparing conditions (with or without a control condition) and a case-study design. The second contrast (contrast B) is that of a study where the researcher takes an 'outsider' position and, after implementing the necessary conditions, collects data from the background without intervening in the actual teaching process versus the researcher who is closely connected to the research setting and exerts a strong control over the ongoing teaching process (referred to as an 'insider' position). The third contrast (contrast C) concerns the data that are collected: 'objective' measures of process and output variables or more 'subjective'

self-reported perceptions. Finally, the fourth contrast (contrast D) addresses the type of analyses that are performed: quantitative or qualitative.

The first (Chapter 5) and the second (Chapter 6) study can be situated within a quasi-experimental approach (contrast A), taking an outsider position (contrast B), collecting objective measures of output (Chapter 5) or of process and output (Chapter 6), combined with a minor focus on student perceptions in Chapter 5 (contrast C). Finally, the data regarding the perceptions as well as the objective measures are analysed in a quantitative way (contrast D). The third study on the contrary was a case study (contrast A) conducted from an insider position (contrast B), collecting mainly perception data (contrast C) which are analysed in a quantitative as well as a qualitative way (contrast D).

The strength of the outsider approach (contrast B) of the first two studies is that it provides a view on what is possible in a realistic setting, with a 'normal' teacher who is not specifically interested in peer assessment, and without an enthusiastic researcher having considerable input in the ongoing teaching processes. Moreover, due to the quasi-experimental approach (contrast A) and the 'objective' measurements (contrast C), these studies provide 'hard evidence' of the impact of certain peer assessment features. Finally, the data collection in these studies was minimally intrusive (we collected student artefacts, and only at the end of the course a short questionnaire was administered), making students less aware of their status as a participant in a study, thereby avoiding to a certain extent the possibility of a Hawthorne effect, and minimizing the disturbance of the normal teaching and learning processes.

The weaknesses of this type of studies, however, are associated with the same features that also define their strengths. For instance, the outsider position (contrast B) and the main focus on output measures (contrast C), result in a limited access of the researcher to information about other factors that might have influenced the outcomes of the study (e.g., how did the teacher guide students in the peer assessment process, what happened when the time schedule became tight, what solutions were used for problems encountered on the way, what did students really think of the innovation?). Furthermore, although the use of 'average' classrooms and 'average' teachers can give a realistic view on the possibilities of peer assessment, just using one single teacher is still a major limitation. Having one and the same teacher teaching all conditions is a good way to exclude a differentiating teacher

effect between the conditions, but since there is only one teacher involved a generalisation to other teachers is still problematic: there might still be a general teacher effect that is related to for instance the way a certain teacher interacts with his students, or to his own beliefs in the effectiveness of the different conditions (Draper, 2006). One cannot guarantee that a similar study with another teacher would lead to the same results. To exclude this kind of teacher effect, several replications are needed with different teachers.

To some extent, the study in Chapter 6 can be considered a replication of the study in Chapter 5; however, the research questions did not completely overlap. In our search for a second research setting, however, we encountered the difficulty associated with conducting multiple replications of experimental studies in realistic educational settings (contrast A). Teachers are reluctant to allow the use of a control group design, in that they wish not to put some of their students at a possible disadvantage. The creation of experimental conditions suffers the same problem. Treating some of your classes different then others is a sensible issue in education, subject to ethical objections. Therefore, the use of multiple replication studies is not straightforward when adopting an experimental approach. Another disadvantage concerning the experimental approach is that, by splitting the available sample in several conditions, the sample size within each condition decreases considerably. And raising the overall sample size is only possible to the extent that additional classes can be found that are enrolled in the same curriculum and are taught by the same teacher. In Chapter 6, we intercepted this 'disadvantage' of a fairly small sample size by adopting a repeated measures approach to increase the power of the study. The study of Chapter 5 proved to have sufficient power to rely on the pretest and posttest data of the exams. Although a final risk of adopting an experimental approach (contrast A) within a normal educational setting is that 'conditions meet at the playground', and share information that is meant to be available only in one condition, this latter risk is less likely when it concerns several types of feedback, because students are not expected to share much of this personal feedback.

