Assessment in constructivist learning environments

ISBN 90-5278-549-X
ISBN 978-90-5278-549-3

Druk: Datawyse / Universitaire Pers Maastricht
Omslag: Daan & Gijs

Assessment in constructivist learning environments

Studies about perceptions and assessment in a constructivist learning environment in relation to students' study outcomes

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit Maastricht,
op gezag van de Rector Magnificus,
Prof. mr. G.P.M.F. Mols
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen
op donderdag 7 september 2006 om 14.00 uur

door

Gerard van de Watering

Promotor

Prof. dr. F.J.R.C. Dochy

Beoordelingscommissie

Prof. dr. P.L.H. Van den Bossche LL.M. (voorzitter)
Prof. dr. H. Baert (Katholieke Universiteit Leuven)
Prof. dr. W. Gijselaers
Prof. dr. T. Spronken

Content

DANKWOORD

Ik ben de vele mensen om me heen dankbaar voor de hulp die ik heb mogen ontvangen bij de totstandkoming van dit proefschrift. Op de eerste plaats ben ik Karien zeer dankbaar. Het proefschrift heeft niet alleen van mij veel inzet gevraagd. Ik werd al die tijd door jou uit liefde gesteund. Je hebt veel geduld met me gehad. Ik heb je liefde en opoffering wel gevoeld maar niet altijd onder woorden kunnen brengen wat dat voor mij betekende en nog betekent. Dankzij jou heb ik veel tijd en energie kunnen steken in het vervolmaken van het proefschrift. Ook jij bent op de proef gesteld. Ik prijs me gelukkig met jou en met onze kinderen. Want ook voor Daan en Gijs was ik niet altijd grijpbaar.

Zeer veel dank ben ik verschuldigd aan mijn promotor Filip Dochy. Het was me een waar genoegen om onder je supervisie mijn eerste stappen te zetten in de onderzoekswereld. De wijze waarop je mij begeleid hebt kan ik het best omschrijven met het woord plezierig. Ik ben er trots op om deelgemaakt te hebben uit jouw team van onderzoekers.

Speciale dank voor David Gijbels, Janine van der Rijt en Piet Van de Bossche voor de vele uren die we samen hebben doorgebracht aan het uitwerken van onze onderzoeken. Ze waren motiverend en inspirerend.

Met plezier denk ik ook terug aan alle collega's van Edit (Educational Innovation and Information Technology) van de Faculteit der Rechtsgeleerdheid van de Universiteit van Maastricht en de goede sfeer waarin we hebben geleefd. Niet te vergeten alle docenten en onderwijsondersteuners waarmee ik de afgelopen tijd heb samengewerkt. Vele van jullie maken van het onderwijs een prachtig vakgebied. De probleemstellingen en de discussies zijn een bron van inspiratie geweest. Een aantal van die probleemstellingen is terug te vinden in het proefschrift.

Als laatste wil ik speciaal mijn ouders bedanken. Jullie hebben me altijd gesteund. Wat kan een mens zich nog meer wensen?

Ik ben vooral trots op mezelf. Trots op het hetgeen ik gepresteerd heb.

Introduction

When I started my professional career in 1995, a lot of attention in the field of educational development was being paid to independent learning environments. In these environments, students were supposed to have a more active role in the classroom, to have more responsibility for their own learning processes and more insights into their own metacognitive skills. Although the positive effects for the learner of such elements of autonomous or independent learning were already well known and described, this concept came into prominence in the early 1990s because of the notion that we now live in a knowledge society, so that knowledge learned at school will date quickly. For this reason, life long learning is needed for graduates to keep up with ongoing developments in society. A decade later various concepts, such as student centred learning environments, new learning environments and powerful learning environments, with a shift from cognitive psychology to constructivist psychology (Cooper 1993, Martin, 1997), are used interchangeably. Nowadays the concept 'constructivist learning environment' is still used to point out the importance of learning environments in which the students' learning, defined as an active knowledge building process, is the core issue (Tynjälä, 1999). Consequently, the role of the teacher has become a more facilitative one (Jonassen, Peck & Wilson, 1999).

Constructivism and problem-based learning

There are, of course, different viewpoints about constructivism and the debates regarding its operationalisation are still going on (Gijbels, 2005). To some, for example Wilson (1997), it is not so obvious what a constructivist learning environment looks like, because one should ask in each case how construction of meaning can be facilitated in the best way. According to his viewpoint, teaching may include all kind of activities, including drill and practice, a lecture or prepared assignments (Tynjälä, 1999; Wilson & Lowry, 2001). But a constructivist learning environment should always be a 'place where learners may work together, draw upon (information) resources using a variety of tools, supporting each other in their guided pursuit of learning goals and problem-solving activities' (Wilson, 1995). According to Lin et al. (1996) a constructivist learning environment should be a learning community in which students have the opportunity to plan, organize,

monitor, and revise their own research and problem solving in a collaborative way. To others, for example Savery and Duffy (1996), it is more obvious that constructivism implies specific learning activities or instructional principles, such as: (1) anchoring of all learning in large tasks or problems; (2) activating the learners; (3) challenging and supporting the learner's thinking; (4) authentic tasks or problems; (5) reflecting the complexity of the real world; (6) the learners' ownership of the problem solving process; and (7) the opportunity to reflect on the content and the learning process, as it is embedded in problem-, case- or project-based learning (Brown & King, 2000; Dekeyser & Baert, 1999). In fact, according to Savery and Duffy (2001), problem-based learning is one of the best exemplars of a constructivist learning environment.

Nowadays, problem solving is seen as one of the key skills in education. Other examples of key skills include working with others and reflecting on one's own work in order to improve learning and performance (Savin-Baden, 2000). According to Dunlap and Grabinger (1996), problem solving skills are essential for today's workforce. Modern societies are complex and fast developing, competitive, and multicultural, with shifting values and a profusion of information provided by media and information technology. To live and work in such a society, one should learn to "*think critically and to analyze and synthesize information in order to solve technical, social, economic, political and scientific problems*" (Dunlap & Grabinger; 1996, p. 65). Compared to the past, in today's society there are more problems and more choices, and individuals are held more responsible for their choices. Where there was only one correct answer in the past, nowadays countless answers are possible. For students this means that they have to learn to be able to work with others constructively in complex situations to solve problems. Problem-based learning "*can help students to learn with complexity to see that there are no straightforward answers to problem scenarios, but that learning and life take place in contexts, contexts which affect the kinds of solutions that are available and possible*" (Savin-Baden; 2000, p. 5).

Problem-based learning has its origins in the 1950s in Canadian medical schools because of dissatisfaction with the medical learning environment (Barrows, 1996). Since then it has evolved and is implemented on different levels of education in different ways in many disciplines throughout the world for a variety of reasons (Gijselaers, 1995). Despite the fact that many variations of problem-based learning are applied, a basic definition of problem-based learning can be given by means of its core characteristics (Barrows, 1996; Dochy, Segers, Van den Bossche & Gijbels, 2003). Problem-based learning has six core characteristics which are in line with the previous described constructivist instructional principles: (1) learning is student centred; (2) learning occurs in small groups; (3) a tutor guides or facilitate the learning processes; (4) problems are used as stimuli for learning; (5) authentic problems are used; and (6) new information is acquired through self directed learning. The advantages of problem-based learning are critically described in a meta analysis, from the angle of assessment, by Gijbels, Dochy, Van den Bossche and Segers (2005). According to this analysis, students in problem-based learning

environments can perform better on different levels of assessment (the analysis distinguishes three levels of assessment: the understanding of concepts; understanding of principles; and application) when compared to students in more traditional learning environments. For an optimum effect, the learning environment should pay attention to the assessment of the application of knowledge. The students also have to adopt an active learning attitude towards this problem-based learning environment (Moust, Bouhuijs & Schmidt; 2001). But it is also dependent on whether the teacher (the tutor) is able to activate the students in a suitable way.

The empirical studies in this dissertation took place in a learning environment with the above mentioned characteristics. Research into problem-based learning is important in two ways. The first is the argument that the potential of problem-based learning is not fully realized (see, for example, Savin-Baden, 2000). Secondly, there is a growing interest in this learning environment, especially in secondary and higher (vocational) education because of the introduction of competencies in these fields. As a consequence, research into exactly how problem-based learning works, and how the learning of students can be activated, is a necessity for both secondary and higher education to be able to implement competence based learning.

Problem-based learning and assessment

It is generally believed that assessment has an important impact on instruction and learning (Gibbs, 1999; Scouller, 1998). In fact assessment can be a very powerful means to focus students on their learning. For example, research has shown that thinking about the consequences of the assessment system in advance is important to the understanding of student behaviour (Boud, 1990). In many cases the assessment system encourages the reproduction of knowledge and surface learning strategies (Dochy, 2005; Tynjälä, 1999). Even if the learning environment promotes deep level learning by means of tasks in which students have to solve problems and apply knowledge in new situations, if the assessment is nothing more than the reproduction of facts then students will, in the end, use surface learning strategies. With the term constructivist alignment, Biggs (1996, 2001, 2003) focuses on the importance of aligning the curriculum objectives, the teaching method or learning environment, and the assessment, to enhance appropriate learning activities. The fit between the learning environment's objectives and the assessment is considered to be a 'magic bullet' in improving learning (Cohen, 1987). For problem-based learning environments, this also means that the main purpose is to make the assessment congruent with the instruction and to align the assessment to what students should be learning: if solving problems is the aim of the learning environment, then students have to learn how to acquire the knowledge and skills necessary to solve problems. The way in which students are capable of solving problems should also be an important part of the assessment (Biggs, 2001). Therefore, nowadays the six core characteristics of problem-based learning described by Barrows (1996) are completed by adding a seventh characteristic: the assessment requires the assessees to apply their knowledge in authentic (professional) and important problem-solving situations (see, for example, Segers, Dochy & De Corte, 1999).

In general, traditional teacher-made assessments do not always measure the intended learning outcomes (Race, 1999) and the consequence for students is that they often focus on memorizing the study material by adopting a surface approach to learning (Biggs, 1996; Entwistle & Entwistle, 1991). Assessment should be a learning experience, encouraging students to use higher order thinking skills. During the assessment, students ought to discover relationships between ideas that they had not previously come across while studying (Nevo, 1995). In most cases, however, assessment is regarded by students as something you have to pass, which does not motivate them towards further learning. Bennet (2000) illustrates this by comparing the behaviour of sea lions, performing tricks for the reward of fish, to the behaviour of students, working to gain marks. The assessments in this dissertation are a combination of two traditional assessment formats: essay or open ended questions and multiple-choice questions. As valid assessments should do, all the assessments try to cover the full spectrum of the cognitive learning objectives (in terms of Bloom's taxonomy (Bloom, 1956), knowledge and higher order skills) by means of: reproduction based questions (assessing the students' ability to recall information); comprehension based questions (assessing the students' understanding of basic concepts and principles); and application based questions (assessing the students' ability to use learned material in new and concrete situations). Although this assessment format works particularly well with problem-based learning environments (Driessen, van der Vleuten & van Berkel, 1999), improvements can be made to optimise the fit between assessments and learning environments. Instead of changing the whole assessment system, such as implementing portfolio and performance assessment, small changes in assessment can be made. Gibbs and Simpson (2003) presented a framework on strategic changes in assessment, to bridge the gap between assessment and learning and instruction. In this framework, the focus lies on feedback. A progressive step in the assessment system is the opportunity for students to receive feedback on their work by the integration of assessment tasks into the learning process. Feedback is considered to be an important aspect to enhance learning. It also can help students to internalise the standards and notion of quality (Gibbs, 1999). In this way, the teaching staff can make sure students will have an accurate picture of their learning progress and will be more likely to achieve their full potential.

Assessment and perception: Beauty lies in the eye of the beholder

The truth is not always experienced in the same way by all people. Different people will perceive the truth differently. We earlier stated that if the assessment is nothing more than the reproduction of facts, students will, in the end, use reproducing learning strategies. It would have been more accurate to have stated that if the students perceive the assessment as nothing more than the reproduction of facts, students will, in the end, use reproducing learning strategies. Furthermore, the question is not only how students perceive the assessment, but also how students perceive their learning environment or their learning strategies. Research into students' perceptions of a learning environment reveals its impact on the way students cope with that learning environment and how they cope with the learning

outcomes (Segers & Dochy, 2001). The effects of implementation of constructivist learning environments do not always demonstrate the expected outcomes (Birenbaum, 2000). On the one hand this can be because the assessment system and the learning environment are not aligned. Even when there is a fit between the assessment and the learning environment, however, there can be disappointing results in terms of educational change because the students' perceptions did not change accordingly (Lawness & Richardson, 2002). It seems that previous learning experiences have an important influence on students' learning strategies and influence their learning outcomes (Segers, 1996). Perceptions of the appropriateness of the workload, and the clarity of the goals in the learning environment and assessment system, also influence students' study approaches and learning strategies (Nijhuis, Segers & Gijselaers, 2005).

From discussion with students and the teaching staff, and also by means of observation of the educational processes and assessment construction processes, I am convinced that students' perceptions of the assessment, or the consequences of that assessment, play an important role in how they deal with it and sometimes how they cope with it. Personal experiences of assessment related issues do not always help to create an appropriate perception of the assessment, especially when these issues are not put into the correct context. For example, a lot of students and teachers perceive multiple-choice exams only as a way to assess reproductive knowledge. Multiple-choice questions can, indeed, be a very good means to assess the students' ability to recall information. For students, it is also very possible to achieve high grades on these kind of assessments by using a surface approach to learning. And, alas, this is also the case most of the time in practice: multiple-choice exams consist to a large extent of reproduction based questions. Thus, from students' own experiences, multiple-choice exams can be considered as ways to assess reproductive knowledge. These students may have problems, however, when an assessment consisting of multiple-choice questions is announced, but it then asks them for something else than their ability to recall information. This is not only because the questions are asking something from them other than what they expected, but also because the students' expectations have led them to use inappropriate learning strategies. From the teachers' point of view, open ended questions and essay questions are often seen as the only possible means to assess their subject in a suitable way. In most cases this perception originated from the experiences these teachers had as students. These assessment formats were the only ones used in their study, as a student they were able to pass these assessments successfully, and they were satisfied with the results. With the help of open ended or essay questions it is, indeed, very possible to assess students' higher order thinking skills. And for teachers it is also very possible to promote a deep approach to learning. But in practice, the open ended questions in teacher-made assessments do not always assess higher order thinking skills as intended by their constructors, and assess nothing more than the reproduction of taught skills by means of the tasks undertaken during contact hours. Teachers must also have the opportunity to reflect on the assessment contents, the assessment outcomes, and the aims of the learning environment, to get a clear perception of their assessments. From a study by MacLellan (2001), it is clear that there are differences between students'

perceptions of the assessment environment and what is intended by the teaching staff.

Studies by Scouller and Prosser (1994) and Scouller (1998) showed that a student's perception of assessment is one of the variables influencing performance, directly or indirectly. These studies examined the students' perceptions of assessment in relation to their learning strategies, and to their performances, and found that poorer performances were related to the use of unsuitable study approaches due to incorrect perceptions of the assessment. Better performances were positively related to correct perceptions and the use of suitable study approaches.

Structure of the dissertation and the research questions

As the title of this dissertation has already indicated, its main scope focuses on students' and teachers' perceptions of assessment and the relationship between these perceptions and learning outcomes. The studies took place in a constructivist learning environment, as designed into problem-based learning environment at the faculty of law of the University of Maastricht. This dissertation consists of seven chapters. Firstly, an introduction to the different studies in the dissertation is given. This gives an overview of the studies, describes the structure of the studies in the dissertation and presents the research questions of the different studies.

The first study is in chapter 1 and it concerns students' perceptions of their learning environments. Students in a problem-based learning environment and in a conventional lecture-based environment were asked, using a questionnaire consisting of seven key factors of constructivist learning environments, about their experiences of the educational practice. Learning environments based on constructivism have the potential to improve the educational outcomes for students in higher education. Moreover, constructivism is the underlying theory referred to when the beneficial effects of problem-based learning are postulated. We assumed that the extent to which students perceive constructivist principles to be present in the learning environment will be related to the expected effects of the learning environment. The main aim of the first study is, therefore, to verify whether students in a problem-based learning environment perceive the learning environment to be more constructivist, when compared to the perceptions students have of a conventional lecture-based environment. A question of particular interest in this study is for which factors the differences between the problem-based learning environment, and the conventional lecture-based environment, are the largest.

In chapter 2, students' approaches to learning in a problem-based learning environment, and the students' learning outcomes, were the subject of study. In general, the use of deep learning approaches are associated with higher quality learning outcomes and a surface approach with lower learning outcomes, but it is also suggested that the assessment system plays a major role in this relationship. Deep approaches to learning can also be regarded as an outcome of a high quality learning environment. A constructivist learning environment, if perceived as such,

should promote appropriate approaches to learning and therefore students should have adopted a deep approach to learning in the problem-based learning environment under study. In this study, the students' approaches to learning were measured with the revised two factor study process questionnaire (R-SPQ-2F). In order to distinguish between students' assessment outcomes on different levels of the knowledge structure, Sugrue's taxonomy (a model of cognitive components of problem-solving) was used. With this taxonomy, each question in the final exam was categorized as 'understanding of concepts', 'understanding of the principles that link concepts', or as 'linking of concepts and principles to application'. The purpose of the study was to explore further the relationship between students' approaches to learning and their quantitative learning outcomes, as measured by the assessment, from the perspective of the different components of problem-solving, using Sugrue's taxonomy.

The third study, in chapter 3, has two purposes. The first is to gain more insight into students' actual assessment preferences and perceptions of assessment. The second is to explore the effects of these preferences and perceptions on the students' assessment outcomes. Constructivist learning environments such as problem-based learning environments are, as in this study, not always accompanied by new methods of assessment to align learning more closely with assessment. In this study the assessment combined two assessment formats and the question types were directed towards different cognitive process levels. Parts of the Assessment Preferences Inventory (API) were used to answer the four research questions in this study. From different studies regarding assessment preferences, it seems that students prefer assessment formats which reduce stress and anxiety and it is assumed that students will perform better on their preferred assessment formats. It also seems that the students' perceptions of assessment are not always correct, which can lead to poor performance. Four research questions were formulated to investigate these assumptions: Firstly, which assessment preferences do students have? Secondly, how did students perceive the assessment? Thirdly, in what ways are students' assessment preferences related to their assessment results? And fourthly, in what way are students' perceptions of the assessment related to their assessment results?

Chapter 4 elaborates further on the students' and teachers' perceptions of assessments. This chapter has three parts. The first part is a theoretical introduction. The second part is a review of research into teachers' and students' perceptions of item difficulty. The third part is an empirical study of the ability of students and teachers to estimate item difficulty correctly. Research on item difficulty and the perceptions of teachers and students of item difficulty is relevant because little is known about the degree to which assessments in higher education are correctly targeted at the students' levels of competence. The central question to be answered is whether teachers and students perceive item difficulty correctly. In the empirical study four research questions were formulated: firstly, to what extent are teachers accurate in estimating the difficulty level of assessment items? Secondly, to what extent do students' perceptions of the difficulty level of the assessment items correspond with their actual difficulty levels? Thirdly, to what extent do the students' perceptions of the difficulty levels of the assessment items differ from the

teachers' estimations of the item difficulties? And finally, what relationship exists between students' perceptions of the difficulty levels of the assessment items and their performances on the assessment?

In chapter 5 teachers observed their students during a course in a problem-based learning setting and classified these students at the end of the course into four distinct groups, perceived as barely competent, moderately competent, highly competent, and students with test anxiety. The purpose of this study was to investigate to what extent different parts of the assessment, multiple-choice questions and essay questions, discriminate between less and more competent students, as based on the teachers' perceptions. It was expected that the multiple-choice questions would be equally less discriminating between all the perceived groups of students than the essay questions. The validity of the observations in the course was also questioned in this study: were the results course specific or could the same findings also be found in previous courses?

The purpose of the study in chapter 6 was to gain more insight into the effects of the implementation of written assessment-tasks in a problem-based learning environment through quantitative and qualitative data. The implementation of assessment-tasks is considered to be a necessary move, in line with constructivist learning theories, to improve learning and to align assessment more with the learning environment. With assessment-tasks cognitive processes, such as analysing, integrating, synthesising, evaluating and problem solving, can be assessed more extensively than with, for example, just a final exam. In this study students were stimulated to produce qualitative learning activities by means of the assessment-tasks and received a 'bonus point' for the final exam if the assessment-tasks were shown to be of sufficient quality and effort. The influence of the implementation of the assessment-tasks on students' performances and on students' and teachers' perceptions were investigated by means of two research questions: firstly, do students who undertake the assessment-tasks do better in their final exam compared to students who do not? Secondly, what are the most important concerns in students' and teachers' perceptions of the assessment-tasks?

In the last chapter, a summary of the conclusions and discussions of the studies in this dissertation will be given, as well as recommendations for educational practice and future research.

References

Barrows, H.S. 1996. Problem-based learning in medicine and beyond. In L. Wilkerson & W.H. Gijselaers (Eds.), *Bringing problem-based learning to higher education: Theory and practice.* New directions for teaching and learning, No. 68 (pp. 3-13). San Francisco, CA: Jossey-Bass.

Bennett, M. (2000). Assessment to promote learning. *The Law Teacher*, *34* (2), 167-174.

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

Biggs, J. (2001). The reflective institution: Assuring and enhancing the quality of teaching and learning. *Higher Education*, *41*, 221-238.

Biggs, J. (2003). *Teaching for Quality Learning at University* (2nd ed.). Buckingham: SRHE and Open University Press.

Birenbaum, M. (2000, September). *New Insights into Learning and Teaching and the Implications for Assessment.* Keynote address at the 2000 conference of the European Association of Research on Learning and Instruction Special Interest Group on Assessment and Evaluation, Maastricht, The Netherlands.

Bloom, B.S. (Ed.). (1956). *Taxonomy of educational objectives. Handbook I: The cognitive domain.* New York: McKay.

Boud, D. (1990). Assessment and the promotion of academic values. *Studies in Higher Education*, *15* (1), 101-111.

Brown, S.C., & King, F.B. (2000). Constructivist Pedagogy and How We Learn: Educational Psychology Meets International Studies. *International Studies Perspectives*, *1*, 245-254.

Cohen, S.A. (1987) Instructional alignment: searching for a magic bullet, *Educational Researcher*, *16* (8), 16-20.

Cooper, P.A. (1993). Paradigm Shifts in Designed Instruction: From Behaviorism to Cognitivism to Constructivism. *Educational Technology*, *33* (5), pp. 12-18.

Dekeyser, L., & Baert, H. (1999). Projectonderwijs: leren en werken in groep. Leuven: Acco.

Dochy, F. (2005, August). *Learning lasting for life and assessment: How far did we progress?* Presidential address at the 20th conference of the European Association of Research on Learning and Instruction, Nicosia, Cyprus. Retrieved November 11, 2005, from http://www.earli.org/conferences/ previuous_conferences/earli_2005/ presidential_adress

Dochy, F., Segers, M., Van den Bossche, P., & Gijbels, D. (2003) Effects of problem-based learning: A Meta-analysis. *Learning and instruction*, *5* (13), 533-568.

Driessen, E., van der Vleuten, C., & van Berkel, H. (1999) Beyond the multiple-choice v. essay questions controversy: combining the best of both worlds, *The Law Teacher*, *33* (2), 159-171.

Dunlap, J. C., & Grabinger, R. S. (1996). Rich Environments for Active Learning in the higher education classroom. In B. G. Wilson (Ed.), *Constructivist learning environments: Case studies in instructional design* (pp. 65-82). Englewood Cliffs NJ: Educational Technology Publications.

Entwistle, N., & Entwistle, A. (1991). Contrasting forms of understanding for degree examinations: the student experience and its implications. *Higher Education*, *22*, 205–277.

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 and Open University Press.

Gibbs, G., & Simpson, C. (2003). Does your assessment support your students' learning? *Learning and Teaching in Higher Education*, *1* (1), Retrieved December 5, 2005, from http://www.open.ac.uk/science/fdtl/documents/

Gijbels, D. (2005). *Effects of new learning environments: Taking students' perceptions, approaches to learning and assessment into account* [Dissertation]. Maastricht: Maastricht University.

Gijbels, D. Dochy, F., Van den Bossche, P., & Segers, M. (2005). Effects of problem-based learning: A meta-analysis from the angle of assessment. *Review of Educational Research*, *75* (1), 27-61.

Gijselaers, W. (1995). Perspectives on problem-based learning. In W. Gijselaers, D. Tempelaar, P. Keizer, J. Blommaert, E. Bernard & H. Kasper (Eds.), *Educational innovation in economics and business administration: The case of problem-based learning* (pp. 39-52). Boston, MA: Kluwer Academic Publishers

Jonassen, D., Peck, K., & Wilson, B. (1999). *Learning with technology: A constructivist perspective*. Upper Saddle River, NJ: Prentice Hall.

Lawness, C.J., & Richardson, J.T.E. (2002). Approaches to studying and perceptions of academic quality in distance education. *Higher Education*, *44*, 257-282.

Lin, X., Bransford, J., Hmelo, C., Kantor, R., Hickey, D., Secules, T., Petrosino, A., Goldman, S., & The Cognition and Technology Group at Vanderbilt (1996). Instructional design and development of learning communities: An invitation to a dialogue. In B. Wilson (Ed.), *Constructivist learning environments*. Englewood Cliffs, NJ: Educational Technology Publications.

Martin, R. (1997). *Constructivism and Transformative Learning Theories*. Retrieved December 2, 2005, from http://www.inspiredinside.com/learning/library/index.htm

MacLellan, E. (2001). Assessment for learning: the differing perceptions of tutors and students. *Assessment & Evaluation in Higher Education*, *26*, 307-318.

Moust, J.H.C., Bouhuijs, P.A.J., & Schmidt, H.G. (2001). *Problem-based learning: A student guide*. Groningen: Wolters-Noordhoff.

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

Nijhuis, J.F.H., Segers, M.S.R., & Gijselaers, W.H. (2005). Influence of redesigning a learning environment on student perceptions and learning strategies. *Learning Environment Research*, *8* (1), 67-93.

Race, P. (1999). Why Assess Innovatively? In S. Brown & A. Glasner (Eds.), *Assessment Matters in Higher Education* (pp. 57-70). Buckingham: SHRE and Open University Press.

Savery, J.R., & Duffy, T.M. (1996). Problem-based learning: An instructional model and its constructivist framework. In B. Wilson (Ed.), *Constructivist Learning Environments: Case Studies in Instructional Design* (pp. 135-148). Englewood Cliffs, NJ: Educational Technology Publications.

Savery, J.R., & Duffy, T.M. (2001). *Problem Based Learning: An instructional model and its constructivist framework.* Bloomington: CRLT Technical Report No. 16-01. Retrieved November 30, 2005, from http://crlt.indiana.edu/publications/journals/

Savin-Baden, M. (2000). *Problem-based Learning in Higher Education: Untold Stories.* Buckingham: SRHE and Open University Press.

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

Scouller, K.M., & Prosser, M. (1994). Students' experiences in studying for multiple choice question examinations. *Studies in Higher Education*, *19* (3), 267-279.

Segers, M. (1996). Assessment in a problem-based economics curriculum. In M. Birenbaum & F. Dochy (Eds.), *Alternatives in Assessment of Achievements, Learning Processes and Prior Learning* (pp. 201-226). Boston: Kluwer Academic Press.

Segers, M., & Dochy, F. (2001). New assessment forms in Problem-based Learning: the value-added of the students' perspective. *Studies in Higher Education*, *26* (3), 327-343.

Segers, M., Dochy, F., & De Corte, E. (1999). Assessment practices and students' knowledge profiles in a problem-based curriculum. *Learning Environments Research*, *12* (2), 191-213.

Shuell, T.J. (1993). Toward an integrated theory of teaching and learning. *Educational Psychologist*, *28* (4), 291-311.

Thompson, K. (2001) Constructivist Curriculum Design for Professional Development: A Review of the Literature. *Australian Journal of Adult Learning*, *41* (1), 95-105. Retrieved December 3, 2005 from http://hrast.pef.uni-lj.si/~joze/podiplomci/FRI/ Constructivism.htm

Tynjälä, P. (1999). Towards expert knowledge? A comparison between a constructivist and a traditional learning environment in the university. *International Journal of Educational Research*, *31*, 357-442.

Wilson, B. (Ed.). (1996). *Constructivist Learning Environments: Case Studies in Instructional Design.* Englewood Cliffs, NJ: Educational Technology Publications.

Wilson, B.G. (1995). Metaphors for instruction: Why we talk about learning environments. *Educational Technology*, *35* (5), 25-30. Retrieved December 22, 2005 from http://carbon.cudenver.edu/~bwilson/metaphor.html

Wilson, B.G. (1997). Reflections on Constructivism and Instructional Design. In C. R. Dills & A. J. Romoszowski (Eds.), *Instructional Development Paradigms* (pp. 63-80). Englewood Cliffs NJ: Educational Technology Publications.

Wilson, B., & Lowry, M. (2001). Constructivist learning on the Web. In L. Burge (Ed.), *Learning technologies: Reflective and strategic thinking*. San Francisco, CA: Jossey Bass, New Directions for Adult and Continuing Education. Retrieved December 2, 2005, from http://ceo.cudenver.edu/~brent_Wilson/WebLearning.html

Chapter 1

Constructivist Learning Environments: The students' perspective[1]

Abstract

Research into students' perceptions of their learning environments reveals the impact of these perceptions on the way students cope with these learning environments. Consequently, students' perceptions affect the results of their learning. This study aims to investigate whether students in a problem-based learning environment perceive it to be more constructivist when compared with the perceptions students have of a conventional lecture-based environment. Using a questionnaire consisting of seven key factors of constructivist learning environments, the results show that students in the problem-based learning environment perceive it to be more constructivist when compared to the perceptions of students in a conventional lecture-based environment. The difference was statistically significant for four of the seven factors. According to the effect size, as measured by the *d*-index, the difference in perception between the two groups was greatest for the factor 'conceptual conflicts and dilemmas'.

[1] Based on: Gijbels, D., van de Watering, G., & Dochy, F. (2006). New learning environments and constructivism: The students' perspective. *Instructional Science, 34* (3), 213-226.

Introduction

In recent years, education has frequently been blamed for graduates not being sufficiently able to apply their knowledge to solve complex problems in a working context. The development and implementation of instructional practices that will foster students' skills to communicate, think and reason effectively, make judgements about the accuracy of large volumes of information, solve complex problems and work collaboratively in diverse teams, remains an important challenge for today's higher education (Pellegrino, Chudowsky & Glaser, 2001). Constructivist Learning Environments (CLEs), based on constructivist theory, claim to develop an educational setting in which to reach this goal, making the students' learning the core issue and defining instruction as enhancing learning (Lea, Stephenson & Troy, 2003). The gap between educational practice and the theory of constructivism seems to be difficult to bridge, however (De Corte, 2000). One major problem is that it has been difficult to characterise a constructivist learning environment (Windschitl, 2002). Constructivism can be seen as an umbrella term that groups learning perspectives with the same basic assumption about learning: the understanding that knowledge is actively constructed by the learner (Birenbaum, 2003; Harris & Alexander, 1998; Tynjälä, 1999). In this sense all learning environments are constructivist since, even in teaching situations such as drill and practice, students are constructing knowledge and this is simply because that is the way the mind operates (von Glaserfeld, 1993).

The many discussions between the different theoretical positions of constructivism, all with varying emphases, have inhibited the narrowing of the bridge between theory and practice (Kennedy, 1997). Different perspectives of constructivism emphasise either individual cognitive processes, such as cognitive constructivism which is concerned with the knowledge construction of the individual, or social co-constructions of knowledge, such as social constructivism which stresses the collaborative processes in knowledge building (Windschitl, 2002). Despite many animated discussions, there seems to be no incompatibility amongst the theories and integrative approaches seem to be developing (Resnick, 1994; Vosniadou, 1996; Tynjälä, 1999). Despite ongoing debates, the constitution of the instructional principles of constructivist theory, which guide the nature and quality of educational materials and the learning environment, remains unclear (Harris & Alexander, 1998; Tenenbaum, Naidu, Jegede & Austin, 2001). Both teachers and researchers are in need of more concrete anchors to support their thoughts and actions in applying constructivism to educational practice (Windschitl, 2002).

Key factors of constructivist learning environments

During the past decade, some authors attempted to define the key features of constructivist learning environments and developed questionnaires to evaluate their presence in daily educational practice. Taylor, Fraser and Fisher (1997) developed the new Constructivist Learning Environment Survey (CLES), based on the original

CLES (Taylor & Fraser, 1991), to assess the degree to which students in secondary education perceive a mathematics or science learning environment to be consistent with the key dimensions of constructivism. Small-scale qualitative studies, as well as large-scale quantitative studies, were conducted. In the qualitative studies, the researchers visited classrooms as participant-observers, interviewed students and teachers and analysed documents from the curriculum. In both cases, the focus was on the way the students made sense of responding to CLES items and how the data from the CLES were compatible with the observations. The qualitative part resulted in a 30 item questionnaire, divided into 5 scales of 6 items. The statistical characteristics of this questionnaire were determined in two large-scale quantitative surveys. The final version of their survey consists of 5 scales of 6 items, each to be answered on a 5-point Likert scale. The 5 scales that Taylor and colleagues identified for secondary education are: (1) personal relevance, (2) uncertainty, (3) critical voice, (4) shared control and (5) student negotiation. Taylor et al. (1997, p. 300) argue that, because of the satisfactory internal consistency and factorial validity of the 5 scales, the CLES can be used "*to monitor the development of constructivist learning environments in school science in Western cultures*".

In the field of university teaching, Tenenbaum, Naidu, Jegede and Austin (2001) recently empirically defined and examined key features of constructivist learning environments and their incorporation into two different learning environments (on-campus and distance learning). In the first phase of their study, they carried out a survey using an international electronic mailing list to explore the concept of constructivism, the processes underlying constructivist learning, and its facilitation. In the second phase, they elaborated further on the key features of constructivism in the learning environment and developed a questionnaire using the results of phase 1. A subsidiary aim of this second, quantitative phase was the development of a questionnaire that could be used by other researchers in different educational settings to investigate the presence and/or absence of constructivist practices. The results of the study in both phases were very similar and resulted in a survey containing thirty 5-point Likert scale questions. Seven key factors of constructivist learning environments underlie this questionnaire: (1) arguments, discussions, debates; (2) conceptual conflicts and dilemmas; (3) sharing ideas with others; (4) materials and measures targeted toward solutions; (5) reflections and concept investigation; (6) meeting student needs; and (7) making meaning, real-life examples. Comparison of students' perceptions of the seven factors in different units within the same educational setting revealed that the extent to which the seven factors were experienced differ between various units. Furthermore, comparison between the designers' perceptions and the students' perceptions indicated that the seven factors are not very strongly present in the learning environment from the perceptions of the students, despite the belief of the designers that they had created the learning environments in such a way. The difference was clearest for the factors 'sharing ideas with others' and 'making meaning, real life examples'.

Constructivist learning environments: the case of problem-based learning

Generally, the theory of constructivism is frequently referred to when discussing Constructivist Learning Environments (CLEs). Constructivist learning environments, such as project-based education, case-based learning and problem-based learning are claimed to have the potential to improve the educational outcomes for students in higher education (Lea, Stephenson & Troy, 2003; Simons, van der Linden & Duffy, 2000). Problem-based learning is probably the best known example of a CLE claiming to be highly consistent with constructivist features (Birenbaum, 2003; Hendry, Frommer & Walker, 1999; Russell, Creedy & Davis, 1994; Savery & Duffy, 1995; Segers, Dochy & De Corte, 1999). Although new in some aspects, problem-based learning (known as PBL) is generally based on ideas that originated earlier and have been nurtured by different researchers (Ausubel, Novak & Hanesian, 1978; Bruner, 1959, 1961; Dewey, 1910, 1944; Piaget, 1954; Rogers, 1969). PBL originated in the 1950s and 1960s. Nowadays, PBL is developed and implemented in a wide range of domains. In spite of the many variations of PBL that have evolved, Barrows (1996) describes a core model of PBL in which six fundamental characteristics can be distinguished. The first characteristic is that learning needs to be student-centred. Secondly, learning has to occur in small student groups, under the guidance of a tutor. The third characteristic refers to the tutor as a facilitator or guide. Fourthly, authentic problems are encountered in the learning sequence, before any preparation or study has occurred. Fifthly, the problems encountered are used as a tool to achieve the required knowledge and the problem-solving skills necessary to eventually solve the problem. Finally, new information is acquired through self-directed learning.

Although all CLEs are designed to educate students to analyse and solve problems in an efficient way, empirical studies regarding the effects of such learning environments do not always demonstrate the expected learning outcomes (Segers, 1996). Understanding and improving educational effects demands a 'multi-directional attack' (Goodyear & Hativa, 2002). Research shows that the way the learning environment is perceived by the students, rather than the factual curriculum, affects to a large extent how students cope with the learning environment and, consequently, their learning results (Brekelmans, van den Eeden, Terwel & Wubbels, 1997; Entwistle & Tait, 1990; Fraser, Walberg, Welch & Hattie, 1987; Segers & Dochy, 2001). It follows that educational interventions will be ineffective unless they modify students' perceptions in the intended way. A recent study of students' perceptions of PBL (Dochy, Segers, Van den Bossche & Struyven, 2005) indicated that students perceive the characteristics of the problem-based learning environment, translated into statements, as being present and of high consequence for their learning. If one ponders the implementation of new learning environments, a major question is whether students from new learning environments achieve goals in a more effective way than students who receive more conventional instruction. Conventional instruction methods are those that are marked by large group lectures and instructor-provided learning objectives and assignments (Albanese & Mitchell, 1993). Because constructivism is the underlying theory referred to when superior effects of new learning environments are

postulated, we assume that the extent to which students perceive the constructivist principles in the learning environment as being present will be related to the expected effects of the learning environment.

The main aim of the present study is, therefore, to verify whether students in new learning environments perceive the learning environment to be more constructivist when compared to the perceptions students have of a conventional lecture-based environment. A question of particular interest is for which factors the differences between the constructivist learning environment and the conventional lecture-based environment are the largest. The CLE used in this study is highly consistent with the characteristics of PBL, as will be outlined below.

Method

Participants

The participants in this study were 229 students studying in a problem-based curriculum and 188 students in a lecture-based curriculum. Students studied in two different universities offering bachelor and masters law programs. In both groups, students studied law and were enrolled in a course on the topic of private law (including the history of private law), situated for both universities in second semester of the second year of their undergraduate law studies.

Instrument

The students completed the questionnaire developed by Tenenbaum, Naidu, Jegede & Austin (2001) to obtain a view of students' perceptions of the presence of constructivist practices and principles in their learning environments. The original questionnaire was translated into Dutch by the first author. Four expert educational scientists were given the questionnaire in order to decide if the translation was accurate and phrased clearly enough. To check the latter, the questionnaire was also presented to a small group of students. This resulted in a final translation of the original questionnaire. An example for each factor is presented in Table 1.1. The first factor, 'arguments, discussions, debates', stresses learning as an active and cumulative construction of knowledge. The extent to which students are confronted with conceptual conflicts indicating that knowledge is not certain is captured by the second factor, 'conceptual conflicts and dilemmas'. The third factor, 'sharing ideas with others', deals with learning as a cooperative process. The goal-oriented aspect of learning is covered by the fourth factor, 'materials and measures targeted toward solutions'. The fifth factor, 'motivation toward reflections and concept investigation', asks about the extent to which meta-cognitive aspects of learning are stimulated. The student-centred character of the learning process is stressed in the sixth factor, 'meeting student needs'. Finally, the seventh factor, 'making meaning, real-life examples', deals with the contextual aspect of learning. Confirmatory Factor Analysis (CFA) was used to verify whether the original factor structure could be validated. The value for the Root Mean Square Error of Approximation (*RMSEA* = .07) indicates that the data set fits the 7-factor model fairly well

(sufficient fit values are smaller than .08, Browne & Cudeck, 1993; Guay, Marsh & Boivin, 2003) whereas the χ^2/df value (2.82) exceeds the guideline of $\chi^2/df < 2$ somewhat. The latter was also the case in the original questionnaire (Tenenbaum, et al., 2001).

Table 1.1. Main characteristics and an example for each factor of the questionnaire used in the study

Factor	Main characteristics	Scale example
Arguments, discussions, debates	Learning as an active and cumulative construction of knowledge	The unit allowed for constant exchange of ideas between student and teacher
Conceptual conflicts and dilemmas	Confrontation with conceptual conflicts: knowledge is not certain	The unit caused confusion among conceptual ideas
Sharing ideas with others	Learning is cooperative	The unit allowed social interaction
Materials and resources targeted toward solutions	Learning is goal-oriented	The unit included relevant examples
Motivation toward reflections and concept investigation	Motivating the meta-cognitive aspects of learning	The unit encouraged me to examine several perspectives of an issue
Meeting student's needs	The student-centeredness of the learning environment	The unit took into consideration my needs and concerns
Making meaning, real-life examples	The contextual aspect of learning	The unit was rich in examples

The Cronbach's alpha coefficient of .91 indicated a high overall reliability of the translated questionnaire. The alpha coefficients of the subscales are also all judged to be acceptable for assessing differences between groups (Mehrens & Lehmann, 1991): the arguments, discussions, debates scale: .79; the conceptual conflicts and dilemmas scale: .66; the sharing ideas with others scale: .76; the materials and measures targeted toward solutions scale: .60; the reflections and concept investigation scale: .79; the meeting student needs scale: .74; and the making meaning, real-life examples scale: .62.