The strength of the case study approach (contrast A) in the third study is that it enables the researcher to get a more in depth insight in the topic and to develop a richer understanding of what is going on when implementing peer assessment in a classroom. The insider position of the researcher (contrast B) even boosts this strength, since it gives the researcher more

information on the disturbing influence of certain environmental factors and on the supportive role of some specific features of the innovation. Moreover, since the researcher can intervene when difficulties arise, and can provide a maximum of support to the collaborating teachers, this type of research reveals information on the effect that can be realised under 'ideal' circumstances, surpassing the growing pains of an innovation and exceeding the fact that not all teacher are as skilful in using it in the way as it was meant. Furthermore, this study investigated the perspectives of the first party involved, namely the students, which can provide a new and important insight in the innovation going beyond the objective measurements of effects (contrast C). Moreover, a triangulation of different methods of data-collection (closed-ended questions with different starting points, as well as an additional inductive approach based on open-ended questions), together with a variation in analyses techniques (quantitative as well as qualitative) (contrast D) makes it a strength of this third study that it provides a fine-grained insight in this student perspective, with the inductive approach leaving room for new directions that were not yet thought of when developing the deductive instruments.

At the same time, however, the research design of the third study also had certain disadvantages. The in-depth focus on one case (contrast A) excluded the collection of information regarding other peer assessment practices, that differ in one or several features from the present, such as the setting (university students in educational sciences), the contact arrangements (written and oral feedback), or the quality control measures (awarding marks for good feedback). And since there was no comparison possible with a control group, no 'hard evidence' could be collected about positive or negative effects in relation to a more 'traditional' teaching approach (contrast A). Moreover, the specific constellation of peers in the peer assessment arrangement (two assessors per assessee), and the specific relationship with staff assessment in this study (supplementary), did not allow to investigate the individual impact of the feedback of one peer or the teacher on performance outcomes. At the level of perceptions (contrast C), however, we could ask students to separate and compare both. Some might raise the question whether perceptions 'correspond' with reality. Nevertheless, we argue that students' perceptions, whether or not they correspond to a reality that can be observed objectively, are an important reality too, since they have

a large mediating impact on the learning processes and their outcomes (Entwistle, 1991).

Both the quantitative, deductive approach of these perceptions as well as the qualitative, inductive approach carry risks of misinterpretation (contrast D). The translation of our theoretical framework into Likert-scale items in the first approach impoverishes the meaning, and creates the risk that students do not 'read' the items in the same way as we intended them to represent constructs of our framework. On the other hand, letting students freely express their opinions does not escape from the interpretation risk, since in that case, the researcher might read student messages otherwise then they were intended by the students. We have to admit that having several researchers code and interpret the same qualitative data would have been a valuable addition. However, in this study we did choose for combining different measures in a mixed methods approach. The fact that we found converging results strengthens our confidence in the validity of the measures.

Finally, two 'expectation effects' (Draper, 2006) might have played a role in the results of the third study. Due to the fact that students were intensively questioned about their experiences and perceptions on peer assessment, there is no doubt that they were conscious about their status as a participant in a research. This fact alone might have had an effect on their behaviour and their perceptions: the so-called Hawthorne effect. However, inducing a Hawthorne effect might be the trade-off of every systematic attempt to collect in-depth research data in education. Since too many issues that are of interest to researchers of educational innovations cannot be observed objectively, asking students deliberately about their experiences and perceptions is inevitable in a first phase of the research. To control for a Hawthorne effect one can, in subsequent research phases, lower the control over the setting, and start to rely on more subtle sources of information that can be collected with less intrusive methods, and which can even be collected by the teacher, allowing the researcher to adopt an outsider position. However, in the first phase one has to explore too many issues to get an idea of where to look, and what to ask, and this is not feasible in a subtle way. To address the possible Hawthorne effect, we therefore suggest that a follow-up study of the same case tries to look for confirmation or denial of the conclusions that were formulated. Since there will be no researcher at the foreground and the learning environment will loose its status of 'educational

innovation', the chance that the new cohorts of students will consider themselves as 'participants of a study' will diminish.

The second possible expectation effect is associated to the position that the researcher took in the third study. To avoid using the authority of the staff members in stimulating students to take part in the data collection, in order to assure students beliefs in the confidentiality of their answers, the researcher used her 'personal enthusiasm' to convince students to participate. She got in touch with the students and was visible to them throughout the study, for example in the training session, the feedback sessions, the interviews and the questionnaire administration. Although one of the advantages of this approach was that we reached a remarkably high response rate for questionnaires that were administered outside class time, and we collected extended answers to the open questions, it also carried a risk. We noticed that several students started to sympathize with the researcher. An indicator of this sympathy was found in the open questions at the end of the questionnaires, in which students could add whatever comment they wished to make to the researcher. It happened that students included small jokes about the process or success wishes concerning the research. On the one hand, this indicates that participants felt at ease, increasing the possibility that they would freely speak their mind, but on the other hand it could lead to an 'experimenter effect' resulting in more positive answer patterns regarding peer assessment because they wanted to please the researcher. Although we also asked students repeatedly about negative experiences of peer feedback, thereby indicating that also this information was important to us, we cannot reject the possibility of a positive bias due to an experimenter effect. The solution that was described with regard to the Hawthorn effect would in the meantime also take care of this experimenter effect. So, this is an extra reason to perform a follow-up study as described earlier.

Finally, in all three studies a halo-effect might have played. This effect refers to the attractiveness of a novel experience. Students might have liked peer assessment because it was something new, not because of the intrinsic characteristics of peer assessment. If effects and positive perceptions are completely attributable to a halo-effect, we would expect that a long term exposure to peer assessment will decrease the positive results. Further research should therefore address this question of the long term effects of peer assessment.

At last, it should be noticed that a final strength of this dissertation is that it combined two approaches of empirical research. Some of the disadvantages of the first, experimental, outsider, objective and quantitative approach, such as the lack of a fine-grained view or the small sample size, are compensated by the advantages of the case-based, insider, perceptions-directed and mixed methods approach of the third study. The same reasoning goes for some of the disadvantages of the approach of the third study, like the experimenter effect or the lack of 'hard' evidence, that are compensated to a certain extent by the first study. Combining these studies, as was done in the previous section on the results of this dissertation, delivers an added-value to the singular studies. Triangulation of evidence across projects introducing similar innovations can enhance the level of credibility of that evidence, even when the combination of evidence is not as straightforward as in many conventional meta-analyses, which aggregate effect sizes from very similar experiments using the same outcome measures (Kember, 2003). This, however, thus not exclude that further research in peer assessment is still needed to reach full insight in the domain.

REFERENCES

Abson, D. (1994). The effects of peer evaluation on the behaviour of undergraduate students working in tutorless groups. In H. C. Foot, C. J. Howe, A. Anderson, A. K. Tolmie, & D. A. Warden (Eds.), *Group and interactive learning* (pp. 153-158). Southampton: Computational Mechanics.

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington: American Educational Research Association, American Psychological Association, & National Council on Measurement in Education.

Askham, P. (1997). An instrumental response to the instrumental student: Assessment for learning. *Studies in Educational Evaluation, 23,* 299-317.

Atlas.ti Scientific Software Development (2006). Atlas.ti (Version 2.0) [Computer software]. Berlin.

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment & Evaluation in Higher Education, 27,* 427-441.

Bangert-Drowns, R., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61,* 213-238.

Biggs, J. (1996a). Assessing learning quality: Reconciling institutional, staff and educational demands. *Assessment & Evaluation in Higher Education, 21,* 5-15.

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

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.

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

Birenbaum, M. (2007). Evaluating the assessment: Sources of evidence for quality assurance. *Studies in Educational Evaluation, 33,* 29-49.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5,* 7-74.

Blackboard Inc. (2005). *Blackboard Academic Suite. Instructor Manual*. Washington, DC: Blackboard Inc.

Bloxham, S. & West, A. (2004). Understanding the rules of the game: Marking peer assessment as a medium for developing students' conceptions of assessment. *Assessment and Evaluation in Higher Education, 29,* 721-733.

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

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education, 22,* 151-167.

Boud, D. (1986). *Implementing student self-assessment*. Sydney: Higher Education Research and Development Society of Australia.