Procedure

In both groups, the questionnaires were administered to all students who were present during one of the meetings near the end of their course. Participation was voluntary and confidential. Students were told that their responses would remain anonymous.

Learning environments

The CLE in this study can be seen as a variant of a PBL course, structured as follows. Over 8 weeks, students worked on a topic in the area of private law. During these 8 weeks the students worked twice a week for two hours in small groups (maximum 19 students) on different tasks, guided by a tutor. As well as

these tutorial groups, they were enrolled in somewhat bigger practical classes (38 students) for 2 hours a week and another 4 hours a week (2 sessions of two hours) in large class lectures. Assessment for this course took place by means of a written exam, immediately after the course.

Students in the conventional lecture-based curriculum worked over 12 weeks of the course on a topic in the area of private law. During these 12 weeks, the students attended lectures of 2 hours each, twice a week. Assessment for this course took place by means of a written exam, in the examination period at the end of the year.

Results

The students' responses were analysed by means of a one-way multivariate analysis of variance (MANOVA), followed by analyses of variances (ANOVA) using the Bonferroni method on each dependent variable. Calculation of effect sizes (*d* value) was used to examine the possible differences between the two in the respective factors. Guidelines for the interpretation of the *d* values generally take *d* = .2 as a small effect, *d* = .5 as a moderate effect and *d* = .8 as a large effect (Cohen, 1988; Kirk, 1996).

Table 1.2. Means and Standard Deviations of the seven key components of constructivist learning environments in the two groups (*d* = effect size)

	Traditional		PBL		d^a
Dimensions	*M*	*SD*	*M*	*SD*	
1. Arguments, discussions, debates	3.15	.62	3.40	.71	.30
2. Conceptual conflicts and dilemmas	2.68	.73	3.37	.69	.81
3. Sharing ideas with others	3.00	.70	3.61	.61	.75
4. Materials and measures targeted toward solutions	3.43	.67	3.57	.64	.17
5. Motivation toward reflections and concept investigation	3.04	.61	3.18	.67	.18
6. Meeting student needs	2.67	.65	2.99	.66	.40
7. Making meaning, real-life examples	3.42	.61	3.58	.59	.21

[a] Effect sizes calculated following Green and Salkind (2003, p. 153).

Preliminary analysis of the data involved inspection of the normality and homogeneity of the variance assumptions. Normal plots, stem-and-leaf plots and the calculation of skewness and kurtosis were used to check the normality of distribution. To test the equality of group variances the Levene statistics were calculated. All assumptions for the analysis were met.

The results of the MANOVA showed significant differences between the two learning environments on the dependent measures (Wilks's $\Lambda = .66$, $F(7, 407) = 29.66$, $p < .01$). The multivariate η^2 based on Wilks's Λ was quite strong, .34 (Green & Salkind, 2003). Table 1.2 contains the means and the standard deviations of the seven key components of constructivist learning environments in the two groups. All mean differences between the NLE group and the conventional lecture-

based group are accompanied by a small to large effect size. The effect size is about d = .2 for the factors 'materials and measures targeted toward solutions', 'motivation toward reflections and concept investigation' and 'making meaning, real-life examples'. Somewhat larger effect sizes (about d = .4) are found for the factors 'arguments, discussions, debates' and 'meeting student needs'. According to the large effect sizes (about d = .7) , the difference in perceptions between the two groups is most salient for the factors 'motivation toward conceptual conflicts and dilemmas' and 'sharing ideas with others'. From the results in Table 1.2, it seems clear that students in CLEs perceive their learning environment to be more constructivist, compared to the perceptions students have of a conventional lecture-based environment. Using the Bonferroni method, each ANOVA was tested at the .007 level (.05/7). The results of this analysis showed significant differences between the two groups on four of the seven factors: the first factor (arguments, discussions, debates; $F(1, 413) = 13.39$, $p < .007$, $\eta^2 = .03$), the second factor (conceptual conflicts and dilemmas; $F(1, 413) = 94.92$, $p < .007$, $\eta^2 = .19$), the third factor (sharing ideas with others; $F(1, 413) = 87.77$, $p < .007$, $\eta^2 = .18$) and the sixth factor (meeting student needs; $F(1, 413) = 24.92$, $p < .007$, $\eta^2 = .06$). For the other factors, the CLE and the conventional lecture-based learning environment group did not differ significantly from each other. It should be noted, however, that although students perceive the factor 'meeting students needs' as being more present in the CLE, the mean score for this dimension (2.99) is low. On the other hand, it is striking that the factors 'materials and measures targeted toward solutions' and 'making meaning, real-life examples' are perceived by the students in the conventional lecture-based environment as relatively highly present (with mean scores respectively of 3.43 and 3.42).

Conclusion and discussion

Research into students' perceptions of a learning environment reveals its impact on the way students cope with that learning environment and, consequently, their learning results (Brekelmans, van den Eeden, Terwel & Wubbels, 1997; Entwistle & Tait, 1990; Fraser, Walberg, Welch & Hattie, 1987; Segers & Dochy, 2001). This article investigated whether students in a problem-based learning environment perceive their learning environment as more constructivist compared to the perceptions that students have of a conventional lecture-based environment. Learning environments based on constructivism have the potential to improve the educational outcomes for students in higher education (Lea, Stephenson & Troy, 2003). The CLE used in this study was a variant of PBL, which is claimed to be consistent with constructivist features (Birenbaum, 2003; Hendry, Frommer & Walker, 1999; Russell, Creedy & Davis, 1994; Savery & Duffy, 1995; Segers, Dochy & De Corte, 1999). Moreover, constructivism is the underlying theory referred to when the superior effects of PBL are postulated (Dochy, Segers, Van den Bossche & Gijbels, 2003). Of particular interest was the question of for which factors the differences between students' perceptions of the CLE and the conventional lecture-based environment were the largest.

Using the questionnaire of Tenenbaum, Naidu, Jegede and Austin (2001) to probe into students' perceptions of their learning environments, it became clear that students in the CLE perceive their learning environment to be more constructivist when compared to the perceptions students have of a conventional lecture-based environment. According to the effect size as measured by the *d*-index, the difference in perception between the two groups was most salient for the factor 'conceptual conflicts and dilemmas'. Tenenbaum et al. (2001) argue that this factor, stressing the idea that knowledge cannot be found 'out there' and consequently is not certain, represents the constructivist approach more than others. A second factor, called 'sharing ideas with others' also clearly distinguished between the two learning environments. A recent study by Chernobilsky, Dacosta and Hmelo-Silver (2004) indicated that effective cooperative learning communities function better and are associated with more meaningful knowledge construction. These two factors determine the strength of PBL in incorporating constructivist principles. Tutors should be aware of the importance of facilitating these two factors to create a well functioning, cooperative tutorial group that promotes meaningful knowledge construction.

Although the students' perceptions differed significantly on four of the seven factors in the questionnaire and effect sizes varied from sufficient to large, the differences between the two learning environments are not 'extremely' large. For the conventional lecture-based course this means that, according to the perceptions of the students, constructivist principles are also partly incorporated. For the CLE this means that, if the learning environment claims to be highly consistent with constructivist features, at least in the perception of the students, a lot of opportunities still remain to be taken up. In particular, the factor 'meeting students' needs' was only moderately present in the CLE. This indicates that students in the CLE only had a relatively small say in the learning process. On the other hand, the conventional lecture-based environment succeeds in paying relatively large amounts of attention to the factors 'materials and measures targeted toward solutions' and 'making meaning, real-life examples', indicating that working with real-life contexts and authentic problems are not the restricted hallmark of new learning environments.

The contents of some courses lend themselves more easily to a constructivist approach than others. The question of whether the contents of the courses were sufficiently comparable should, therefore, be discussed. Although a lot of attention was paid to selecting comparable courses, no two courses are exactly the same. Nevertheless, the significant differences between units reported in the study by Tenenbaum, Naidu, Jegede and Austin (2001) involved differences in disciplines such as arts, business, education, commerce and engineering, while in our study both courses covered the topic of private law at the level of a second year law course at the university.

It should be noted that students' perceptions are not only based on the actual learning environment, but are also based on their former learning experiences and recent experiences (Segers & Dochy, 2001). In both the CLE and the conventional lecture-based environment, the courses involved in this study took place during the last part of the second year. As a consequence, the learning experiences of the

students in the conventional lecture-based group are based on other lectures in the first and second years of the curriculum, while the learning experiences of students in the CLE group are based on other courses in the PBL curriculum. It is possible that students in the CLE group judged the course under study as less (or more) constructivist, when compared to previous courses in the PBL curriculum. This is also the case for the students in the conventional lecture-based group: there is a possibility that the course under study was more (or less) congruent with previous courses in the lecture-based curriculum. Students' perceptions of a PBL course after experiences in a conventional lecture-based curriculum and students' perceptions of a lecture-based course after experience in a PBL curriculum would probably show a bigger gap between the two learning environments. This was also seen in the results of a recent study by Dochy, Segers, Van den Bossche and Struyven (2005).

As students' perceptions of the learning environment are seen as a powerful factor in the way that students cope with that learning environment, it follows that educational interventions will be less effective if they don't succeed in modifying students' perceptions in the intended way. Research into students' perceptions provides us with more information on the way new learning environments are perceived in the intended – constructivist – way. The CLE under study in this paper, a variant of PBL, is perceived by the students to be more constructivist than the conventional lecture-based environment under study. However, it seems that students' perceptions of constructivist principles in the learning environment are triggered by a greater variety in learning environments. Therefore, a global implementation of problem-based curricula, although perceived as more constructivist by the students, is not recommendable. Rather, we believe that future research on CLE should focus on the engineering of an optimal mix of learning environments and take into account students' perceptions of the blend of lectures, problem- and case-based learning groups, practical work, task-oriented learning, workplace learning, online learning opportunities, etc.

References

Albanese, M.A., & Mitchell, S. (1993). Problem-based learning: A review of literature on its outcomes and implementation issues. *Academic Medicine*, *68*, 52-81.

Ausubel, D., Novak, J., & Hanesian, H. (1978). *Educational Psychology: A cognitive view* (2nd ed.). New York: Holt, Rinehart & Winston.

Barrows, H.S. (1996). Problem-based learning in medicine and beyond. In L. Wilkerson & W.H. Gijselaers (Eds.), *Bringing problem-based learning to higher education: Theory and Practice.* New directions for teaching and learning**,** No. 68 (pp. 3-13). San Francisco, CA: Jossey-Bass.

Birenbaum, M. (2003). New insights into learning and teaching and their implications for assessment. In M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimising new modes of assessment: in search for qualities and standards* (pp. 13-36). Dordrecht: Kluwer Academic Publishers.

Brekelmans, M., van den Eeden, P., Terwel, J., & Wubbels, T. (1997). Student characteristics and learning environment interactions in mathematics and physics education: A resource perspective. *International Journal of Educational Research*, *27* (4), 283-292.

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & R. Stine (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Bruner, J.S. (1959). Learning and thinking. *Harvard Educational Review*, *29*, 184-192.

Bruner, J.S. (1961). The act of discovery. *Harvard Educational Review*, *3*, 21-32.

Chernobilsky, E., Dacosta, M.C., & Hmelo-Silver, C.E. (2004). Learning to talk the educational psychology talk through a problem-based course. *Instructional Science*, *32*, 319-356.

Cohen, J. (1988). *Statistical power analysis for the behavorial sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

Dewey, J. (1910). *How we think.* Boston: Health & Co.

Dewey, J. (1944). *Democracy and education.* New York: Macmillan.

Dochy, F., Segers, M., Van den Bossche, P., & Gijbels, D. (2003). Effects of problem-based learning: A meta-analysis. *Learning and Instruction*, *13* (5), 533-568.

Dochy, F., Segers, M., Van den Bossche, P., & Struyven, K. (2005). Students' perceptions of a problem-based learning environment. *Learning Environments Research*, *8* (1), 41-66.

Entwistle, N.J., & Tait, H. (1990). Approaches to learning, evaluations of teaching, and preferences for contrasting academic environments. *Higher Education*, *19* (2), 169-194.

Fraser, B.J., Walberg, H.J., Welch, W.W., & Hattie, J.A. (1987). Syntheses of educational productivity research. *International Journal of Educational Research*, *11* (2), 145-252.

Green, S.B., & Salkind, N.J. (2003). *Using SPSS for Windows and Macintosh. Analyzing and understanding data* (3rd ed.). New Jersey: Pearson Education.

Goodyear, P., & Hativa, N. (2002). Introduction: Research on teacher thinking, beliefs and knowledge in higher education. In N. Hativa & P. Goodyear (Eds.), *Teacher thinking, beliefs and knowledge in higher education* (pp. 1-13). Dordrecht: Kluwer academic publishers.

Guay, F., Marsh, H.W., & Boivin, M. (2003). Academic Self-concept and Academic Achievement: Developmental Perspectives on their Causal Ordering. *Journal of Educational Psychology*, *95* (1), 124-136.

Harris, K.R., & Alexander, P.A. (1998). Integrated, constructivist education: Challenge and reality. *Educational Psychology Review*, *10* (2), 115-127.

Hendry, G.D., Frommer, M., & Walker, R.A. (1999). Constructivism and problem-based learning, *Journal of Further and Higher Education*, *23* (3), 359-371.

Kennedy, M.M. (1997). The connection between research and practice. *Educational Researcher*, *26* (7), 4-12.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56* (5), 746-759.

Lea, S.J., Stephenson, D., & Troy, J. (2003). Higher education students' attitudes toward student-centred learning: beyond 'educational bulimia'? *Studies in Higher Education*, *28* (3), 321-334.

Mehrens, W.A., & Lehmann, I.J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). Fort Worth: Holt, Rinehart & Winston.

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

Piaget, J. (1954). *The construction of reality in the child.* New York: Basic Books.

Resnick, L.B. (1994). Situated rationalism: Biological and social preparation for learning. In L.A. Hirschfeld & S.A. Gelman (Eds.), *Mapping the mind* (pp. 474-494). New York: Cambridge University Press.

Rogers, C.R. (1969). *Freedom to learn.* Colombus, Ohio: Charles E. Merill Publishing Company.

Russell, A.L., Creedy, D., & Davis, J. (1994). The use of contract learning in PBL. In S.E. Chen, S.E. Cowdroy, A.J. Kingsland & M.J. Ostwald (Eds.), *Reflections on problem based learning* (pp. 57-72). Sydney: Australian Problem Based Network.

Savery, J.R., & Duffy, T.M. (1995). Problem-based learning: An instructional model and its constructivist framework. *Educational Technology*, *35* (5), 31-38.

Segers, M. (1996). Assessment in a problem-based economics curriculum. In M. Birenbaum & F. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior learning* (pp. 201-226). Boston: Kluwer Academic Press.

Segers, M., & Dochy, F. (2001). New assessment forms in problem-based learning: The value-added of the students' perspective. *Studies in Higher Education*, *26* (3), 327-343.

Segers, M., Dochy, F., & De Corte, E. (1999). Assessment practices and students' knowledge profiles in a problem-based curriculum. *Learning Environments Research*, *12* (2), 191-213.

Simons, R.J., van der Linden, J., & Duffy, T. (2000). New learning: Three ways to learn in a new balance. In R.J. Simons, J. van der Linden & T. Duffy (Eds.), *New learning* (pp. 1-20). Dordrecht: Kluwer Academic Publishers.

Taylor, P.C., & Fraser, B.J. (1991, April). *Development of an instrument for assessing constructivist learning environments.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Taylor, P.C., Fraser, B.J., & Fisher, D.L. (1997). Monitoring constructivist classroom learning environments. *International Journal of Educational Research*, *27* (4), 293-302.

Tenenbaum, G., Naidu, S., Jegede, O., & Austin, J. (2001). Constructivist pedagogy in conventional on-campus and distance learning practice: an exploratory investigation. *Learning and Instruction*, *11* (2), 87-111.

Tynjälä, P. (1999). Towards expert knowledge? A comparison between a constructivist and a traditional learning environment in the University. *International Journal of Educational Research*, *33* (5), 355-442.

Von Glaserfeld, E. (1993). Questions and answers about radical constructivism. In Tobin, K. (Ed.), *The practice of constructivism in science education* (pp. 23-38). Hildale, NJ: Lawrence Erlbaum.

Vosniadou, S. (1996). Towards a revised cognitive psychology for new advances in learning and instruction. *Learning and Instruction*, *6* (2), 95-109.

Windschitl, M. (2002). Framing constructivism in practice as the negotiation of dilemmas: An analysis of the conceptual, pedagogical, cultural, and political challenges facing teachers. *Review of Educational Research*, *72* (2), 131-175.

Chapter 2

The relationship between students' approaches to learning and the assessment of learning outcomes[2]

Abstract

The purpose of the present study is to gain more insight into the relationship between students' approaches to learning and students' quantitative learning outcomes, as a function of the different components of problem-solving that are measured within the assessment. Data were obtained from two sources: the revised two factor study process questionnaire (R-SPQ-2F) and students' scores in their final multiple-choice exam. Using a model of cognitive components of problem-solving translated into specifications for assessment, the multiple-choice questions were divided into three categories. Three aspects of the knowledge structure that can be targeted by assessment of problem-solving were used as the distinguishing categories. These were: understanding of concepts; understanding of the principles that link concepts; and linking of concepts and principles to application conditions and procedures. The 133 second year law school students in our sample had slightly higher scores for the deep approach than for the surface approach to learning. Plotting students' approaches to learning indicated that many students had low scores for both deep and surface approaches to learning. Correlational analysis showed no relationship between students' approaches to learning and the components of problem-solving being measured within the multiple-choice assessment. Several explanations are discussed.

---

[2] Based on: Gijbels, D., van de Watering, G., Dochy, F., & Van den Bossche, P. (2005). The relationship between students' approaches to learning and the assessment of learning outcomes. *European Journal of Psychology of Education*, *XX* (4), 327-341.

Introduction

Since its original publication, nearly 30 years ago, the paper by Marton and Säljö (1976) has served as an impetus for the study of students' approaches to learning in order to search for the fundamental differences students have in their approaches to engaging in learning tasks (Biggs, 1987). The study by Marton and Säljö (1976) introduced two concepts which have been widely used in educational research: 'deep' and 'surface' approaches to learning. The concept of the deep approach is associated with students' intentions to understand and construct the meaning of the content to be learned, whereas the concept of the surface approach refers to students' intentions to learn by memorizing and reproducing the factual contents of the study materials.

The original Gothenburg group looked at students' ways of approaching learning in a more qualitative way (Marton, 1981). Others, like the research group of Entwistle in the United Kingdom (Entwistle & Ramsden, 1983) or Biggs and his colleagues in Australia (1987), developed questionnaires and investigated the approaches in a more quantitative way. Although there are substantial differences between the aims, methods, and results of the different studies, they all have in common the dichotomy between a deep approach and a surface approach in students' learning (Prosser & Trigwell, 1999). Besides these two core concepts of approaches to learning, a kind of mixed approach to learning, called the strategic (or achieving) approach, is often identified (Biggs, 1993; Entwistle, 1991). The strategic approach can take place through either deep or surface processing, in line with the demands of the context (Mäkinen, 2003).

An interesting question during this time has been the relationship between students' approaches to learning and students' learning outcomes. Although the results seem to be inconsistent, the use of a deep learning approach is, in general, associated with higher quality learning outcomes and a surface approach with lower quality learning outcomes (Crawford, Gordon, Nicholas & Prosser, 1998; Hazel, Prosser & Trigwell, 1996; Snelgroove & Slater, 2003; Trigwell & Prosser, 1991; Van Rossum & Schenk, 1984; Zeegers, 2001). Van Rossum and Schenk (1984) used the Structure of the Observed Learning Outcome (SOLO) taxonomy to describe the quality of the learning outcomes of 69 first-year psychology students. The SOLO taxonomy consists of five structural categories of learning outcomes, going from the lowest level: 'pre-structural' (an irrelevant response), to the most complete level, called 'extended abstract' (Biggs & Collis, 1982). Their results show a clear positive relationship between the observation of a deep study approach and high quality learning outcomes. The difference in quantitative learning outcomes (using average exam scores) between students using the surface or the deep approach was only significant for questions measuring insight, not for questions measuring the reproduction of knowledge. Trigwell and Prosser (1991) studied the relationship between the observed approaches to learning and the learning outcomes of 122 first-year nursing students. Using the SOLO taxonomy, they found a positive correlation between a deep approach to learning and high qualitative levels in learning outcomes, but no such correlation to quantitative

differences in outcome. There were no relationships found between surface approaches to learning and qualitative or quantitative outcome measures. In a later study in the field of biology, Hazel, Prosser and Trigwell (1996) also made use of the SOLO taxonomy to analyse the learning outcomes, complimented with concept maps and phenomenographic methods. The 272 students involved in this study ended up in two clusters. In the first cluster, there was a relationship between low outcome measures, low scores on deep approaches and high scores on surface approaches. On the other hand, the second cluster reported high outcome scores related to low surface approach scores and high deep approach scores. In the field of mathematics, Crawford and colleagues (1998) found strong correlations between 300 first-year students' observed approaches to learning and their final percentage mark in their first year mathematics course. Relatively high scores on the surface approach subscale were related to low marks in the final exam, while relatively high scores on the deep approach to learning subscale were related to higher final exam scores. In a longitudinal study with 200 first-year science students, Zeegers (2001) used Biggs' (1987) Study Process Questionnaire (SPQ) and annual grade point average (GPA) scores to evaluate the predictive value of the SPQ scales on students' learning outcomes. The results showed a consistent positive correlation between the deep approach to learning and assessment outcomes. Snelgrove and Slater (2003) also used the SPQ (Biggs, 1987) with 300 nursing students and found the deep factor to be positively and significantly correlated with average grade performance.