Boud, D. (1994). The move to self-assessment: Liberation or a new mechanism for oppression? In P. Armstrong, B. Bright, & M. Zukas (Eds.), *Reflecting on Changing Practices, Contexts and Identities* (pp. 10-14). Leeds: Department of Adult Continuing Education, University of Leeds.

Boud, D. (1995). Assessment and learning: Contradictory or complementary? In P. Knight (Ed.), *Assessment for Learning in Higher Education* (pp. 35-48). London: Kogan Page.

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

Boud, D. & Brew, A. (1995). Developing a typology for learner self assessment practices. *Research and Development in Higher Education, 18,* 130-135.

Boud, D. & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education, 31,* 399-413.

Breitmeyer, B. J., Ayres, L., & Knafl, K. A. (1993). Triangulation in qualitative research: evaluation of completeness and confirmation purposes. *IMAGE Journal of Nursing Scholarship, 25,* 237-243.

Butler, D. L. & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65,* 245-281.

Carless, D., Joughin, G., & Mok, M. M. C. (2006). Learning-oriented assessment: principles and practice. *Assessment & Evaluation in Higher Education, 31,* 395-398.

Cheng, W. & Warren, M. (1999). Peer and teacher assessment of the oral and written tasks of a group project. *Assessment & Evaluation in Higher Education, 24,* 301-314.

Cheng, W. & Warren, M. (2000). Making a difference: Using peers to assess individual students' contributions to a group project. *Teaching in Higher Education, 5,* 243-255.

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98,* 891-901.

Cole, D. (1991). Change in self-perceived competence as a function of peer and teacher evaluation. *Developmental Psychology, 27,* 682-688.

Conway, J. M. & Huffcutt, A. I. (1997). Psychometric properties of multi-source performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10,* 331-360.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale: Erlbaum.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58,* 438-481.

Dart, B., Burnett, P. C., Purdie, N., Boulton-Lewis, G., Campbell, J., & Smith, D. (2000). Students' conceptions of learning, the classroom environment, and approaches to learning. *Journal of Educational Research, 93,* 262-270.

De Corte, E. (1996). Instructional psychology: Overview. In E. De Corte & F. E. Weinert (Eds.), *International encyclopedia of developmental and instructional psychology* (pp. 33-43). Oxford: Elsevier Science.

De Corte, E. (2000). Marrying theory and the improvement of school practice: a permanent challenge for instructional psychology. *Learning and Instruction, 10,* 249-266.

Deci, E. L. & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*.

Dierick, S. & Dochy, F. (2001). New lines in edumetrics: New forms of assessment lead to new assessment criteria. *Studies in Educational Evaluation, 27,* 307-329.

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

Dochy, F. & Moerkerke, G. (1997). Assessment as a major influence on learning and instruction. *International Journal of Educational Research, 27,* 415-431.

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.

Draper (2006). The Hawthorne, Pygmalion, placebo and other effects of expectation. http://www.psy.gla.ac.uk/~steve/hawth.html [On-line].

Entwistle, N. (1991). Approaches to learning and perceptions of the learning environment. *Higher Education, 22,* 201-204.

Entwistle, N. (2000). Approaches to studying and levels of understanding: The influences of teaching and assessment. In J. Smart (Ed.), *Higher Education: Handbook of theory and research (XV)* (pp. 156-218). New York: Agathon Press.

Falchikov, N. (1995a). Improving feedback to and from students. In P. Knight (Ed.), *Assessment for Learning in Higher Education* (pp. 157-166). London: Kogan Page.

Falchikov, N. (1995b). Peer feedback marking: Developing peer assessment. *Innovations in Education & Training International, 32,* 175-187.

Falchikov, N. (1996). Improving learning through critical peer feedback and reflection. In *Different approaches: Theory and practice in Higher Education. Proceedings HERDSA Conference 1996.* Perth. (available at http://www.herdsa.org.au/confs/1996/falchikov.html).

Falchikov, N. (1993). Group-process analysis - Self and peer assessment of working together in a group. *Educational & Training Technology International, 30,* 275-284.

Falchikov, N. & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70,* 287-322.

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

Forbes, D. & Spence, J. (1991). An experiment in assessment for a large class. In R. Smith (Ed.), *Innovations in engineering education* (pp. 97-101). London: Ellis Horwood.

Frederiksen, J. R. & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18,* 27-32.

Gibbs, G. (1999). Using assessment strategically to change the way students learn. In S. Brown & A. Glasner (Eds.), *Assessment matters in Higher Education: Choosing and using diverse approaches* (pp. 41-53). Buckingham: SRHE & Open University Press.

Gibbs, G. & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education, 1,* 3-31.

Gielen, S., Dochy, F., & Dierick, S. (2003). Evaluating the consequential validity of new modes of assessment: The influence of assessment on learning, including pre-, post-, and true assessment effects. In M. Segers, F. Dochy, & E. Cascallar (Eds.), *Optimizing new modes of assessment: In search of qualities and standards* (pp. 37-54). Dordrecht: Kluwer Academic.

Gielen, S., Dochy, F., & Onghena, P. (2007). An inventory of peer assessment diversity. In S. Gielen, *Peer assessment as a tool for learning* (pp. 67-94). Unpublished doctoral dissertation, Katholieke Universiteit Leuven, Leuven, Belgium.

Gielen, S., Dochy, F., Onghena, P., Janssens, S., Schelfhout, W., & Decuyper, S. (2007). A complementary role for peer feedback and staff feedback in powerful learning environments. In S. Gielen, *Peer assessment as a tool for learning* (pp. 157-199). Unpublished doctoral dissertation, Katholieke Universiteit Leuven, Leuven, Belgium.

Gielen, S., Dochy, F., Onghena, P., Struyven, K., Smeets, S., & Decuyper, S. (2007). Goals of peer assessment and their associated quality concepts. In S. Gielen, *Peer assessment as a tool for learning* (pp. 41-66). Unpublished doctoral dissertation, Katholieke Universiteit Leuven, Leuven, Belgium.

Gielen, S., Peeters, E., & Tops, L. (in preparation). The impact of a peer review experience on the use of self-regulation strategies.

Gielen, S., Tops, L., Dochy, F., Onghena, P., & Smeets, S. (2007). Peer feedback as a substitute for teacher feedback. In S. Gielen, *Peer assessment as a tool for learning* (pp. 95-124). Unpublished doctoral dissertation, Katholieke Universiteit Leuven, Leuven, Belgium.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. London: Falmer.

Habermas, J. (1987). *Knowledge and human interests*. (Translated by Shapiro, J. ed.) London: Polity Press.

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

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77,* 81-112.

Higgins, R. (2000). "Be more critical!": Rethinking assessment feedback. In *Paper presented at the British Educational Research Association Conference.* Cardiff University.

Hounsell, D. (1987). Essay writing and the quality of feedback. In J. Richardson, M. W. Eysenck, & D. W. Piper (Eds.), *Student Learning: research in education and cognitive psychology* (Milton Keynes: Open University Press.

Jacobs, G., Curtis, A., Braine, G., & Huang, S.-Y. (1998). Feedback on student writing: Taking the middle path. *Journal of Second Language Writing, 7,* 307-317.

Johnson, J., Olson, A., & Courtney, C. (1996). Implementing multiple perspective feedback: An integrated framework. *Human resource management review, 6,* 253-277.

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

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319-342.

Keller, J. M. (1983). Motivational design of instruction. In C. M. Reigeluth (Ed.), *Instructional design theories and models* (pp. 383-434). Hillsdale: Erdbaum.

Kember, D. (2003). To control or not to control: The question of whether experimental designs are appropriate for evaluating teaching innovations in higher education. *Assessment & Evaluation in Higher Education, 28,* 89-101.

Kim, M. (2005). *The effects of the assessor and assessee's roles on preservice teachers' metacognitive awareness, performance, and attitude in a technology-related design task.* Unpublished doctoral dissertation. Florida State University.

Kluger, A. N. & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin, 119,* 254-284.

Langan, A. M., Wheater, C. P., Shaw, E. M., Haines, B. J., Cullen, W. R., Boyle, J. C. et al. (2005). Peer assessment of oral presentations: effects of student gender, university affiliation and participation in the development of assessment criteria. *Assessment & Evaluation in Higher Education, 30,* 21-34.