Recently, Watkins (2001) conducted a cross-cultural meta-analysis in which the relationship between students' approaches to learning and their academic performance was one of the central questions. It was hypothesised that surface approaches to learning would be significantly negatively correlated with students' grades, whilst the deep approach would be positively related with academic achievement. The results of his study were rather disappointing, although in the expected direction, with correlations of -.11 for surface and .16 for deep approaches. In the literature, assessment is generally blamed for such disappointing results. Although a deep approach to learning is expected to lead to higher achievement (both in terms of higher quality outcomes and grades), the assessment system does not always reward the deep approach (Biggs, 1987; Marton & Säljö, 1976; Scouller, 1998; Scouller & Prosser, 1994). Entwistle, McCune and Hounsell (2003, p. 90) suggest that research findings vary "*due to differences in the extent to which understanding is explicitly rewarded in the assessment procedure*". A recent study by Minbashian, Huon and Bird (2004) tried to investigate this moderating effect of the type of exam questions in a study involving 49 third year psychology students using Entwistle and Tait's (1994) Revised Approaches to Studying Inventory and short essay questions. However, the hypothesis that a deep approach would be more effective for questions of higher cognitive order than for questions of lower cognitive order could not be confirmed: the observed relationship was not significant and was in the opposite direction.

The present study

The relationship between students' approaches to learning and the assessed (quantitative) learning outcomes is of interest to the present study. Today's stated learning outcomes in higher education are, to a large extent, congruent with trends in the marketplace. "*With more and more routine jobs being turned over to robots and other automated devices, the jobs left for humans tend to be less routine requiring more problem-solving skill for adequate job performance*" (Gagné, Yekovich & Yekovich, 1993, p. 210). In essence, a primary goal in higher education seems to be to enable students to solve complex problems in an efficient way (Engel, 1997; Gagné et al., 1993; Poikela & Poikela, 1997; Segers, 1997).

The literature on problem-solving is characterized by a wide variety of theoretical frameworks (e.g. De Corte, 1996; Glaser, Raghavan & Baxter, 1992; O'Neil & Schacter, 1997; Schoenfeld, 1985; Smith, 1991). Despite their differences in details and terminology, all models agree that an organized and structured domain-specific knowledge base and meta-cognitive functions that operate on that knowledge are essential components of successful problem-solving. There is also a fairly broad consensus that motivation and beliefs account for differences in problem-solving. As a consequence, the purpose of the present study is to explore further the relationship between students' approaches to learning and their quantitative learning outcomes, from the perspective of the different components of problem-solving that are measured with the assessment.

Research context

The study was conducted in a European law school using problem-based learning (PBL). Educating for successful problem-solvers is one of the main goals of PBL (Dochy, Segers, Van den Bossche & Gijbels, 2003). Although originally developed for medical training in Canada, the orthodox version of PBL has been modified and applied globally in many disciplines (Gijselaers, 1995). The present study took place in a course on public law. Students had to work in small tutorial groups (12-18 students) and met twice a week under the supervision of a teacher (tutor). During each session, students were confronted with a range of tasks which they had to analyse and solve by formulating 'learning goals' for self-study. In the next session, students reported their findings and started to analyse new problems. As well as this, students were enrolled on a weekly basis in somewhat larger 'practical groups' (24-36 students) and had one lecture a week. During the course, students had the opportunity to complete 3 assessment tasks on a voluntary basis. These could result in a bonus, which was added to the score of the final exam.

Method

Participants

The sample consisted of 133 second-year Law students (65% females and 35% males, mean age: 20.6) who were enrolled for the first time in a second year course on public law, using PBL. The students were divided into 17 small groups that were tutored by 7 teachers.

Instruments

Data were obtained from two sources: a questionnaire and students' final exam results for the course.

The questionnaire was a Dutch translation of Biggs, Kember and Leung's (2001) Revised two Factor Study Process Questionnaire (R-SPQ-2F). The R-SPQ-2F is a more refined version of Biggs' (1987) original Study Process Questionnaire (SPQ). In the theoretical framework of the SPQ, three approaches to learning (surface, deep and achieving) are proposed, each with a motive and strategy subscale. Kember and Leung (1998) conducted a study with over 7000 Hong Kong students which investigated the construct and internal reliability of the SPQ. The results indicated that a model with two factors had the best fit. Other studies, including cross-cultural research, have also shown a two factor solution with deep and surfaces approaches, rather than the initial three factor solution, accounted for most of the variance (Snelgrove & Slater, 2003; Watkins & Regmi, 1996; Zhang, 2000). Biggs and colleagues (2001) accordingly refined the SPQ. The revised two factor SPQ consists of 20 items which are scored on a 5 point Likert scale and categorizes students into two different types of approaches to learning: 'surface learning approaches' and 'deep learning approaches', each containing two subscales, 'motive' and 'strategy'. The study of Biggs and colleagues (2001) indicated that the R-SPQ-2F had reasonable Cronbach's alpha values for scale reliability and desirable goodness of fit with the intended two factor model. Leung and Chan (2001) investigated the psychometric properties and applicability of the R-SPQ-2F in the Hong Kong Chinese context. Their results also indicated reasonably good reliability coefficients and goodness of fit for the two factor model. Our Dutch translation of the questionnaire resulted in acceptable Cronbach's alpha values for the 2 factor model: surface learning approaches (Cronbach's alpha = .75) and deep learning approaches (Chronbach's alpha = .73). The subscales deep motive (Chronbach's alpha = .60), deep strategy (Chronbach's alpha = .54), surface motive (Chronbach's alpha = .65) and surface strategy (Chronbach's alpha = .48) had lower reliability coefficients and are not used for further analysis. Confirmatory Factor Analysis (CFA) using LISREL 8.52 was performed to verify whether the two factor structure could be validated (Jöreskog & Sörbom, 2002). The results indicated that the data set fits the two factor model fairly well ($\chi^2/df = 1.64$, *RMSEA* = .07). Sufficient fit values are smaller than 2.0 for the first (Dolmans, Wolfhagen, Scherpbier & van der Vleuten, 2003; Tenenbaum, Naidu, Jegede & Austin, 2001),

and smaller than .08 for the Root Mean Square Error of Approximation (Browne & Cudeck, 1993; Guay, Marsh & Boivin, 2003; Sachs & Gao, 2000).

The final exam consisted of 40 multiple-choice questions (Cronbach's alpha = .70). In order to distinguish between the different components of problem-solving for each question in the final exam, we used Sugrue's (1993, 1995) model of cognitive components of problem-solving. Sugrue translated her model into specifications for the assessment of the main cognitive components of problem-solving, and is therefore useful for our purpose. The assumption made by Sugrue is that successful problem-solving in a given domain results from the interaction of knowledge structure, meta-cognitive functions and motivation. For each of the three categories of cognitive components, Sugrue describes a limited set of variables that should be targeted by assessment. In relation to the final exam used in our study, the knowledge structure is of special interest. Three levels which the assessment can appeal to are distinguished in the knowledge structure. These three levels are presented in Figure 2.1, which gives an overview of possibilities for the assessment within a 'selection' format, of which multiple-choice questions are obviously the most well-known example (Sugrue, 1995).

Levels in the knowledge structure	
Concepts	- Select examples of concepts - Distinguish between examples that are and are not instances of the concept of interest
Principles	- Select best/similar/dissimilar problems - Select best prediction - Select best explanation for event
Application	- Select correct procedure for identifying instances - Select most appropriate procedure to change the state of a concept by manipulating another

Figure 2.1. Construct-by-format matrix for measuring constructs related to the knowledge structure with selection-formatted questions (after Sugrue, 1995).

At the first level, assessment of the understanding of concepts, which can be defined as "*a category of objects, events, people, symbols or ideas that share common defining attributes or properties and are identified by the same name*" (Sugrue, 1993, p.9) is the core issue. In this case, students are confronted with several examples of the concept and asked to select those which are instances of the concept of interest. At the second level, understanding of the principles that link concepts, or in other words the organisation of the knowledge structure, is the subject of assessment. Sugrue (1993, p. 9) defines a principle as "*a rule, law, formula, or if-then statement that characterizes the relationship (often causal) between two or more concepts. Principles can be used to interpret problems, to guide actions, to troubleshoot systems, to explain why something happened, or to predict the effect a change in some concept(s) will have on other concepts.*" In this case, students could be asked to select the most appropriate prediction or solution from a list of given descriptions of an event. The third and final level targets the linking of concepts and principles to application conditions and procedures by assessment. A 'procedure' is defined as "*a set of steps that can be carried out either to classify an instance of a concept or to change the state of a concept to effect a*

change in another" (Sugrue, 1993, p. 22) and 'conditions' as "*aspects of the environment that indicate the existence of an instance of a concept, and/or that a principle is operating or can be applied and/or that a particular procedure is appropriate*" (Sugrue, 1993, p. 22). At this level, the organized knowledge is applied under appropriate circumstances. A student can be asked to select the most appropriate procedure for a given task in order to reach a particular goal.

A major benefit of Sugrue's model is that it can easily be used to classify questions. The model allows the use of different assessment reviewers for one assessment, even if the reviewers have little subject knowledge. Two reviewers categorized the questions in the final exam separately. After that, items that were differently classified were discussed until a clear consensus was reached. Finally, 17 questions were classified as being at the 'concepts' level, 11 questions at the 'principles' level and 12 questions at the 'application' level.

Procedure

Students were asked to complete the R-SPQ-2F questionnaire during one of the tutorial sessions near the end of a second year law course. The final exam was administered one week after the end of the course.

Results

Results were plotted and analysed by means of descriptive statistics for the measures used in the present study and by correlation analysis to probe into the relationships between students' approaches to learning and the different components of problem-solving measured within the final exam.

Table 2.1. Descriptive statistics for the main measures used

Variable	*M*	*SD*	%
Deep approach	2.99	.51	
Surface approach	2.21	.59	
Concepts mark	12.60	2.27	(74.12%)
Principles mark	7.24	2.01	(65.82%)
Application mark	7.52	1.82	(62.67%)
Total mc-exam mark	27.36	4.91	(68.40%)

Table 2.1 presents descriptive statistics for the measures used in the present study. Students' scores for deep approaches were higher than their scores for surface approaches in our sample. For the assessments, students had highest average scores for the questions measuring concepts (74.12% of the questions correct). The second highest scores were obtained for questions measuring principles (65.82% of the questions correct). The questions measuring application had the lowest scores (62.67% of the questions correct).

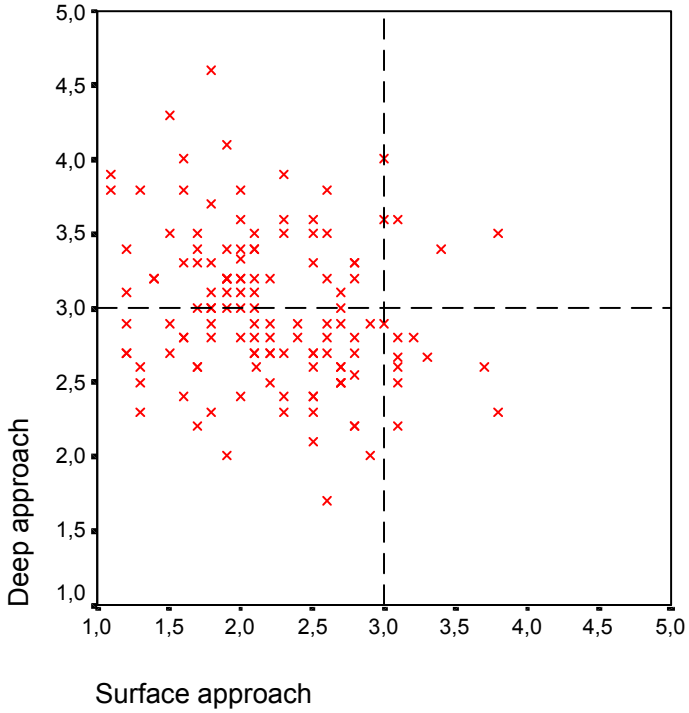

Figure 2.2. Plot of study approaches

The plot in Figure 2.2 indicates that most students fitted into two groups: a group of students with high scores for deep approach and low scores for surface approach and a group with low scores for both the deep and surface approach. Very few students employed high levels of both deep and surface approaches to learning. The group of students that had high scores for the surface approach and low scores for the deep approach to learning is also small. Further analysis indicated that for the surface approach to learning, the mean score of women (M = 2.07, SD = .59) differs significantly from men's score (M = 2.43, SD = .53, $F(1, 129) = 12.03$, $p <$ .01). The deep approach to learning shows a statistically significant relationship to students' ages: the older the students, the more deep approaches to learning are used (r = .22, p = .01).

The correlations of the main variables used in this study are presented in Table 2.2. The interrelationships between the three categories of measured components of problem-solving and the total exam grade are all high and statistically significant. However, neither students' final exam grades, nor their sub-results on questions in the exam asking for different components of problem-solving, are significantly (p-values all exceed .05) related to the extent to which they use either deep or surface approaches to learning.

Table 2.2. Correlations among the measures used in the present study

Variable	1	2	3	4	5	6
1. Deep approach	1.000					
2. Surface approach	-.232**	1.000				
3. Concepts mark	-.031	-.161	1.000			
4. Principles mark	.040	-.119	.585**	1.000		
5. Application mark	-.088	.020	.366**	.446**	1.000	
6. Total mc-exam	-.031	-.116	.837**	.845**	.722**	1.000

** Correlation is significant at the .01 level

Discussion

In the present study we wanted to gain more insight into the relationships between students' approaches to learning and the different components of problem-solving that were being measured using a multiple-choice assessment. The students in our sample showed slightly higher scores for a deep approach than for a surface approach to learning. However, plotting students' approaches to learning indicated that a lot of students had low scores for both deep and surface approaches. Previous research has shown that a profile which consists of low (or high) scores on both deep and surface approaches is quite typical of novice students even though this kind of combination could be entitled 'disintegrated' or 'dissonant' or 'not yet established' (Entwistle, Meyer & Tait, 1991; Lindblom-Ylänne & Lonka, 1999, Lonka & Lindblom-Ylänne, 1996). A recent study with 110 (first-, second-, and third-year) law students enrolled in problem-based courses in legal history and communication skills for lawyers (Lindblom-Ylänne, 2003) revealed that 23% of the students showed clearly dissonant study orchestrations. These students seemed to "*lack the metacognitive skills to evaluate how functional their study practices were in their learning environment, and admitted to having problems with their study strategy. Many of the students realised that their study methods were not suitable for studying law, but they did no know how to develop them.*" (p.73)

Further analysis of our data indicated that male students adopted a significantly higher level of surface approaches and that older students adopted significantly deeper approaches to learning. The first contradicts prior research by Richardson (1993), which showed no consistent evidence of significant differences between men's and women's approaches to learning. The latter is in line with Richardson's later (1995) research which indicated that older students are more meaning oriented when studying.

The results of our correlational analysis indicated no relationships between students' approaches to learning and the components of problem-solving being measured within the multiple-choice assessment. From our data, it is impossible to associate the expected employment of deep learning approaches with higher assessment outcomes. The view that within the same question format (i.e. multiple-choice questions), students with different approaches to learning would score differently on questions measuring different components of problem-solving is also not supported.

Our results are in line with those of Minbashian, Huon and Bird (2004): namely, there is no evidence that a deep approach to learning would be more effective for questions assessing more complex components of problem-solving. One of the explanations they give is that the wording of the questions (from what we deduced by means of Sugrue's (1993, 1995) model of the components of problem-solving being measured) is by itself not sufficient to influence the nature of students' responses. The method of assessment probably has more influence on the way students study for, and respond to, exam questions (Minbashian et al., 2004). Related to this, students' perceptions of the method of assessment (i.e. multiple-choice questions) could be seen as a mediating factor (Segers, Dochy & Cascallar, 2003). A recent review (Struyven, Dochy & Janssens, 2003) indicated that students' perceptions of assessment have considerable influences on students' approaches to learning. Scouller (1998) found that success in multiple-choice examinations was related to the perception of the questions as assessing lower levels of cognitive processes and the non-employment of deep strategies. Although we did not take students' perceptions of the assessment into account, it is possible that students do not differentiate between the different questions within the same assessment method.

Another possible explanation for the lack of clear results could be the effectiveness of the classification model. We used Sugrue's (1993, 1995) model of the cognitive components of problem-solving to categorize the different questions in the multiple-choice exam, according to the three components of problem-solving that were to be measured. Although the model seems clear and exhaustive for multiple-choice questions, the two assessment reviewers reported difficulties in categorizing some questions. In their opinion, questions asking for 'the reproduction of facts', although important for assessment according to most of the law teachers, at first sight had no place in the model. Since the difference between 'a concept' and 'a fact' appeared difficult to explain, after discussing these questions the reviewers agreed to classify them in the category of 'understanding concepts'. Sugrue (1995) remarks that her model lends itself extremely well to domains such as science, mathematics, economics and geography, but that it might not be easy to use in other domains such as history. The reviewers' difficulties in classifying some questions indicate that law could also be added as a domain for which the model is complicated.

Another problem is that one can classify questions in terms of components of problem-solving being measured but one can not be sure in a multiple-choice exam that, when a student gives the wrong answer, he fails to achieve the components of problem-solving being measured by the question. When a student doesn't understand two related concepts in a familiar problem, the student will fail to select similar problems, not because (s)he does not understand the relationship between the two concepts, but simply because (s)he does not have basic understanding of the concepts. This would mean that it is very difficult to investigate the relationship between students' approaches to learning and components of problem-solving being measured within a multiple-choice setting. Furthermore, it suggests that the students' answers should be the unit of analysis, rather than the questions. However, this problem could be solved in the assessment construction process by including a

multiple-choice question for each of the three components of problem-solving for each subject tested in the assessment.

The question whether it is at all possible to measure deep-level processing of knowledge as well as problem-solving skills by multiple-choice questions should also be raised here. Although it is argued that multiple-choice questions can be appropriate for assessing the understanding and application of knowledge as well as the capability to analyse situations and solve problems (Anderson & Krathwohl, 2001; Haladyna, 2004), others, like Driessen and van der Vleuten (2000) state that it is only possible to assess higher cognitive skills if multiple-choice questions are combined with another type of assessment like essay question using problem vignettes.

As well as the method of assessment, the content and method of teaching also influence the way in which students study for and respond to exam questions (Minbashian, Huon & Bird, 2004). The present study was carried out within the context of a second year law course. A recent study by Mäkinen and Olkinuora (2003) in Finland found that, in contrast to the situation in the faculty of medicine, first year law school students' study credits were negatively correlated with a deep learning orientation, whereas the grades of second year law school students were positively correlated with a deep learning orientation. Not only the content of teaching, but also the teaching method in our study, problem-based learning, must be taken into consideration. According to Biggs (2003), problem-based learning is an instructional approach that has the potential to facilitate deep approaches to learning. Although on average students had slightly higher scores for a 'deep approach' than for a 'surface approach', there is no tendency towards a use of deep approaches to learning, despite the problem-based learning environment. Seemingly, the current conditions of teaching and assessment did not make all students decide that a deep approach would give the best results, as indeed it didn't.

Interesting questions for future research in this respect would be what kind of influences the tutor or the tutorial group has on students' approaches to learning in problem-based learning environments. Trigwell, Prosser and Waterhouse (1999) conducted an empirical study which showed that approaches to teaching are associated with approaches to learning: teacher-centred approaches to teaching are related to a surface approach to learning. Conversely, student-centred approaches to teaching were related to deeper approaches to learning. In legal education, the difference between 'traditional PBL tutors' and tutors adopting the 'Socratic method' is well known and could be a possible moderator of students' approaches to learning (Liddle, 1999, 2000).

The tutorial group also influences students' approaches to learning and the outcomes. A recent study by Lindblom-Ylänne, Pihlajamäki and Kotkas (2003) showed that if students in a PBL group participate more evenly and actively in the discussions they achieve higher grades as a group.

Finally, like gender and age, some other elements in the learning-environment (such as the possibility students had to make three assessment tasks during the course) will have had an influence on students' approaches to learning and possibly also indirect on their final exam. More general factors such as prior academic achievement or GPA (Snelgrove & Slater, 2003; Young, 1993, Zeegers, 2001), self-

confidence (Watkins & Biggs, 1996) and academic self-efficacy (Pintrich & De Groot, 1990) are potential moderators in the relation between students' approaches to learning and students' quantitative learning outcomes which should be subject of future research.