Leki, I. (1990). Potential problems with peer responding in ESL writing classes. *CATESOL Journal, 3,* 5-19.

Lewin, A. & Zwany, A. (1976). Peer nominations: A model, literature critique and a paradigm for research. *Personnel Psychology, 29,* 423-447.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16,* 14-16.

Liu, C.-C. & Tsai, C.-M. (2005). Peer assessment through web-based knowledge acquisition: tools to support conceptual awareness. *Innovations in Education & Teaching International, 42,* 43-59.

Lockhart, C. & Ng, P. (1995). Analyzing talk in ESL Peer Response groups: Stances, functions and content. *Language Learning, 45,* 605-655.

Magin, D. (2001). Reciprocity as a source of bias in multiple peer assessment of group work. *Studies in Higher Education, 26,* 53.

Magin, D. & Helmore, P. (2001). Peer and teacher assessments of oral presentation skills: how reliable are they? *Studies in Higher Education, 26,* 287-298.

Malcolmson, C. & Shaw, J. (2005). The use of self- and peer-contribution assessments within a final year pharmaceutics assignment. *Pharmacy Education, 5,* 169-174.

Marcoulides, G. A. & Simkin, M. G. (1995). The consistency of peer review in student writing projects. *Journal of Education for Business, 70,* 220-224.

Martens, R. L. & Dochy, F. (1997). Assessment and feedback as student support devices. *Studies in Educational Evaluation, 23,* 257-273.

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

McGourty, J. (2000). Using multisource feedback in the classroom: a computer-based approach. *IEEE Transactions on Education, 43,* 120-124.

Meltzer, L. & Reid, D. K. (1994). New directions in the assessment of students with special needs: the shift toward a constructivist perspective. *The journal of special education, 28,* 338-355.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3 ed., pp. 13-103). New York: American Council on Education/ Macmillan.

Miller, C. M. L. & Parlett, M. (1974). *Up to the mark: A study of the examination game*. London: SRHE.

Miller, P. (2003). The Effect of Scoring Criteria Specificity on Peer and Self-assessment. *Assessment and Evaluation in Higher Education, 28,* 383-394.

Mory, E. H. (2003). Feedback research revisited. In D. H. Jonassen (Ed.), *Handbook of Research for Educational Communications and Technology* (pp. 745-783). New York: Macmillan.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for Performance Assessment. *Review of Educational Research, 62,* 229-258.

Mouton, J. S., Blake, R. R., & Fruchter, B. (1955b). The validity of sociometric responses. *Sociometry, 18,* 181-206.

Mouton, J. S., Blake, R. R., & Fruchter, B. (1955a). The reliability of sociometric measures. *Sociometry, 18,* 7-48.

Narciss, S. (1999). Motivational effects of the informativeness of feedback. In *Annual Meeting of the American Educational Research Association* Montreal.

Neuman, W. (1994). *Social research methods: Qualitative and quantitative approaches*. Boston: Simon & Schuster.

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

Nicol, D. J. & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31,* 199-218.

Nilson, L. B. (2003). Improving student peer feedback. *College Teaching, 51,* 34-38.

Norcini, J. J. (2003). Peer assessment of competence. *Medical Education, 37,* 539-543.

Oldfield, K. A. & MacAlpine, J. M. K. (1995). Peer and self-assessment at tertiary level - an experiential report. *Assessment & Evaluation in Higher Education, 20,* 125-131.

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

Orsmond, P., Merry, S., & Reiling, K. (2002). The use of exemplars and formative feedback when using student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education, 27,* 309-323.

Orsmond, P., Merry, S., & Reiling, K. (1996). The importance of marking criteria in the use of peer assessment. *Assessment & Evaluation in Higher Education, 21,* 239-250.

Pâquet, M. R. & Des Marchais, J. E. (1998). Students' acceptance of peer assessment. *Education for Health: Change in Training and Practice, 11,* 25-35.

Patton, M. Q. (1987). *How to use qualitative methods in evaluation*. Newbury Park: Sage.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know. The science and design of educational assessment*. Washington, DC: National Academy Press.

Pope, N. (2001). An examination of the use of peer rating for formative assessment in the context of the theory of consumption values. *Assessment & Evaluation in Higher Education, 26,* 235-246.

Pope, N. K. L. (2005). The impact of stress in self- and peer assessment. *Assessment & Evaluation in Higher Education, 30,* 51-63.

Popham, W. J. (1997). Consequential validity: Right concern, wrong concept. *Educational Measurement: Issues and Practice, 16,* 9-13.

Prins, F., Sluijsmans, D., & Kirschner, P. A. (2006). Feedback for general practitioners in training: Quality, styles, and preferences. *Advances in Health Sciences Education, 11,* 289-303.

Prins, F., Sluijsmans, D. M. A., Kirschner, P. A., & Strijbos, J. W. (2005). Formative peer assessment in a CSCL environment: a case study. *Assessment & Evaluation in Higher Education, 30,* 417-444.

Pryor, J. & Lubisi, C. (2002). Reconceptualising educational assessment in South Africa -- testing times for teachers. *International Journal of Educational Development, 22,* 673-686.

Purchase, H. C. (2000). Learning about interface design through peer assessment. *Assessment & Evaluation in Higher Education, 25,* 341-352.

Rada, R. & Hu, K. (2002). Patterns in student-student commenting. *IEEE Transactions on Education, 45,* 262-267.

Ramsden, P. (1992). *Learning to teach in higher education*. London: Routledge.

Robinson, J. M. (2002). In Search of Fairness: an application of multi-reviewer anonymous peer review in a large class. *Journal of Further and Higher Education, 26,* 183-192.

Rust, C., Price, M., & O'Donovan, B. (2003). Improving students' learning by developing their understanding of assessment criteria and processes. *Assessment & Evaluation in Higher Education, 28,* 147-164.

Ryan, R. M., Connell, J. P., & Deci, E. L. (1985). A motivational analysis of self-determination and self-regulation in education. In C. Ames & R. Ames (Eds.), *Research on motivation in education: Vol2. The Classroom milieu* (New York: Academic Press.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in education, 5,* 77-84.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18,* 119-144.

Sadler, P. & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11,* 1-31.

Saito, H. & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research, 8,* 31-54.

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

Sambell, K. & McDowell, L. (1997). 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: 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.

SAS Institute Inc. (2004). *SAS/STAT 9.1 User's Guide*. Cary: SAS Institute Inc.

Schelfhout, W., Dochy, F., & Janssens, S. (2004). The use of self, peer and teacher assessment as a feedback system in a learning environment aimed at fostering skills of cooperation in an entrepreneurial context. *Assessment & Evaluation in Higher Education, 29,* 177-201.

Schelfhout, W., Dochy, F., Janssens, S., Struyven, K., & Gielen, S. (2006). Towards an equilibrium model for creating powerful learning environments. Validation of a questionnaire on creating powerful learning environments during teacher training internships. *European Journal of Teacher Education, 29,* 471-505.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education, 35,* 453-472.

Searby, M. & Ewers, T. (1997). An evaluation of the use of peer assessment in higher education: A case study in the School of Music, Kingston University. *Assessment & Evaluation in Higher Education, 22,* 371-383.

Segers, M., Dochy, F., & Cascallar, E. (2003). *Optimizing new modes of assessment: In search of qualities and standards*. Dordrecht: Kluwer Academic.

Segers, M. & Dochy, F. (2001). New assessment forms in problem-based learning: The value-added of the students' perspective. *Studies in Higher Education, 26,* 327-343.

Sitthiworachart, J. & Joy, M. (2003). Deepening computer programming skills by using web-based peer assessment. In *Proceedings of the 4th Annual Conference of the LTSN Centre for Information and Computer Sciences.* NUI Galway (Ireland): LTSN-ICS.

Sivan, A. (2000). The implementation of peer assessment: An action research approach. *Assessment in Education: Principles, Policy & Practice, 7,* 193-213.

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

Sluijsmans, D. (2002). *Student involvement in assessment. The training of peer assessment skills.* Unpublished doctoral dissertation. Open Universiteit Nederland, Heerlen.