To conclude, the second-year law students enrolled in a problem-based course showed slightly higher scores for a deep approach than for a surface approach to learning. However, a lot of students had low scores for both deep and surface approaches, indicating 'dissonant' study strategies. For the first-year law students the faculty recently developed an on-line environment 'legal study- and assessment skills' where students can find information about how to develop suitable study-strategies for their law study. The present study indicates that also second-year law students would benefit from this on-line environment. The results of this study confirm to some extent, previous findings that student approaches to learning are sensitive to the learning context, as well as student age and gender, and that the values for deep and surface learning approaches may be related to academic outcomes. The specific findings here show these correlations to be weak and not statistically significant. It was suggested that students' perceptions of the method of assessment will have had considerable influences on students' approaches to learning. When re-engineering the assessment, at least more authentic assessment-tasks should be added to the multiple-choice examination (Gijbels, van de Watering & Dochy, 2005). Further research should probe into the relationship between students' approaches to learning and their outcomes on and perceptions of a blend of assessment methods.

References

Anderson, L., & Krathwohl, D. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Biggs, J. (1987). *Student approaches to learning and studying.* Melbourne: Australian Council for Educational Research.

Biggs, J. (1993). What do inventories of students' learning processes really measure? A theoretical review and clarification. *British Journal of Educational Psychology*, *63*, 3-19.

Biggs, J. (2003). *Teaching for Quality Learning at University* (2nd ed.). Buckingham: SRHE and Open University Press.

Biggs, J., & Collis, K. (1982). *Evaluating the Quality of Learning: The SOLO Taxonomy.* New York: Academic Press.

Biggs, J., Kember, D., & Leung D.Y.P. (2001). The revised two-factor study process questionnaire: R-SPQ-2F. *British Journal of Educational Psychology*, *71*, 133-149.

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & R. Stine (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Crawford, K., Gordon, S., Nicholas, J., & Prosser, M. (1998). Qualitatively different experiences of learning mathematics at university. *Learning and Instruction*, *8*, 455-468.

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., Segers, M., Van den Bossche, P., & Gijbels, D. (2003). Effects of Problem-based Learning: A Meta-analysis. *Learning and Instruction*, *5* (13), 533-568.

Dolmans, D., Wolfhagen, H., Scherpbier, A., & van der Vleuten, C. (2003). Development of an instrument to evaluate the effectiveness of teachers in guiding small groups. *Higher Education*, *46* (4), 431-446.

Driessen, E., & van der Vleuten, C. (2000). Matching student assessment to problem-based learning: lessons from experience in a law faculty. *Studies in Continuing Education*, *22* (2), 235-248.

Engel, C.E. (1997). Not just a method but a way of learning. In D. Boud & G. Feletti (Eds.), *The Challenge of Problem-based Learning* (2nd ed., pp.17-27). London: Kogan Page.

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

Entwistle, N., & Ramsden, P. (1983). *Understanding Student Learning*. London: Croom Helm.

Entwistle, N., & Tait, H. (1994). *The Revised Approaches to Studying Inventory*. University of Edinburgh: Centre for Research into Learning and Instruction.

Entwistle, N., McCune, V., & Hounsell, J. (2003). Investigating Ways of Enhancing University Teaching-Learning Environments: Measuring Students' Approaches to Studying and Perceptions of Teaching. In E. De Corte, L. Verschaffel, N. Entwistle & J. van Merriënboer (Eds.), *Powerful Learning Environments: Unravelling Basic Components and Dimensions.* Amsterdam: Pergamon, Elsevier Science.

Entwistle, N.J., Meyer, J.H.F., & Tait, H. (1991). Student failure: Disintegrated patterns of study strategies and perceptions of the learning environment. *Higher Education*, *21*, 249 -261

Gagné, E.D., Yekovich, C.W., & Yekovich, F.R. (1993). *The cognitive psychology of school learning* (2nd ed.). New York: HarperCollins College publishers.

Gijbels, D., van de Watering, G., & Dochy, F. (2005). Integrating assessment tasks in a problem-based learning environment. *Assessment and Evaluation in Higher Education*, 3*0* (1), 71-84.

Gijselaers, W. (1995). Perspectives on problem-based learning. In W. Gijselaers, D. Tempelaar, P. Keizer, J. Blommaert, E. Bernard & H. Kasper (Eds.), *Educational Innovation in Economics and Business Administration: the Case of Problem-based Learning* (pp. 39 - 52). Norwell, Mass.: Kluwer.

Glaser, R., Raghavan, K., & Baxter, G.P. (1992). *Cognitive theory as the basis for design of innovative assessment: Design characteristics of science assessments* (CSE Tech. Rep. No. 349). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Guay, F., Marsh, H.W., & Boivin, M. (2003). Academic Self-concept and Academic Achievement: Developmental Perspectives on their Causal Ordering. *Journal of Educational Psychology*, *95* (1), 124-136.

Haladyna, T.M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Hazel, E., Prosser, M., & Trigwell, K. (1996). Student learning of biology concepts in different university contexts. *Research and Development in Higher Education*, *19*, 323-326.

Jöreskog, K., & Sörbom, D. (2002). *LISREL 8.52*. Chicago: Scientific Software International, Inc.

Kember, D., & Leung, D.Y.P. (1998). The dimensionality of approaches to learning: an investigation with confirmatory factor analysis on the structure of the SPQ and LPQ. *British Journal of Educational Psychology*, *68*, 395-407.

Leung, M., & Chan, K. (2001, December). *Construct validity and psychometric properties of the revised two-factor study process questionnaire (R-SPQ-2F) in the Hong Kong context.* Paper presented at the AARE conference, Perth, Australia.

Liddle, M. (1999). Problem based learning in law: Student attitudes. In J. Marsh (Ed.) *Implementing Problem Based Learning Project: Proceedings of the First Asia Pacific Conference on Problem Based Learning* (pp. 235-240). Hong Kong: The University Grants Committee of Hong Kong, Teaching Development Project.

Liddle, M. (2000). Student attitudes toward problem-based learning in law. *Journal on Excellence in College Teaching*, *11* (2), 163-190.

Lindblom-Ylänne, S. (2003). Broadening understanding of the phenomenon of dissonance. *Studies in Higher Education*, *28* (1), 63-77

Lindblom-Ylänne, S., & Lonka, K (1999). Individual ways of interacting with the learning environment - Are they related to study success? *Learning and Instruction*, *9* (1), 1-18.

Lindblom-Ylänne, S., Pihlajamäki, H., & Kotkas, T. (2003). What Makes a Student Group Successful? Student-Student and Student-Teacher Interaction in a Problem-Based Learning Environment. *Learning Environment Research*, *6* (1), 59-76.

Lonka, K., & Lindbom-Ylänne, S. (1996). Epistemologies, conceptions of learning, and study practices in medicine and psychology. *Higher Education*, *31* (1), 5-24.

Mäkinen, J. (2003). *University students' general study orientations. Theoretical background, measurements, and practical implications* (Dissertation). Turku: Turun Yliopisto.

Mäkinen, J., & Olkinuora, E. (2003, August). *Personal experience of studying and study success: A three-years follow-up study of university students.* Paper presented at the 10th biannual conference of the European Association for Research on Learning and Instruction, Padova, Italy.

Marton, F. (1981). Phenomenographic: Describing conceptions of learning. *International Journal of Educational Research*, *19*, 277-300.

Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I. Outcome and process. *British Journal of Educational Psychology*, *46*, 4-11.

Minbashian, A., Huon, G.F., & Bird, K.D. (2004). Approaches to studying and academic performance in short-essay exams. *Higher Education*, *47* (2), 161-176.

O'Neil, H.F., & Schacter, J. (1997). *Test specifications for problem-solving assessment* (CSE Tech. Rep. No. 463). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Pintrich, P.R., & De Groot, E.V. (1990). Motivation and self regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82* (1), 33-40.

Poikela, E., & Poikela, S. (1997). Conceptions of learning and knowledge - impacts on the implementation of problem-based learning. *Zeitschrift fur Hochschuldidactic*, *21* (1), 8-21.

Prosser, M., & Trigwell, K. (Eds.) (1999). *Understanding learning and teaching. The experience in higher education.* Buckingham: The society for research into higher education.

Richardson, J.T.E. (1993). Gender Differences in Response to the Approaches to Studying Inventory. *Studies in Higher Education*, *18* (1), 3-13.

Richardson, J.T.E. (1995) Mature students in higher education: II. An investigation of approaches to studying and academic performance. *Studies in Higher Education*, *20* (1), 5-17.

Sachs, J., & Gao, L. (2000). Item-level and subscale-level factoring of Biggs' Learning Process Questionnaire (LPQ) in a mainland Chinese sample. *British Journal of Educational Psychology*, *70* (3), 405- 418.

Schoenfeld, A.H. (1985). *Mathematical problem solving.* San Diego, CA: Academic Press.

Scouller, K., & Prosser, M. (1994). Students' experiences in studying for multiple choice question examinations. *Studies in Higher Education*, *19* (3)*,* 267-279.

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

Segers, M. (1997). An alternative for assessing problem-solving skills: The overall test. *Studies in Educational Evaluation*, *23* (4), 373-398.

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

Smith, M.U. (1991). *Toward a unified theory of problem-solving: Views from the content domains.* Hilsdale, NJ: Lawrence Erlbaum.

Snelgrove, S., & Slater, J. (2003). Approaches to learning: psychometric testing of a study process questionnaire. *Journal of Advanced Nursing*, *43* (5), 496-505.

Struyven, K., Dochy, F., & Janssens, S. (2003). Students' perceptions about new modes of assessment in higher education: A review. In M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimizing new modes of assessment: In search of qualities and standards* (pp. 171-224). Boston/Dordrecht: Kluwer Academic

Sugrue, B. (1993). *Specifications for the design of problem-solving assessments in science. Project 2.1 designs for assessing individual and group problem-solving*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Sugrue, B. (1995). A theory-based framework for assessing domain-specific problem solving ability. *Educational Measurement: Issues and Practice*, *(14)* 3, 29-36.

Tenenbaum, G., Naidu, S., Jegede, O., & Austin, J. (2001). Constructivist pedagogy in conventional on-campus and distance learning practice: an exploratory investigation. *Learning and Instruction*, *11* (2), 87-111.

Trigwell, K., & Prosser, M. (1991). Relating approaches to study and the quality of learning outcomes at the course level. *British Journal of Educational Psychology, 61,* 265-275.

Trigwell, K., Prosser, M., & Waterhouse, F. (1999). Relations between teachers' approaches to teaching and students' approach to learning. *Higher Education, 37*, 57-70.

Van Rossum, E.J., & Schenk, S.M. (1984). The relationship between learning conception, study strategy and outcome. *British Journal of Educational Psychology*, *54*, 73-83.

Watkins, D. (2001). Correlates of Approaches to Learning: A Cross-Cultural Meta-Analysis. In R.J. Sternberg & L. Zhang (Eds.), *Perspectives on Thinking, Learning, and Cognitive Styles* (pp. 165-196). London: Lawrence Erlbaum Associates.

Watkins, D., & Biggs, J. (Eds.) (1996). *The Chinese learner: Cultural, psychological and contextual influences.* Hong Kong: University of Hong Kong, Comparative Education Research Centre.

Watkins, D., & Regmi, M. (1996). Toward the cross-cultural validation of a Western model of student approaches to learning. *Journal of Cross-Cultural Psychology*, *27* (5), 547-560.

Young, J.W. (1993). Grade adjustment methods. *Review of Educational Research*, *63* (2), 151-165.

Zeegers, P. (2001). Student learning in science: a longitudinal study. *British Journal of Educational Psychology*, *71*, 115-132.

Zhang, L.F. (2000). University students' learning approaches in three cultures: An investigation of Biggs 3P model. *Journal of Psychology*, *134* (1), 37-56.

Chapter 3

Students' assessment preferences, perceptions of assessment and their relationships to study results[3]

Abstract

The purposes of this study are to gain more insight into students' actual preferences and perceptions of assessment, into the effects of these on their performances when different assessment formats are used, and into the different cognitive process levels assessed. Data were obtained from two sources. The first was the scores on the assessment of learning outcomes, consisting of open ended and multiple-choice questions measuring the students' abilities to recall information, to understand concepts and principles, and to apply knowledge in new situations. The second was the adapted Assessment Preferences Inventory (API) which measured students' preferences as a pre-test and perceptions as a post-test. Results show that, when participating in a Constructivist Learning Environment (CLE), students prefer traditional written assessment and questions which are as closed as possible, assessing a mix of cognitive processes. Some relationships, but not all the expected ones, were found between students' preferences and their assessment scores. No relationships were found between students' perceptions of assessment and their assessment scores. Additionally, only forty percent of the students had perceptions of the levels of the cognitive processes assessed that matched those measured by the assessments. Several explanations are discussed.

[3] Based on: Van de Watering, G., Gijbels, D., Dochy, F., & van der Rijt, J. Students' assessment preferences, perceptions of assessment and their relationships to study results. *Submitted to Higher Education*.

Introduction

Assessment is an umbrella term. Understanding of it varies, depending on how one sees the role of the assessment itself in the educational process, as well as the role of the participants (the assessors and the assessees) in the education and assessment processes. The main difference is described in terms of an 'assessment culture' and a 'testing culture' (Birenbaum, 1994, 1996, 2000). The traditional testing culture is heavily influenced by old paradigms, such as the behaviourist learning theory, the belief in objective and standardized testing (Shepard, 2000), and testing being separated from instruction. Multiple-choice and open ended assessments are typical test formats of a testing culture. In the last few decades, developments in society (National Research Council, 2001) and a shift towards a constructivist learning paradigm, combined with the implementation of constructivist learning environments (CLEs), have changed the role of assessment in education. CLEs claim to have the potential to improve the educational outcomes for students in higher education which are necessary to function successfully in today's society (Simons, van der Linden & Duffy, 2000). The most fundamental change in the view of assessment is represented by the notion of 'assessment as a tool for learning' (Dochy & McDowell, 1997). In the past, assessment was primarily seen as a means to determine grades; to find out to what extent students had reached the intended objectives. Today, there is a realisation that the potential benefits of assessing are much wider and impinge on all stages of the learning process. Therefore, the new assessment culture strongly emphasises the integration of instruction and assessment, in order to align learning and instruction more with assessment (Segers, Dochy & Cascallar, 2003). The integration of assessment, learning and instruction, however, remains a challenge for most teachers (Struyf, Vandenberghe & Lens, 2001). In the UK, Glasner (1999) concludes that a number of factors, like the massification of higher education, the declining levels of resources, and concerns about the ability to inhibit plagiarism, are responsible for the persistence of traditional methods of assessment and the absence of widespread innovation. A recent report on final exams in secondary education in the Netherlands indicated that most of the exams consisted primarily of multiple-choice questions and open ended or essay questions, in spite of the effort that had been put into the implementation of new teaching and assessment methods (Kuhlemeier, de Jonge & Kremers, 2004).

Nevertheless, it is generally acknowledged that assessment plays a crucial role in the learning process and, accordingly, on the impact of new teaching methods (Brown, Rust & Gibbs, 1994; Gibbs, 1999; Scouller, 1998). Within the principle of influencing, better known under the appropriate terms 'compatibility', 'feed forward', 'backwash' or 'consequential validity' (Hofstee, 1999; Messick, 1994), three major types of effect can occur: pre assessment effects; true assessment effects; and post assessment effects (Gielen, Dochy & Dierick, 2003). The first type of effect indicates the influence of the perceived state and nature of the assessment on the study approach, especially on the way students prepare themselves for the assessment. According to Boud (1990), students focus on the subjects and levels of

(cognitive) processing asked for in the assessment which will bring in marks (resulting in higher grades). The second type of effect, called the true assessment effect, is described by Nevo (1995) as the way in which, in some cases, the assessment itself encourages the students to use higher order thinking skills, and to make links and discover relationships between ideas that they had not yet discovered while studying. This can result in a rich learning experience for the students. In post assessment, feedback is given about the position of the student with regard to the learning process and this feedback has an influence on learning behaviour. In most cases, post assessment effects are related to formative assessment. In a nutshell, the way students prepare themselves for an assessment depends on how they perceive the assessment (before, during and after the assessment), and these effects can have either positive or negative influences on learning. There also can be a discrepancy between what is actually asked, what students prefer and what students expect to be asked (Broekkamp, van Hout-Wolters, van den Bergh & Rijlaarsdam, 2004). CLEs have been developed in which schools have a balance between a test culture and an assessment culture. The effects of such environments, however, do not always demonstrate the expected outcomes (Segers, 1996). Research results show that educational change only becomes effective if the students' perceptions are also changed accordingly (Lawness & Richardson, 2002; Segers & Dochy, 2001).

As mentioned before, CLEs are not always accompanied by new methods of assessment. In this article, we want to explore which assessment formats are preferred in a CLE, how students perceive rather traditional assessment formats, and what relationships exist between students' preferences, perceptions and their assessment results. Before presenting the results, we will first describe some research into students' assessment preferences and perceptions of assessment.

Assessment preferences

Preference is described in the Webster's Dictionary as "*The act of preferring, or the state of being preferred; the setting of one thing before another; precedence; higher estimation; predilection; choice; also, the power or opportunity of choosing; as, to give him his preference*". It assumes a real or imagined choice between alternatives and the possibility of the rank ordering of these alternatives. More generally, it can be seen as a source of motivation. According to the studies of Ben-Chaim and Zoller (1997), Birenbaum and Feldman (1998), Traub and McRury (1990) and Zeidner (1987) students, especially the males (Beller & Gafni, 2000), generally prefer multiple-choice formats, or simple and de-contextualised questions, over essay type assessments or constructed-response types of questions (complex and authentic).

Traub and McRury (1990), for example, report that students have more positive attitudes towards multiple-choice tests in comparison to free response tests because they think that these tests are easier to prepare for, easier to take, and will bring in relatively higher scores. In the study by Ben-Chaim and Zoller (1997), the examination format preferences of secondary school students were assessed by a questionnaire (Type of Preferred Examinations questionnaire) and structured

interviews. Their findings suggest that students prefer written, unlimited time examinations and those in which the use of supporting material is permitted. Time limits are seen to be stressful and to result in agitation and pressure. Assessment formats which reduce stress will, accordingly, increase the chance of success. Students also vastly prefer examinations which emphasize understanding rather than rote learning.

Birenbaum (1994) introduced a questionnaire to determine students' assessment preferences (Assessment Preference Inventory) for various facets of assessment. This questionnaire was designed to measure three areas of assessment. The first is assessment-form related dimensions such as assessment type, item format/task type and pre assessment preparation. The second was examinee-related dimensions such as cognitive processes, students' role/responsibilities and conative aspects. The final area was a grading and reporting dimension. Using the questionnaire, Birenbaum (1997) found that differences in assessment preferences correlated with differences in learning strategies. Moreover, Birenbaum and Feldman (1998) discovered that students with a deep study approach tended to prefer essay type questions, while students with a surface study approach tended to prefer multiple-choice formats. In their study, questionnaires about attitudes towards multiple-choice and essay questions, about learning related characteristics, and measuring test anxiety, were administered to university students. Test anxiety is another variable that leads to certain attitudes towards assessment formats: students with high test anxiety have more favourable attitudes towards multiple-choice questions whilst those with low test anxiety tend to prefer open ended formats. Birenbaum and Feldman assumed that if students are provided with the type of assessment format they prefer, they will be motivated to perform at their best.

Scouller (1998) investigated the relationships between students' learning approaches, preferences, perceptions and performance outcomes in two assessment contexts: a multiple-choice question examination requiring knowledge across the whole course; and assignment essays requiring in-depth study of a limited area of knowledge. The results indicated that if students prefer essays this is more likely to result in positive outcomes in their essays than if they prefer multiple-choice question examinations.

Beller and Gafni (2000) gave an overview of several studies which analyzed the students' preferences for assessment formats, their scores on the different formats, and the influence of gender differences. In a range of studies they found some consistent conclusions suggesting that, if gender differences are found (and that is not always the case), female students prefer essay formats, and male students show a slight preference for multiple-choice formats (e.g. Gellman & Berkowitz, 1993). Furthermore, male students score better on multiple-choice questions than female students and female students score better than male students on open ended questions than on multiple-choice questions (e.g. Ben-Shakhar & Sinai, 1991).

From studies regarding students' assessment preferences, it seems that students prefer assessment formats which reduce stress and anxiety. It is assumed, despite the fact that there are no studies that directly analyze the preferences of students and their scores on different item or assessment formats, that students will perform better on their preferred assessment formats.

Perceptions of assessment

Perception is commonly defined as 'the act of perceiving; cognizance by the senses or intellect'. According to Brunswik (1956), an eminent expert in the field of cognitive psychology, perception "*emerges as that relatively primitive, partly autonomous, institutionalized, ratiomorphic subsystem of cognition which achieves prompt and richly detailed orientation habitually concerning the vitally relevant, mostly distal aspects of the environment on the basis of mutually vicarious, relatively restricted and stereotyped, insufficient evidence in uncertainty-geared interaction and compromise, seemingly following the highest probability for smallness of error at the expense of the highest frequency of precision*". Recently, several researchers have investigated the role of perceptions of assessment in learning processes.

For example, Scouller and Prosser (1994) investigated students' perceptions of a multiple-choice question examination, consisting mostly of reproduction-oriented questions, to investigate the students' abilities to recall information, their general orientation towards their studies and their study strategies. The students' perceptions do not always seem to be correct: on one hand they found that some students wrongly perceived the examination to be assessing higher order thinking skills. As a consequence, these students used deep study strategies to learn for their examination. On the other hand, the researchers concluded that students with a surface orientation may have an incorrect perception of the concept of understanding, cannot make a proper distinction between understanding and reproduction, and therefore have an incorrect perception of what is being assessed. Nevertheless, in this study no correlations were found between perceptions of the multiple-choice questions and the resulting grades. In the earlier mentioned study by Scouller (1998), relationships were found between students' preferences, perceptions and performance outcomes. Students who prefer multiple-choice question examinations perceive these assessments (actually assessing lower levels of cognitive processing) to be more likely to assess higher levels of cognitive processing than students who prefer essays. Poorer performance, either on the multiple-choice questions or on the essays, was related to the use of an unsuitable study approach due to an incorrect perception of the assessment. Better performance on the essays (actually assessing higher levels of cognitive processing) was positively related to a perception of essays as assessing higher levels of cognitive processing and to the use of a suitable study approach (i.e. deep approach).

In the above mentioned studies, the multiple-choice examinations intentionally assessed the lower levels, and the assignment essays the higher levels, of cognitive processing. It is not always evident that students perceive assessments in the ways that were intended by the staff. MacLellan (2001), for example, used a questionnaire asking students and teaching staff about the purposes of their assessment, the nature and demand level of the tasks which were assessed, the timing of the assessment and the procedures for marking and reporting. The results showed that there are differences in perceptions between students and staff of the use and purposes of the assessment and the cognitive level measured by the assessment. For example, the students perceived the reproduction of knowledge to

be more frequently assessed, and the application, analysis, synthesis and evaluation of knowledge less frequently assessed, than the staff believed.

How do perceptions and preferences relate to each other? According to Birenbaum and Rosenau (in press), students' perceptions of assessment refer to opinions, attitudes, and preferences towards the assessment and its properties. Struyven, Dochy and Janssens (2002) interpret perceptions as a constructivist act of creating meaning in which perceptions are seen as beliefs, opinions, interpretations, ideas, preferences, images and conceptions as a result of experience. In both cases, preferences are seen as a factor in determining the students' perceptions. Apart from the study by Scouller (1998), there has not been much research to investigate this relationship. In our study the questions are: which assessment and question types students prefer in a certain course in a CLE using rather traditional assessment types (multiple-choice questions and essays); how students perceive the assessment when the multiple-choice questions also assess higher levels of cognitive processing; and what effects these may have on the assessment outcomes.

Empirical study

The current study took place in a CLE and was structured as follows. For a period of 7 weeks, students worked on a specific course theme (i.e. 'legal acts'). During these 7 weeks the students worked twice a week for two hours in small groups (maximum 19 students) on different tasks, guided by a tutor. In conjunction with these tutorial groups, they were enrolled in somewhat bigger practical classes (38 students) for 2 hours a week and another 2 hours a week in large class lectures. Assessment took place immediately after the course, by means of a written exam (a combination of multiple-choice questions and essay questions). The CLE in this study is highly consistent with learning environments described under the label of problem-based learning (Gijbels, Dochy, Van den Bossche & Segers, 2005).

In order to gain more insight into the students' actual preferences and perceptions, the effects of these on students' performances on the different formats in the assessment, and the different cognitive process levels that were assessed, four research questions were formulated.

1. Which assessment preferences do students have in a CLE? In more detail: a) which assessment type is preferred; and b) assessments of which cognitive processes are preferred?
2. How did students actually perceive the 'traditional' assessment in the CLE (i.e. the cognitive processes assessed in the traditional assessment)?
3. In what ways are students' assessment preferences related to their assessment results? In more detail, what were: a) the relationships between assessment type preferences and the scores on the assessment format; and b) the relationships between cognitive process preferences and the scores on the different cognitive levels that were assessed?
4. In what way are students' perceptions of assessment related to their assessment results?

The research questions are visualised in Figure 3.1.

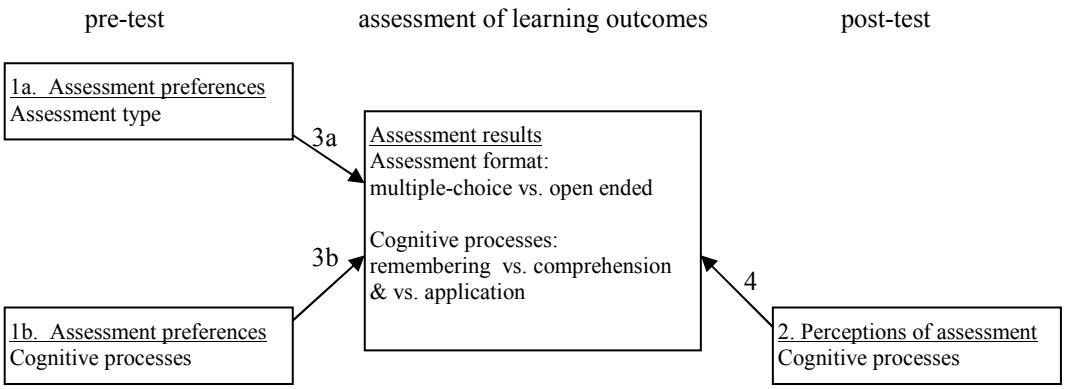

Figure 3.1. Visualised research questions of this study.

Method

Subjects

A total of 210 students, following a course on the theme 'legal acts' in the first year at a Dutch university, participated in the inventory of assessment preferences (pre-test). 392 students underwent the assessment of learning outcomes and 163 students participated in the inventory of perceptions of the questions after the assessment (post-test). In total, 83 students participated on all three occasions.

Procedures and instruments

Assessment preferences (pre-test) were measured by means of the Assessment Preferences Inventory (API; Birenbaum, 1994). This was originally a 67 item Likert-type questionnaire designed to measure seven dimensions of assessment. For our purposes we selected only two dimensions of the questionnaire using a 5 point Likert-scale and translated these items. Minor adjustments were also made to fit the translated questionnaire into the current educational and assessment environment. The used dimensions are: Assessment types (12 questions about students' preferences for different modes of oral, written and alternative tests) and Cognitive processes (15 questions about the preferences for assessing the cognitive processes remembering, understanding, applying, and analysing, evaluating and creating). Students were asked to complete the API questionnaire during one of the tutorial sessions near the end of a first year law course. Our translation of the questionnaire resulted in an acceptable Cronbach's alpha value for the whole questionnaire (Cronbach's alpha = .76). The reliability of the first part of the questionnaire (Assessment type) is moderate (Cronbach's alpha = .61) and that of second part (Cognitive process) is good (Cronbach's alpha = .82).

The traditional assessment was a combination of 6 open ended questions, requiring a variety of short and long answers, and 40 multiple-choice questions

(assessing learning outcomes). The objectives of the assessment were threefold and derived from Bloom's taxonomy (Bloom, 1956; Anderson & Krathwohl, 2001). The first objective was to investigate the student's ability to recall information (in terms of Bloom, knowledge; defined as the remembering of previous learned material). The second was to investigate understanding of basic concepts and principles (comprehension; defined as the ability to grasp the meaning of material). The third was to examine the application of information, concepts, and principles, in new situations (application; refers to the ability to use learned material in new and concrete situations). Cognitive processes such as analysing, evaluating and creating were also put into this last category, so the emphasis of this category is to investigate if students are able to use all their knowledge in concrete situations to solve an underlying problem in the presented cases or problem scenarios. 25% of the assessment consisted of reproduction/knowledge based questions (37.5% of all multiple-choice questions), 20% of comprehension based questions (30% of all multiple-choice questions), and 55% of application based questions (32.5% of all multiple-choice questions and all open ended questions).

The item construction and the assessment composition were carried out by the so-called assessment construction group. This group, consisting of four expert teachers on the subject and one assessment expert, worked according to the faculty assessment construction procedures, using several tools such as a planning meeting, a specification table, a number of item construction meetings, and a pool of reviewers, to compose a valid and reliable assessment. During the item construction meetings, different aspects of the items were discussed: the purpose of the question; the construction of the question (in case of a multiple-choice question the construction of the stem, the construction and usefulness of the correct choice and the distracters); and the objectivity of the right answer. A specification table was used to ensure the assessment was a sound reflection of the content (subject domain) and the cognitive level (cognitive process dimension) of the course. The difficulty of the items was also discussed. Classifications of the items were made during the meetings. After composing the assessment, it was send to four reviewers (in this case two tutors and two professors who were all responsible for some teaching in the course). The reviewers judged the assessment in terms of the usefulness of the questions and the difficulty of the assessment as a whole. The assessment was validated by the course supervisor. On average, students scored 34.8 out of 60 on the total assessment (SD = 9.75); the average score on the open ended part was 12 out of 20 (SD = 4.40); and the average score on the multiple-choice part was 22.8 out of 40 (SD = 6.16). The internal consistency reliability coefficient of the multiple-choice part was also measured and found to be appropriate (Cronbach's alpha = .80). For these students it was the fifth course in the curriculum and the fourth assessment of this kind.

To measure the students' perceptions of the assessment, the 15 questions from the dimension Cognitive process of the API were used (post-test). Basically, this questionnaire asked students which cognitive processes (remembering, understanding, applying, and analysing, evaluating and creating) they thought were assessed by the combination of open ended and multiple-choice questions they had just taken. Students were asked to complete the questionnaire directly after

finishing the assessment. This questionnaire also has an acceptable internal consistency reliability (Cronbach's alpha = .79).

Analysis

Results were analysed by means of descriptive statistics for the measures used in the study and multivariate analysis of variances (MANOVAs) were conducted to probe into the relationships between students' assessment preferences, perceptions and their study results (Green & Salkind, 2003). The research questions will be reported on one by one.

Results

Which assessment preferences do students have?

For the first research question, students were firstly asked about their preferences for assessment types and item format/task types in the current CLE (see Table 3.1).

Table 3.1. Descriptive statistics for students' preferences of assessment type

Rank	Assessment type	*M*	*SD*
	Oral tests	*2.47*	*.92*
12.	Individual oral tests, without supporting material (notes, books).	2.10	1.13
11.	Individual oral tests wherein the questions are given half an hour prior the test, and answers can be prepared without supporting materials.	2.26	1.17
7.	Individual oral tests wherein the questions are given half an hour prior the test, and answers can be prepared with supporting materials.	3.01	1.30
10.	Oral tests, in the form of a group discussion where the instructor observes and assesses the contribution of each of the participants.	2.48	1.36
	Written tests	*3.42*	*.73*
5.	Written tests without supporting materials.	3.39	1.15
6.	Written tests, without a time limit and without supporting material.	3.16	1.24
1.	Written test, with supporting materials.	3.73	1.08
3.	Written tests without a time limit, with supporting materials.	3.47	1.31
4.	Take-home exams.	3.41	1.32
	Alternative tests	*3.20*	*1.09*
2.	Papers/projects.	3.50	1.17
8.	Portfolio (your collected work, finished and in progress).	2.93	1.25
9.	Computerised tests.	2.72	1.22

Students preferred written tests, including take-home exams and papers, in which they are allowed to use supporting materials such as notes and books, as well as papers or projects. Oral tests and the other modes of alternative assessment mentioned in the questionnaire, i.e. computerised tests and portfolios, are not amongst the students' preferences.

Secondly, the students were asked about their preferences in relation to the cognitive processes which were to be assessed. According to the students' preferences, a mix of cognitive processes should be assessed, such as (in order of preference): reproducing; comprehending; problem solving; explaining; drawing conclusions; critical thinking; and applying. Evaluating others' solutions or opinions, scientific investigation, providing of examples and comparing different concepts, were not preferred (see Table 3.2).

Table 3.2. Descriptive statistics for students' preferences of cognitive processes to be measured in assessment

Rank	Cognitive process	*M*	*SD*
	Remember	*3.72*	*.77*
1.	Knowledge questions related to the reading of assignments.	3.84	.92
4.	Questions making an appeal to the reproduction of facts.	3.60	.92
	Understand (exemplifying, comparing, inferring)	*3.33*	*.61*
2.	Comprehension questions related to the material taught by the instructor.	3.76	.89
4.	Questions that require the drawing of conclusions.	3.60	.83
12.	Questions that require the providing of examples.	2.99	.96
13.	Questions that require comparing different concepts/ideas.	2.93	.90
	Apply (implementing), problem solving	*3.61*	*.72*
3.	Questions that require problem solving.	3.70	.85
8.	Questions requiring the application of material learnt during the course to new situations.	3.53	.94
	Analyse (organizing), evaluate (critiquing, checking), create (generate)	*3.10*	*.62*
4.	Questions that require a personal explanation or opinion.	3.60	1.03
7.	Questions that require critical thinking.	3.54	1.01
9.	Questions that require analysis and interpretation.	3.29	.87
10.	Questions that require an overall view of the relationships between all topics learnt.	3.14	.97
10.	Questions that require creativity and imagination.	3.14	1.04
14.	Questions that require scientific investigation.	2.70	1.01
15.	Questions in which you are asked to evaluate others' solutions or opinions.	2.27	1.00

How did students perceive the traditional assessment?

Concerning the second research question, how students actually perceived the assessment they took, students considered it to be primarily a measurement consisting of comprehension- and application-based questions that required the drawing of conclusions, problem solving, analysis, interpretation and critical thinking (see Table 3.3). The measurement was also considered, secondarily, as a measurement of reproduction based questions.

Additionally, correlations between student preferences (see Table 3.2) and perceptions of the assessment (see Table 3.3) turned out not to be significant, suggesting that there is a distinction between students' preferences and their perception of the assessment.

Table 3.3. Descriptive statistics for students' perceptions of cognitive processes measured by the assessment

Rank	Cognitive process	*M*	*SD*
	Remember	*3.37*	*.68*
6.	Knowledge questions related to the reading of assignments.	3.38	.81
8.	Questions making an appeal to the reproduction of facts.	3.36	.82
	Understand (exemplifying, comparing, inferring)	*3.26*	*.58*
1.	Comprehension questions related to the material taught by the instructor.	3.66	.82
3.	Questions that require the drawing of conclusions.	3.59	.80
14.	Questions that require the providing of examples.	2.81	.92
10.	Questions that require comparing different concepts/ideas.	3.00	.86
	Apply (implementing), problem solving	*3.59*	*.70*
3.	Questions that require problem solving.	3.59	.80
2.	Questions requiring the application of material learnt during the course to new situations.	3.60	.79
	Analyse (organizing), evaluate (critiquing, checking), create (generate)	*3.00*	*.56*
13.	Questions that require a personal explanation or opinion.	2.84	.97
6.	Questions that require critical thinking.	3.38	.84
5.	Questions that require analysis and interpretation.	3.49	.74
8.	Questions that require an overall view of the relationships between all topics learnt.	3.37	.91
11.	Questions that require creativity and imagination.	2.97	.94
12.	Questions that require scientific investigation.	2.95	.99
15.	Questions in which you are asked to evaluate others' solutions or opinions.	2.23	1.04

How are students' preferences related to assessment results?

The third research question concerned the relationships between students' assessment preferences and their assessment results. A multivariate analysis of variance (MANOVA) was conducted to evaluate this relationship between the preference for oral assessments and the two assessment scores, between the preference for written assessments and the assessment scores, and the preference for alternative assessments and the assessment scores. The independent variable, the preference for the assessment type (oral, written and alternative), included three levels: students who prefer that assessment type, students who are neutral about that assessment type, and students who do not prefer that assessment type. The two dependent variables were the scores in the multiple-choice questions and the scores in the open ended questions. Results only showed significant differences among the three levels of preferences for written assessments on the assessment scores, Wilks's $\Lambda = .95$, $F(4, 414) = 2.61$, $p < .05$, though the multivariate effect size η^2 based on Wilks's Λ was low, at .03, suggesting the relationship between the preferences and the assessment scores are weak. Table 3.4 contains the means and the standard deviations on the dependent variables for the three levels.

Analysis of variances (ANOVAs) on each dependent variable were conducted as follow-up tests to the MANOVA. Using the Bonferroni method, each ANOVA was tested at the .025 level. The ANOVA on the open ended questions scores was significant, $F(2, 208) = 5.25$, $p < .01$, $\eta^2 = .05$, while the ANOVA on the multiple-choice questions scores was not significant, $F(2, 208) = 2.31$, $p = .10$, $\eta^2 = .02$.

Table 3.4. Means and standard deviations on the dependent variables for the three levels of preference for written assessment

	Multiple-choice questions		Open ended questions	
Preference level	*M*	*SD*	*M*	*SD*
Not preferring (*N* = 17)	25.72	6.69	14.58	4.93
Neutral (*N* = 116)	25.57	4.94	14.12	4.15
Preferring (*N* = 77)	24.01	5.04	12.28	3.93

Post hoc analyses to the univariate ANOVA for the open ended questions scores consisted of conducting pairwise comparisons, each tested at the .008 level (.025 divided by 3), to find which level of preference for written assessments influences the outcome of the open ended questions most strongly. The students who preferred written assessments obtained lower marks on this part of the assessment when compared with those students who are neutral towards them.

Secondly, MANOVAs were conducted to evaluate the relationships between the students' preferences for cognitive processes measured by the assessment (the independent variables: remembering, understanding, applying, and analysing, evaluating and creating), and the scores on the different cognitive levels actually measured by the assessment of learning outcomes (the dependent variables: total scores on the reproduction based questions, comprehension based questions and application based questions). For this purpose, students were also divided into three groups for each cognitive process: students who prefer that cognitive process, students who are neutral to that cognitive process, and students who do not prefer that cognitive process. No significant differences were revealed by means of the MANOVAs, suggesting that no relationship exists between preferences for cognitive processes and the actual outcomes on the different cognitive levels measured with a combination of multiple-choice and open ended questions.

How are students' perceptions and assessment results related?

The fourth research question followed from the findings of Scouller's study (1998) that a mismatch in perception of an assessment will lead to poorer results on the assessment. In order to answer this question, it was assumed that students with a matching perception of the level of the cognitive processes measured by the assessment of learning outcomes (being a clear correspondence between the perceived levels of cognitive processes and the intended levels of cognitive processes by the assessment construction group), will have better results than students with a misperception of the cognitive processes (being a mismatch between the perceived level of cognitive processes and the intended levels of cognitive processes by the assessment construction group). To investigate this assumption, students were divided into three groups: a matching group, made up of students who perceived the assessment more as applying than remembering (*N* = 65; 40%); a mismatching group, consisting of students who perceived the assessment more as remembering than applying (*N* = 36; 22%); and a second mismatching group, of students who perceived the assessment as equally remembering and applying (*N* = 62; 38%) . The dependent variable was the outcome on the total assessment.

Though students with a matching perception scored slightly better on the assessment of outcomes compared to students with a misperception, the differences

were marginal and not significant. So students with a misconception of the level of cognitive processes assessed (i.e. those who perceived the assessment as being more remembering than applying or equally remembering and applying), did not perform significantly worse than students with a matching perception (i.e. those who correctly perceived the assessment to be more applying than remembering).

Discussion

This study was designed to examine which assessment formats students prefer in a CLE, and how students perceive more traditional assessment formats in the CLE used. In addition, the relationships between students' assessment preferences, perceptions of assessment and their assessment results were examined. Four research questions guided this study: 1. which assessment preferences do students have in a CLE?; 2. how did students actually perceive the 'traditional' assessment in the CLE?; 3. in what way are students' assessment preferences related to the assessment results?; and 4. in what way are students' perceptions of assessment related to their assessment results? Below we discuss the results of these questions one by one. Certainly statements about internal validity must be interpreted cautiously as a result of the fact that only 39.9% of the students following the course completed all three research instruments. In view of the specific context of the institution in which the study was conducted and the limited range of subject matters studied, caution must also be exercised where external validity is concerned. Nevertheless we feel that some interesting conclusions can be drawn from this study.

Which assessment preferences do students have in a CLE?

Students who were participating in the described CLE preferred traditional written assessment, as well as alternative assessment such as papers or projects. According to the students, the use of supporting material should be allowed and the questions or tasks should assess a mix of cognitive processes. The preference for assessment formats with the use of supporting material is in line with the studies of Ben-Chaim and Zoller (1997) and Traub and McRury (1990) in which students prefer easy-to-take and stress reducing assessment formats. Additionally, papers and projects can be considered as assessment formats in which, generally speaking, support (by means of materials and fellow students) is allowed. This could be one of the reasons why students prefer these assessment formats. Despite the fact that oral assessments, group discussions and peer evaluation play a more important role in a CLE, these formats or modes are not preferred by our students. This is possibly also because these assessment formats are yet not very common in the curriculum of the described CLE.