Sluijsmans, D., Brand-Gruwel, S., van Merriënboer, J. J. G., & Bastiaens, T. (2002). The training of peer assessment skills to promote the development of reflection skills in teacher education. *Studies in Educational Evaluation, 29,* 23-42.

Sluijsmans, D., Dochy, F., & Moerkerke, G. (1998). Creating a learning environment by using self-, peer- and co-assessment. *Learning environments research, 1,* 293-319.

Sluijsmans, D. & Prins, F. (2006). A conceptual framework for integrating peer assessment in teacher education. *Studies in Educational Evaluation, 32,* 6-22.

Sluijsmans, D. M. A., Brand-Gruwel, S., & van Merriënboer, J. J. G. (2002). Peer assessment training in teacher education: effects on performance and perceptions. *Assessment & Evaluation in Higher Education, 27,* 443-454.

Sluijsmans, D. M. A., Brand-Gruwel, S., van Merriënboer, J. J. G., & Martens, R. L. (2004). Training teachers in peer-assessment skills: effects on performance and perceptions. *Innovations in Education & Teaching International, 41,* 60-78.

Smith, H., Cooper, A., & Lancaster, L. (2002). Improving the quality of undergraduate peer assessment: A case for student and staff development. *Innovations in Education & Teaching International, 39,* 71-81.

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

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

Starren, H. (1998). De toets als hefboom voor meer en beter leren. *Academia,* 26.

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

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

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.

Struyf, E., Vandenberghe, R., & Lens, W. (2001). The evaluation practice of teachers as a learning opportunity for students. *Studies in Educational Evaluation, 27,* 215-238.

Taras, M. (2002). Using assessment for learning and learning from assessment. *Assessment and Evaluation in Higher Education, 27,* 501-510.

Thomas, P. R. & Bain, J. D. (1984). Contextual dependence of learning approaches: The effects of assessment. *Human Learning, 3,* 227-240.

Thomson, K. & Falchikov, N. (1998). 'Full on until the sun comes out': The effects of assessment on student approaches to studying. *Assessment & Evaluation in Higher Education, 23,* 379.

Topping, K. J. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy, & E. Cascallar (Eds.), *Optimizing new modes of assessment: In search of qualities and standards* (pp. 55-87). Dordrecht: Kluwer Academic.

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment & Evaluation in Higher Education, 25,* 149-169.

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

Topping, K. J. (2005). Trends in peer learning. *Educational Psychology, 25,* 631-645.

Trahasch, S. (2004). Towards a flexible peer assessment system. In *Proceedings of the Fifth International Conference on Information Technology Based Higher Education and Training, 2004.* (pp. 516-520).

Tsui, A. B. M. & Ng, M. (2000). Do secondary L2 writers benefit from peer comments? *Journal of Second Language Writing, 9,* 147-170.

Tynjälä, P. (1997). Developing education students' conceptions of the learning process in different learning environments. *Learning and Instruction, 7,* 277-292.

van den Berg, I. (2003). *Peer assessment in universitair onderwijs.* Unpublished doctoral dissertation. Universiteit Utrecht, Utrecht.

van den Berg, I., Admiraal, W., & Pilot, A. (2006b). Peer assessment in university teaching: Evaluating seven course designs. *Assessment & Evaluation in Higher Education, 31,* 19-36.

van den Berg, I., Admiraal, W., & Pilot, A. (2006a). Designing student peer assessment in higher education: analysis of written and oral peer feedback. *Teaching in Higher Education, 11,* 135-147.

Venables, A. & Summit, R. (2003). Enhancing scientific essay writing using peer assessment. *Innovations in Education & Teaching International, 40,* 281-290.

Weaver, M. R. (2006). Do students value feedback? Student perceptions of tutors' written responses. *Assessment & Evaluation in Higher Education, 31,* 379-394.

Wiliam, D. (2006). Does assessment hinder learning? In *Speech delivered at the ETS Europe breakfast salon (11th July 2006).*

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17,* 31-74.

Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15,* 179-200.

Yorke, M. (2003). Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher Education, 45,* 477-501.

Zhang, S. (1995). Reexamining the affective advantage of peer feedback in the ESL writing class. *Journal of Second Language Writing, 4,* 209-222.

Design of the cover: Karmen Buvens