How did students actually perceive the 'traditional' assessment in the CLE?

The 'traditional' assessment, a combination of multiple-choice and essay questions in the CLE, was generally, as intended by the item constructors and the assessment composers, perceived more as assessing the application of knowledge,

problem solving, the drawing of conclusions, and analysing and interpreting, than assessing the reproduction of knowledge. Despite this, there was a clear correspondence between the intended level of cognitive processes and the perceived level of processes in the assessment in only 40% of the cases. In 38% of the cases, students made no distinction between a more reproduction based and a more application based assessment, and 22% of the students perceived the assessment to be more reproduction based. These figures show that it is possible for students to have a clear picture of the demands of assessments. On the other hand, the presence of multiple-choice questions in the assessment (and the apparently persistent perception of these questions created by previous experiences, that this format mainly assesses the reproduction of knowledge), could have caused the overestimating of the reproduction of knowledge. The standard interviews with students about the course and the assessment revealed some insights into how they perceived the assessment. For some students it was unimaginable to use cases or problem scenarios in a multiple-choice assessment. So, when preparing for the assessment, they did not prepare for applying knowledge, meaning that they did not practice their problem solving skills. Students also seem to identify certain questions as being reproduction based, because the questions or the cases used in the assessment resembled the tasks or cases used in the preceding educational programme. These results imply that a lot of students need help in building up a matching perception of what is assessed by means of the assessment formats that are used. Just giving help using examples of assessment items and discussing their answers seems not to be enough for these students. The purpose of the assessment questions and the cognitive processes to be used for answering the question correctly must also made clear and preferably be practiced. This is especially true in cases where a multiple-choice format is used to measure cognitive processes, rather than reproducing knowledge.

In what way are students' assessment preferences related to the assessment results?

Our findings regarding the relationships between assessment type preferences and the resulting scores on the assessment formats showed some significant differences. Strangely, students who preferred written assessments obtained lower marks on both parts of the assessment. Written assessments, especially the multiple-choice format, are often preferred because students think they reduce stress and test anxiety and are easy to prepare for and to take (Traub & McRury; 1990). So it is possible that some students prefer written assessment formats because they are used to it, but they are not good at them.

No significant relationships were found between students' preferences for the cognitive processes measured by the assessment and the scores on the different cognitive levels actually measured by the assessment of learning outcomes. As mentioned before, according to Birenbaum and Feldman (1998), students will be motivated to perform at their best if they are provided with the assessment format they prefer. Our outcomes do not support that finding. Preferences assume a choice. As in most cases, however, the students who were studied could not choose between assessment formats. They had to take the exam as it was presented. It is thus possible that their preferences only reflect what they think is a suitable

assessment format to measure their abilities and to give results which are fair enough.

In what way are students' perceptions of assessment related to their assessment results?

It has been argued by Scouller (1998) that a mismatch in the perception of assessment leads to poorer assessment results. The outcomes of that study were found to be strongly associated with students' general orientations towards study. Students with an orientation towards deep learning approaches will continue to rely on deep strategies when preparing for their examinations. As discussed by Scouller and Prosser (1994), students' perceptions, based on previous experiences with multiple-choice questions, lead to a strong association between multiple-choice examinations and the employment of surface learning approaches, leading to successful outcomes. In both studies (as in other studies of this kind), assessment formats were used in which the multiple-choice assessment was reproduction based and the essay assessment was more application based. In this study, we used an assessment format containing multiple-choice questions and essay questions with a heavy emphasis on application, using problem solving in both parts of the assessment. We found that students with matching perceptions scored slightly better on the total assessment, but not significantly so, compared to students with misperceptions. So we did not find any direct relationships between students' perceptions of assessment and their assessment results. It is possible that students' approaches to learning moderate the relationships between students' perceptions and their assessment results. From previous research, we know that a substantial proportion of the student population in the described CLE do not use appropriate approaches to learning (Gijbels, van de Watering, Dochy & Van den Bossche, 2005). We also now know from our study that a lot of students have a mismatching perception of the cognitive processes measured by the assessment. The relationships between study approaches and perceptions of the assessments used in this study should be further explored, in combination with the findings of Birenbaum and Feldman (1998) and Scouller (1998) about the relationships between study approaches and perceptions of assessment.

References

Anderson, L., & Krathwohl, D. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Beller, M., & Gafni, N. (2000). Can Item Format (Multiple Choice vs. Open-Ended) Account for Gender Differences in Mathematics Achievement? *Sex Roles: A Journal of Research*, 42, 1-21.

Ben-Chaim, D., & Zoller, U. (1997). Examination-type preferences of secondary school students and their teachers in the science disciplines. *Instructional Science*, *25* (5), 347-367.

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing. *Journal of Educational Measurement*, *28*, 23-35.

Biggs, J. (2003). *Teaching for quality learning at university* (2nd ed.). Buckingham: SRHE and Open University Press.

Birenbaum, M. (1994). Toward adaptive assessment - the student's angle, *Studies in Educational Evaluation*, *20*, 239-255.

Birenbaum, M. (1996). Assessment 2000: towards a pluralistic approach to assessment. In M. Birenbaum & F. Dochy (Eds.), *Alternatives in Assessment of Achievement, Learning Processes and Prior Knowledge* (pp. 3-30). Boston: Kluwer Academic.

Birenbaum, M. (1997). Assessment preferences and their relationship to learning strategies and orientations. *Higher education*, *33*, 71-84.

Birenbaum, M. (2000, September). *New Insights into Learning and Teaching and the Implications for Assessment.* Keynote address at the 2000 conference of the European Association of Research on Learning and Instruction Special Interest Group on Assessment and Evaluation, Maastricht, The Netherlands.

Birenbaum, M., & Feldman, R.A. (1998). Relationships between learning patterns and attitudes towards two assessment formats. *Educational Research*, *40* (1), 90-97.

Birenbaum, M., & Rosenau, S. (in press). Assessment Preferences, Learning Orientations, and Learning Strategies of Preservice and Inservice Teachers.

Bloom, B.S. (Ed.). (1956). *Taxonomy of educational objectives. Handbook I: The cognitive domain.* New York: McKay.

Boud, D. (1990). Assessment and the promotion of academic values. *Studies in Higher Education*, *15* (1), 101-111.

Broekkamp, H., van Hout-Wolters, B.H.A.M., van den Bergh, H., & Rijlaarsdam, G. (2004). Teachers' task demands, students' test expectation, and actual test content. *British Journal of Educational Psychology*, *74*, 205-220.

Brown, S., Rust, C., & Gibbs, G. (1994). *Strategies for diversifying assessment in higher education*. Oxford: The Oxford Centre for Staff and Learning Development, Oxford Brookes University.

Brunswik, E. (1956). *Perception & the representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California Press. Retrieved November 10, 2004 from http://www-2.cs.cmu.edu/~mws/definition.html

Dochy, F., & McDowell, L. (1997). Assessment as a tool for learning. *Studies in Educational Evaluation*, *23* (4), 279-298.

Gellman, E., & Berkowitz, M. (1993). Test-item type: what students prefer and why. *College Student Journal*, *27* (1), 17-26.

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 and Open University Press.

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.), *Optimising New Modes of Assessment: In Search of Qualities and Standards* (pp. 37-54). Dordrecht: Kluwer Academic Publishers.

Gijbels, D., Dochy, F., Van den Bossche, P., & Segers, M. (2005). Effects of problem based learning: A meta-analysis from the angle of assessment. *Review of Educational Research*, *75* (1), 27-61.

Gijbels, D., van de Watering, G., Dochy, F., & Van den Bossche, P. (2005). The relationship between students' approaches to learning and the assessment of learning outcomes. *European Journal of Psychology of Education*, *XX* (4), 327-341.

Glasner, A. (1999). Innovations in Student Assessment: a System-wide Perspective. In S. Brown & A. Glasner (Eds.), *Assessment Matters in Higher Education* (pp. 14-27). Buckingham: SRHE and Open University Press,.

Hofstee, W.K.B. (1999). *Principes van beoordeling: methodiek en ethiek van selectie, examinering en evaluatie*. Lisse: Swets and Zeitlinger.

Kuhlemeier, H., de Jonge, A., & Kremers, E. (2004). *Flexibilisering van centrale examens.* Cito: Arnhem.

Lawness, C.J., & Richardson, J.T.E. (2002). Approaches to studying and perceptions of academic quality in distance education. *Higher Education*, *44*, 257-282.

MacLellan, E. (2001). Assessment for learning: the differing perceptions of tutors and students. *Assessment & Evaluation in Higher Education*, *26* (4), 307-318.

Messick, S. (1994). The interplay of evidence and consequences in the validation performance assessments. *Educational Researcher*, *23* (2), 13-22.

National Research Council. (2001). *Knowing what students know: The science & design of educational assessment*. Committee on the foundation of Assessment. J. Pelligrino, N. Chudowski & R. Glaser (Eds). Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

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

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

Scouller, K.M., & Prosser, M. (1994). Students' experiences in studying for multiple choice question examinations. *Studies in Higher Education*, *19* (3), 267-279.

Segers, M. (1996). Assessment in a problem-based economics curriculum. In M. Birenbaum & F. Dochy (Eds.), *Alternatives in Assessment of Achievements, Learning Processes and Prior Learning* (pp. 201-226). Boston: Kluwer Academic Press.

Segers, M., & Dochy, F. (2001). New assessment forms in Problem- based Learning: the value- added of the students' perspective. *Studies in Higher Education*, *26* (3), 327-343.

Segers, M., Dochy, F., & Cascallar, E. (2003). The era of assessment engineering: changing perspectives on teaching and learning and the role of new modes of assessment. In M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimising New*

Modes of Assessment: In Search of Qualities and Standards (pp. 1-12). Dordrecht: Kluwer Academic Publishers.

Shepard, L.A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29* (7), 4-14.

Simons, R.J., van der Linden, J., & Duffy, T. (2000). New learning: Three ways to learn in a new balance. In R.J. Simons, J. van der Linden & T. Duffy (Eds.), *New Learning* (pp. 1-20). Dordrecht: Kluwer Academic Publishers.

Struyf, E., Vandenberghe, R., & Lens, W. (2001). The evaluation practice of teachers as a learning opportunity for students. *Studies in Educational Evaluation*, *27* (3), 215-238.

Struyven, K., Dochy, F., & Janssens, S. (2002, August). *Students' Perceptions about Assessment in Higher Education: a Review.* Paper presented at the Joint Northumbria/Earli SIG Assessment and Evaluation Conference: Learning communities and assessment cultures, University of Northumbria, Newcastle.

Traub, R. E., & MacRury, K. (1990). Multiple choice vs. free response in the testing of scholastic achievement. In K. Ingenkamp & R.S. Jager (Eds.), *Tests und Trends 8: Jahrbuch der Pädagogischen Diagnostik* (pp. 128-159). Weinheim und Basel: Beltz.

Webster's Dictionary (n.d.). Retrieved December 20, 2004, from http://www. wordiq.com /definition/Preference

Zeidner, M. (1987). Essay versus multiple choice type classroom exams: the students' perspective. *Journal of Educational Research*, *80* (6), 352-358.

Chapter 4

Teachers' and students' perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items[4]

Abstract

In today's higher education, high quality assessments play an important role. Little is known, however, about the degree to which assessments are correctly aimed at the students' levels of competence in relation to the defined learning goals. This article reviews previous research into teachers' and students' perceptions of item difficulty. It focuses on the item difficulty of assessments and students' and teachers' abilities to estimate item difficulty correctly. The review indicates that teachers tend to overestimate the difficulty of easy items and underestimate the difficulty of difficult items. Students seem to be better estimators of item difficulty. The accuracy of the estimates can be improved by: the information the estimators or teachers have about the target group and their earlier assessment results; defining the target group before the estimation process; by the possibility of having discussions about the defined target group students and their corresponding standards during the estimation process; and by the amount of training in item construction and estimating. In the subsequent study, the ability and accuracy of teachers and students to estimate the difficulty levels of assessment items was examined. In higher education, results show that teachers are able to estimate the difficulty levels correctly for only a small proportion of the assessment items. They overestimate the difficulty level of most of the assessment items. Students, on the other hand, underestimate their own performances. In addition, the relationships between the students' perceptions of the difficulty levels of the assessment items and their performances on the assessments were investigated. Results provide evidence that the students who performed best on the assessments underestimated their performances the most. Several explanations are discussed and suggestions for additional research are offered.

---

[4] Based on: Van de Watering, G., & van der Rijt, J. Teachers' and students' perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. *Accepted for publication in Educational Research Review*.

Introduction

During the last decade, higher education has been changing substantially. In today's higher education, new teaching methods are used to emphasise the importance of the learning process and the construction of knowledge. Assessment plays a major role in these innovative processes. Moreover, assessment is increasingly becoming central to the whole process of higher education in enhancing the quality of educational provision (Brown & Glasner, 1999). Biggs (2003) introduced the term 'constructive alignment', meaning that appropriate learning will be supported when all aspects of the teaching system - the objectives, the teaching method and the assessment - act in accord. Others, like Gibbs (1999) and Scouller (1998) acknowledge that assessment plays a crucial role in the learning process and on the impact of teaching methods, particularly new methods. Consequently, high quality assessment should assess the desired performance standards, academic standards or the students' levels of competence in such a way that students' knowledge, skills, abilities and aspects of professional expertise can be judged.

Recently there has been growing attention to academic standards related to the connection of students' entry-levels and their assessment outcomes. For example in several countries, such as the USA, there is an increasing interest in academic standards due to concerns about high dropout rates after the first year (Barefoot, 2004). In other countries, such as some European countries, more attention is paid to the selection of first year students to stimulate the study progress of students. In the UK, similar attention is paid to academic standards due to concerns about fairness of admission processes by changes in the discourses of the academic standards and there is discussion about the power the academic community has over the, sometimes undefined, standards (Leathwood, 2005; Stowell, 2004). Certainly, there is a strong need to define academic standards, and assessment reflecting these standards, in an appropriate way. Assessment, especially when used for selection purposes, has to be of high quality. Within the framework of high quality assessments, combined with attention to the sometimes implicit performance standards, conducting research on item difficulty and the perception of teachers and students of item difficulty is relevant. Little is known about the degree to which assessments in higher education are correctly aimed at the students' levels of competence. Therefore, the present study focuses on the item difficulty of assessments in higher education. The central question to be answered is whether teachers and students have correct perceptions of the item difficulty.

To illustrate the importance of item difficulty, we first introduce the research into assessment quality, standard setting and standard setting procedures. Then, in two subsequent sections, we review the main findings of previous research into teachers' perceptions of item difficulty, followed by a review of results of research into students' perceptions of item difficulty. Finally, in a third part, an empirical study is presented, investigating the accuracy of teachers' estimations and students' perceptions of difficulty.

PART I: THEORETICAL INTRODUCTION

Assessment quality and assessment difficulty

A high quality assessment should be a valid and reliable measurement. One important aspect of validity and reliability, which can affect the quality of the assessment and also can be linked to performance or academic standards, is the difficulty of the tasks or items in the assessment. When constructing the assessment, the entry levels of the students and the desired standards have to be borne in mind to control the difficulty of the items in the assessment. Item difficulty, indicated as item difficulty index or *p* value, can mathematically be defined as the proportion of assessees who answered the item correctly, and assessment difficulty can be defined as the average of the item difficulties or the ratio between the average score and the total assessment score (Allen & Yen, 1979; Mehrens & Lehmann, 1991). According to this definition the *p* value lies between 0 and 1 and it means that the higher the *p* value, the easier the item or assessment is. Depending on the purpose of the assessment, the constitution of the group and the item formats used in the assessment, the item difficulty or the assessment difficulty should lie between specific minimum and a maximum values.

The authors of the Standards for educational and psychological testing (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1985) define validity, denoted as construct validity, as "*a unitary concept, requiring multiple lines of evidence, to support the appropriateness, meaningfulness of the specific inferences made from test scores*" (p. 9). Terms such as appropriateness, adequacy, relevance, meaningfulness, utility of scores and interpretability are closely linked to the concept of validity. According to Messick (1989), validity always refers to "*the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores*" (p. 13). For an assessment of learning goals, the assessment is valid if it measures what it is intended to measure, namely to what extent the students have mastered that which is described in the learning goals. In an item construction and assessment composition process, it is important to determine the degree to which the items or tasks are relevant to the specific goals being assessed (Allen & Yen, 1979; Mehrens & Lehmann, 1991; Messick, 1989). A valid assessment reflects the material that is covered in the educational programme, taking the academic standard of the students into account. When the level of cognitive processes in assessment items does not correspond with the level of cognitive processes in the tasks of the educational programme, an assessment is not valid. For example an assessment containing too many difficult or too many easy questions does not correspond with the level of cognitive processes that should be assessed.

An assessment is not reliable when it contains too many items which a proportion of students cannot answer correctly. In classical test theory it is assumed that an observed score on an assessment is the sum of a true score component and a measurement error component (Feldt & Brennan, 1989). The measurement error is partly systematic and partly random and can be represented statistically. A method to determine the reliability of an assessment is Coefficient alpha, using variances of

item scores and variances of assessment scores. Stanley (1971) categorized sources of variance for scores on tests as possible sources of error. Among these sources are factors affecting performance, such as motivation and emotional strain, but also error due to chance elements of particular test items, such as guessing. The difficulty of an assessment, or several items in the assessment, can decrease the reliability of the assessment in two ways. Firstly, if the assessment is more difficult than expected by students, this can cause confusion, decrease of motivation, loss of concentration, uncertainty, anxiety, etcetera and as a consequence, in Stanley's terms, this means more error. Secondly, especially in multiple-choice assessment formats, there is the possibility of guessing. If the items are more difficult, this means more students will be guessing and this adds more random error to the variance of the assessment score (Bereby-Meijer, Meijer & Flascher, 2002).

Standard setting, standard setting procedures and item difficulty

In the construction of assessment items and the composition of an assessment, attention should be paid to the difficulty level. An assessment can contain some easy items as well as some difficult items, but the overall difficulty of the assessment should be adjusted to the level of the student population taking the assessment. To evaluate whether the difficulty of an item, or the whole assessment, is suited to the level of the students taking the assessment, a measure of the item difficulty is very useful.

At the start of the assessment construction process, item constructors and assessment composers have to make decisions about the required level of competence to meet the performance standard (Abbot, 2003). Ideally, there must be agreement and convergence about the expected students' performance standards in order to associate these with the difficulty level. Hence, in the assessment literature and research, different concepts are used such as 'average student' and the 'minimally competent candidate', also denoted as 'borderline students' (Chang, 1999; van de Watering & Claessens, 2003; Verhoeven, Verwijnen, Muijtjens, Scherpbier & van der Vleuten, 2002). Common definitions of these concepts are very useful because this seems to increase consensus in estimating item difficulty (Hurtz & Auerbach, 2003). In fact, if there is no definition already available then teachers, judges or experts should derive a definition themselves as a group in order to make sure that they are all using the same definition (e.g. Fehrmann, Woehr & Arthur, 1991). Most important in this is that judges are able to envisage the skills and competences, or to understand the cognitive levels, of their students and are able to communicate with them on the same level in order to make accurate estimations of the students' performances on the assessment items (Plake & Impara, 2001; Webb & Farivar, 1999).

The estimation of the assessment difficulty is often based on a hypothetical student who has studied a hypothetical number of hours. For example, in the case of teacher-made assessments, the item constructors and assessment composers usually keep an 'average student' in mind. An assessment should have a certain level of difficulty, enabling an average student who has done the amount of work required, to pass the assessment. Average students can be defined as those who are generally

sufficiently prepared for participation in the classes and have confidence in their own capacities to study for the assessment. It tends to be rather predictable for the teaching staff when, and about which topics, these students need further explanation (Van de Watering & Claessens, 2003). Item difficulty estimates for the average performance level of students can be determined by estimating the probability that an average student in a course is able to answer the assessment items correctly (Chang, 1999). In the determination of the difficulty level of assessment items, the 'minimally competent candidate' or the 'borderline examinee' is often kept in mind too. A minimally competent candidate would only just pass the assessment. According to Verhoeven et al. (2002), a borderline examinee can be described as a student who spends an average amount of time studying, whose knowledge is just sufficient to pass at the perceived level, but who frequently has difficulty in scoring above the cut-off score.

In standard setting procedures, the evaluation of the difficulty level of items is seen as a significant part of the determination of cut-off scores. To be able to determine standards, it is important to know the item characteristics that affect the difficulty levels of items (Goodwin, 1996). In different standard setting methods, for example in the Nedelsky method (Nedelsky, 1954), the Angoff method (Angoff, 1971) and the Ebel method (Ebel, 1972), judges have to indicate the difficulty level of test items whilst keeping a hypothetical student in mind. In the most often used method, that of Angoff, the 'minimally competent candidate' is kept in mind (Chinn & Hertz, 2002). Teachers have to signify the test items such a hypothetical student would answer correctly and incorrectly. The mean of the judges' estimated probabilities is the minimum passing score (Abbot, 2003).

PART II: LITERATURE REVIEW

Review of previous research into teachers' perceptions of item difficulty

The use of these different methods for standard setting illustrates the importance of the correct estimation of the difficulty level by teachers. Major decisions which will affect the students' futures are based on these predictions, (Irwin, Plake & Impara, 2000). Although the estimation of the item difficulty level is important in different standard setting methods and in the construction of assessments, there are only a few studies about item and test (assessment) difficulty. To find relevant studies a wide variety of computerised databases were utilised including Educational Resources Information Center (ERIC), ISI Web of Knowledge, Science Direct, Online Contents and Google Scholar. The following keywords were used: *difficulty level*, *assessment difficulty*, *item difficulty*, *performance standard*, *standard setting procedures*, *item construction*, *expectancy*, *accuracy*, *perceptions* and *higher education*. Regarding the context of our study only a few articles were found relevant. Next, the 'snowball method' was employed and the references in the selected articles for additional works were reviewed. Little is known about teachers' abilities to accurately estimate assessment and item difficulties during assessment and item construction processes. More research regarding the ability to estimate item difficulty has been done by researchers

investigating various aspects of the modified Angoff standard setting method because estimating the difficulty of items for a certain group of students is an explicit part of that method. Recently, the ability of teachers (in the Angoff method usually called judges or panellists) to make accurate item performance estimations has been questioned (Goodwin, 1999; Impara & Plake, 1998; Plake & Impara, 2001; Shepard, 1995). The task of conceptualizing minimally competent candidates and predicting their item performance was, according to Shepard (1995) and Plake and Impara (2001), very complex for the judges. According to the National Research Council (1999, p. 167 in Plake & Impara, 2001) it was "*difficult and confusing*" and, according to Berk (1996, p. 216, in Goodwin, 1999), this task requires judges "*to perform a nearly impossible cognitive task*".

Shepard (1995) examined the internal consistency of judges' ratings using the data from the National Academy of Education (NAE) Panel and demonstrated that judges underestimated the difficulty of hard items and the easiness of easy items for the minimally competent candidate. Shepard assumed that the judges were unable to hold the hypothetical minimally competent candidate in mind and therefore were unable to estimate the difficulty of assessment items accurately. Shepard's finding was supported by several studies, showing that judges' estimates about the difficult and easy items were inappropriate. The difficulty level of the difficult items was often underestimated and the difficulty of easy items was overestimated (e.g. Chang, 1999; Goodwin, 1999; Impara & Plake, 1998; Mattar, 2000).

Impara and Plake (1998) examined the ability of teachers to conceptualize minimally competent students or borderline students and the teachers' abilities to estimate item performance. In this study the teachers were familiar with their own students and the nature of the criterion variable. They asked teachers prior to an assessment to estimate the performance levels for two groups of students: students who were in their opinion 'borderline students' and the total group of students. The results showed that teachers were able to discriminate between minimally competent students and other students. Teachers were quite accurate in ranking items according to difficulty, but their estimation of the actual levels of the students' performances was not accurate. In line with Shepard's findings, teachers underestimated the performance of the borderline students (estimated mean of a total score of 50 was 13.01, actual mean was 22.52) but to some extent overestimated the performance of the total group of students (estimated mean was 36.34, actual mean was 32.69). The authors defined three levels of accuracy of the item performance estimates: estimates more than 10% (in terms of the *p* value .10) under the actual student performance (the actual *p* value) were classified as underestimates, meaning for these items the students' performances were underestimated by the teachers; accurate estimates were those within 10% (in terms of the *p* value .10) of the actual student performance and estimates more than 10% (in terms of the *p* value .10) over the actual student performance were considered to be overestimates, meaning for these items the students' performances were overestimated by the teachers. The smaller the mean difference, the higher the accuracy of the predicted assessment difficulty is. The mean absolute difference of the teachers' average differences between the predicted item difficulty and the actual item difficulty across 50 items was -.17 (or 17% under the actual student

performance) for the borderline group of examinees and .09 (or 9% over the actual student performance) for the total group of examinees. For the borderline group of students, the teachers estimated 22.6% accurately and 66% of the predictions were underestimates. For the total group 41% were considered accurate and 48% were overestimates.

Goodwin (1999) also compared the item estimates for minimally competent examinees with estimates made for the total group of examinees. After the assessment was administered, a group of judges discussed data about the recent group of examinees, followed by training including practicing the standard-setting steps, and rated the items of the assessment. The average difference between the judges' estimations (predicted *p* values) and the actual performance of the borderline group of examinees (actual *p* values) was .03 for the borderline group of examinees and .12 for the total group of examinees. The accuracy of the judges' estimates for the total group of examinees and for the borderline group of examinees was classified into the same three levels that Impara and Plake (1998) used in their study: underestimates, accurate estimates and overestimates. For the borderline group of students 61.4% of the predictions were accurate, 27.9% were overestimates and 10.7% were underestimates. For the total group 39.3% of the estimates were accurate and 60.7% were overestimates. These results are opposite to the findings of Impara and Plake (1998). According to Goodwin these results may suggest that judges, who are often experts in their field with a lot of experience and training, have too much knowledge or have too high expectations of the examinees. Although judges are closely associated with the educational programme, there remains a large difference in cognitive levels between judges and the students.

According to the above studies, it seems that it is difficult for judges to make accurate estimations. Several single studies and two reviews prove that discussion among experts and training does lead to more accurate judgements and increases consensus. The Hurtz and Auerbach (2003) meta-analysis also adds that this leads to higher cut-off scores.

In another study, Plake and Impara (2001) investigated the reliability and accuracy of item performance estimations. In their study all teachers received training and information about the group of candidates during a 2-day workshop. They practiced making item performance estimates and together elicited the knowledge, skills and abilities of the minimally competent candidate. To investigate whether the item performance estimates were consistent, they compared the estimates of panellists for two years. Results showed that these estimations were consistent and reliable across different panels. They also found that the judges' estimations of the item performance of the minimally competent candidate corresponded with the actual performances of the minimally competent candidates. The mean absolute value of this difference was .07 for each year. These results supported validity and differ from the findings in the study by Impara and Plake (1998). According to Plake and Impara (2001) these differences are due to the received information and group discussions about the minimally competent candidate, the received training and the opportunity to practice the item estimation process.

Bateman and Griffin (2003) also investigated the appropriateness of professional expertise in determining levels of performance. In their study the judges were experts in the subject matter and were the constructors of the items. The experts were carefully selected and they all received on-the-job training in item construction and the setting of standards. They found, by comparing the judges' estimates of difficulty to the outcomes of item response analysis, that the judges were able to make accurate judgments about the difficulty of the items.

Moreover, two review studies reveal interesting results. In a meta-analysis done by Hurtz and Auerbach (2003), 38 different studies from the past 30 years, using the Angoff method and with useful data, were reviewed. The impact of common procedural modifications on the level of the resulting cut-off scores and the degree of consensus amongst the judges was evaluated. The authors concluded that, when judges have the opportunity to discuss their estimates, the consensus increases, but this also tends to result in higher standards and cut-off scores. In addition, the highest standards were found, with the highest degree of consensus, in those studies in which judges used a common definition of the target group and later had the opportunity to discuss their estimates. On the other hand, they concluded that providing judges with normative data about the *p* values will lead to lower standards.

Finally, Brandon (2004) reviewed several empirical studies on topics related to the modified Angoff standard setting method. The topic of the validity of judges' estimates confirmed, on the whole, the findings of Shepard (1995): his findings about deviation of judges' estimates from empirical item *p* values (item estimate accuracy) showed that overall item estimation did not have the desired validity. Validity, meaning a small mean difference between the judges' estimates and the empirical item *p* values, is clearly supported in cases where judges are very well trained, information about the assessment data of the candidates is given to the judges, and standards are extensively discussed.

Review of previous research into students' perceptions of item difficulty

As well as teachers' perceptions of item difficulty, students' perceptions have been scrutinised in several studies. In the study of Chang (1999), three groups of master and doctoral education majors participated. These graduate students were divided into three groups of judges and asked to estimate the item difficulty of multiple-choice questions using the Angoff method and/or the Nedelsky method in a course they were taking. The main difference between the Nedelsky method and the Angoff method in this case was that in the Nedelsky method students additionally had to estimate whether an average student could eliminate the alternatives of the multiple-choice questions successfully. The mean absolute discrepancy between the judges' ratings and the actual performance levels (intra-judge inconsistency) for the Angoff and Nedelsky methods were, for the first group of judges, who used both methods, .25 and .13 respectively. For the second and third groups, half of the judges were assigned to the Angoff method and the other half to the Nedelsky method and matched by assessment scores. For two different assessments the mean intra-judge inconsistency were respectively .21 and .20 for

the Angoff method and .10 and .14 for the Nedelsky method. The Nedelsky method produces more accurate estimates, but the findings about both methods in Chang's study do not indicate a high degree of validity.

Research by Lee and Heyworth (2000) investigated the students' perceptions of the problem difficulty of mathematic test items, along with the teachers' abilities to estimate the problem difficulty. In their study, in order to develop a measure of difficulty, they identified different factors, called complexity factors, that affected the difficulty of items. Students indicated the difficulty of the problems in a mathematics test while solving the items. Teachers completed a questionnaire in which they had to indicate the difficulty of the problems, as well as identifying factors that could influence the estimation of the difficulty of the items. Among these complexity factors were the perceived number of difficult steps during the problem solving process, the number of steps required to finish the problem, and the students' degrees of familiarity with the question. The complexity factors were seen to be important in the prediction of the problem difficulty by the teachers. Results showed that the estimated problem complexity, the students' performances (item difficulty ratio), the teachers' estimations, and the students' perceptions of problem difficulty, were highly correlated. Despite the fact that teachers were asked to think more thoroughly about the difficulty level of the items by means of the questionnaire, students' predictions of their performance were more accurate than the teachers' estimations.

Verhoeven, Verwijnen, Muijtjens, Scherpbier and Van der Vleuten (2002) compared the item estimates and the cut-off scores set by item writers with the estimates and cut-off scores set by recently graduated students. The item writers and graduates were asked to estimate for each item the probability of a borderline examinee answering the item correctly. Results showed that the item estimates and cut-off scores set by recently graduated students were more credible than the standards set by the item writers. The authors suggested that the graduate students were more familiar with examinations which they had taken when they were students and were more common with the educational context in which the examinations were taken. There was a wider range between the highest and lowest estimate for the item writers than for the graduates. A possible reason is that the group of graduates was more homogeneous. The item writers differed in their backgrounds and levels of experience with students. Item writers were inclined to overestimate the performance of students. They might have based their judgments on "*what they think students ought to know*" (Verhoeven et al., p. 865).

Summary of the previously reviewed studies

In summary, although the outcomes of previous research into teachers' perceptions of item difficulty are not consistent, most research shows that estimating the difficulty of items or assessment standards is a difficult job. Teachers tend to overestimate the difficulty of easy items and underestimate the difficulty of difficult items. The accuracy of the estimates can be improved by the information the estimators (judges, teachers, assessment constructors) have about the target group (total group of students, borderline group students, average students) and

former assessment results, defining the target group before the estimation process, the possibility of having discussions amongst the estimators about the defined target group of students and its corresponding standards during the estimation process, the amount of training in item construction, and practice with estimating. However, students seem to be better estimators of item difficulty. Nevertheless, studies of students' perceptions are also inconclusive.

PART III: EMPIRICAL STUDY

Investigating the accuracy of teachers' estimations and students' perceptions of difficulty

To verify and further investigate the accuracy of teachers' estimations during assessment and item construction processes and students' perceptions of item and assessment difficulty an empirical study was conducted. The purposes of this third empirical part were (a) to examine the accuracy of the item constructors' and assessment composers' estimations of the difficulty levels of the assessment items and the assessment difficulty, (b) to assess to what extent the students' perceptions of the difficulty levels of the assessment items correspond to the statistical difficulty levels, (c) to examine to what extent the students' perceptions of the difficulty levels of the assessment items differ to the item constructors' and assessment composers' estimations of the item difficulties, and (d) to investigate the relationships between students' perceptions of the difficulty levels of the assessment items and their performances on the assessment.

Method and Instrumentation

Participants

Both teachers and students from a faculty of law at a Dutch university participated in this study. Firstly, the teachers participated as item constructors, assessment composers or assessment reviewers in a first year bachelor programme at a university. According to the assessment construction procedures, the item construction and the assessment composition took place in a so-called assessment construction group. This group consisted of at least three item constructors or assessment composers who were expert teachers in the subject matter of the course, one assessment expert and at least two assessment reviewers, who judged the test on the usefulness of the assessment items. Three assessment construction groups, consisting of six, seven and four members respectively, participated in this study.

Students enrolled in a first year bachelor programme also took part in this study. 223 students participated and completed assessment A and the questionnaire assessing the students' perceptions of the difficulty levels of the assessment items. Another group of 138 students completed assessment B and the questionnaire. A final group of 198 students completed assessment C and the questionnaire.

Finally, 30 students were randomly selected to participate in one of three focus group interviews.

Variables and procedure

The three assessments (tests A, B and C), which were constructed by assessment construction groups, were each composed of a set of 40 multiple-choice questions (four-choice items) and one open question. For assessments A and B the open question consisted of 2 case studies and for assessment C the open question consisted of 6 sub-questions. During three assessment construction meetings (of two hours) different aspects of the assessment items were discussed in order to compose a valid and reliable assessment. The meetings discussed the purposes of the questions, the construction of the questions (in case of a multiple-choice question the construction of the stem, the construction and usefulness of the correct choice and the distracters), and the objectivity of the right answer. A table of specification was used to ensure the assessment was a sound reflection of the content (subject domain) and the cognitive level (cognitive process dimension) of the course. The primary purpose of assessments A and B was to apply information, concepts and principles in new situations (application, in terms of Bloom (1956); refers to the ability to use learned material in new and concrete situations). The principal purposes of assessment C were to investigate the students' abilities to recall information (knowledge; defined as the remembering of previous learned material) and to understand basic concepts and principles (comprehension; defined as the ability to grasp the meaning of material). The reliability of the assessments (Cronbach's alpha coefficient) varied from .70 (assessment A) to .82 (assessment C).

To assess the teachers' perceptions of the difficulty level of the assessment items for students, a questionnaire was developed. Every item constructor or assessment composer was asked to rate each assessment item by means of a 3-point rating scale ranging from (1) *difficult*, (2) *not difficult and not easy*, to (3) *easy.* During the assessment construction meetings, where they discussed the constructed assessment items, they individually estimated the difficulty levels of the assessment items. The assessment reviewers estimated the difficulty levels of the items during a review of the composed assessment.

To assess the students' perceptions of the difficulty levels of the assessment items, a student questionnaire was developed. Students were asked to indicate how difficult they thought every item of the assessments (assessment A, B or C) was on a 3-point rating scale ranging from (1) *difficult*, (2) *not difficult and not easy*, to (3) *easy*.

The actual difficulty levels of the assessment items were indicated by the *p* values of each item. The *p* value can be described as the proportion of correct answers to an assessment item. The assessment performance of students was determined by the total scores on the assessment (assessment A, B or C). The total score on an assessment consisted of the score on the 40 multiple-choice questions (maximum score of 40 points) added to the score on the open question (maximum score of 20 points).

Focus group interviews with three randomly selected groups of students (N = 10) were held to in order to gain an insight into the students' perceptions of the assessment difficulty. Two researchers and one assessment composer interviewed the students. All of the interviews were audio taped and transcribed for further analysis.

Analysis

To examine the accuracy of the teachers' estimations of the difficulty levels of the assessment items, the teachers' ratings were compared with the actual p values of each assessment item. A similar comparison was made between the students' perceptions of the item difficulties and the actual item difficulty levels.

Rating scale analyses were conducted to map the differences between the item constructors and assessment composers' estimations, the students' perceptions of the difficulty level of the assessments and the statistical difficulty levels of the items.

Results

Firstly, the research questions will be reported on individually by means of quantitative data. Finally, the results of the qualitative data will be discussed to gain more insight into the most important concerns in students' perceptions of item difficulty.

The accuracy of teachers' estimations

To answer the first research question, to what extent item constructors, assessment composers and assessment reviewers are accurately able to estimate the difficulty levels of the assessment items, the three levels of accuracy as defined by Impara and Plake (1998) were used. The number of overestimated, accurately estimated, and underestimated items, as well as the mean differences between the estimated difficulty levels and the p values, are presented in Table 4.1.

For assessment A, 26.2% of the items were estimated accurately, 31.0% were underestimates and 42.9% were overestimates. For assessment B, 26.2% were also estimated accurately, 23.8% were underestimates and 50.0% were overestimates. For assessment C, 32.6% were estimated accurately, 30.4% were underestimates and 37.0% were overestimates. For all three assessments most items were overestimates, meaning for most items the students' performances were overestimated by the teachers. Or in terms of difficulty: the items were more difficult for students than perceived by the teachers. Though, the mean difference between the estimated difficulty level and the actual p value suggests that the estimation of the difficulty of the complete assessment is appropriate (values lie between -.10 and .10). The closer the values lie to 0, the more accurate the estimation is.

Table 4.1. The accuracy of teachers' estimates for assessments A, B and C.

	Assessment A			Assessment B			Assessment C		
	N items	% items	Mean difference	N items	% items	Mean difference	N items	% items	Mean difference
Underestimated	13 (1 case)	31.0	-.27	10 (1 case)	23.8	-.28	14 (1 open)	30.4	-.24
Accurate	11 (1 case)	26.2	-.02	11 (1 case)	26.2	.01	15	32.6	.02
Overestimated	18	42.9	.33	21	50.0	.26	17 (5 open)	37.0	.31
Total	42 (2 cases)	100.0	.05	42 (2 cases)	100.0	.06	46 (6 open)	100.0	.05

Table 4.2. The accuracy of students' estimates for assessments A, B and C.

	Assessment A			Assessment B			Assessment C		
	N items	% items	Mean difference	N items	% items	Mean difference	N items	% items	Mean difference
Underestimated	25 (1 case)	59.5	-.22	29 (1 case)	69.0	-.23	28 (2 open)	60.9	-.25
Accurate	12 (1 case)	28.6	.01	12 (1 case)	28.6	-.03	12 (2 open)	26.1	-.01
Overestimated	5	11.9	.17	1	2.4	.14	6 (2 open)	13.0	.17
Total	42 (2 cases)	100.0	-.11	42 (2 cases)	100.0	-.17	46 (6 open)	100.0	-.14

The accuracy of students' estimations

For the second research question, assessing to what extent the students' perceptions of the difficulty level of the assessment items correspond to the actual difficulty levels, the number of overestimated, accurately estimated and underestimated items, as well as the mean differences between the estimated difficulty levels and the actual *p* values, were also calculated and are presented in Table 4.2.

For all three assessments, most perceptions of the items were underestimates (assessment A: 59.5%; assessment B: 69.0%, and assessment C: 60.9%) meaning for most items the performances were underestimated by the students themselves, leading to a mean difference between the estimated difficulty levels and the actual *p* values lower than -.10 (or more than 10% under the actual student performance) for all three assessments. In terms of difficulty this means, that students perceived the items as more difficult than the items were actually.

Differences between students' and teachers' estimations

For the third research question, to assess to what extent the students' perceptions of the difficulty levels of the assessment items differ to the teachers' estimations of the item difficulties, the results of the first two research questions were combined and compared to each other. The results are presented in Figures 4.1 to 4.3. The figures show that teachers gave higher ratings for nearly every item of assessments A, B and C compared to the students. In their eyes, the assessment items were less difficult than perceived by the students. In particular, for the most difficult items (items with a *p* value lower than .40) of the assessments the items were more difficult for the students than estimated by the teachers (see Figures 4.1, 4.2 and 4.3). For the students these items were also more difficult than perceived by themselves. In general, however, students gave lower ratings to the items than the teachers; they perceived most of the items in the assessments as more difficult. In addition, the figures show a clear relationship between the actual difficulties of the items and the differences between the students' perceptions of the item difficulties and the actual *p* values: the more easy the items, the greater the difference. Clearly, the easiest items were less difficult than perceived by the students in all three assessments.

The relationship between students' estimates and assessment performance

For the fourth research question, to investigate the relationship between students' perceptions of the difficulty levels of the assessment items and their performances on the assessments, students were first divided into 5 groups, reflecting the rank order scores. Group 1 consisted of the 20% lowest scoring students and group 5 was the 20% highest scoring students. Secondly, for each group and each assessment, the mean differences between the students' perceptions of the item difficulties and the actual item difficulties, as well as the pass rate of these groups, were calculated. The results are presented in Table 4.3.

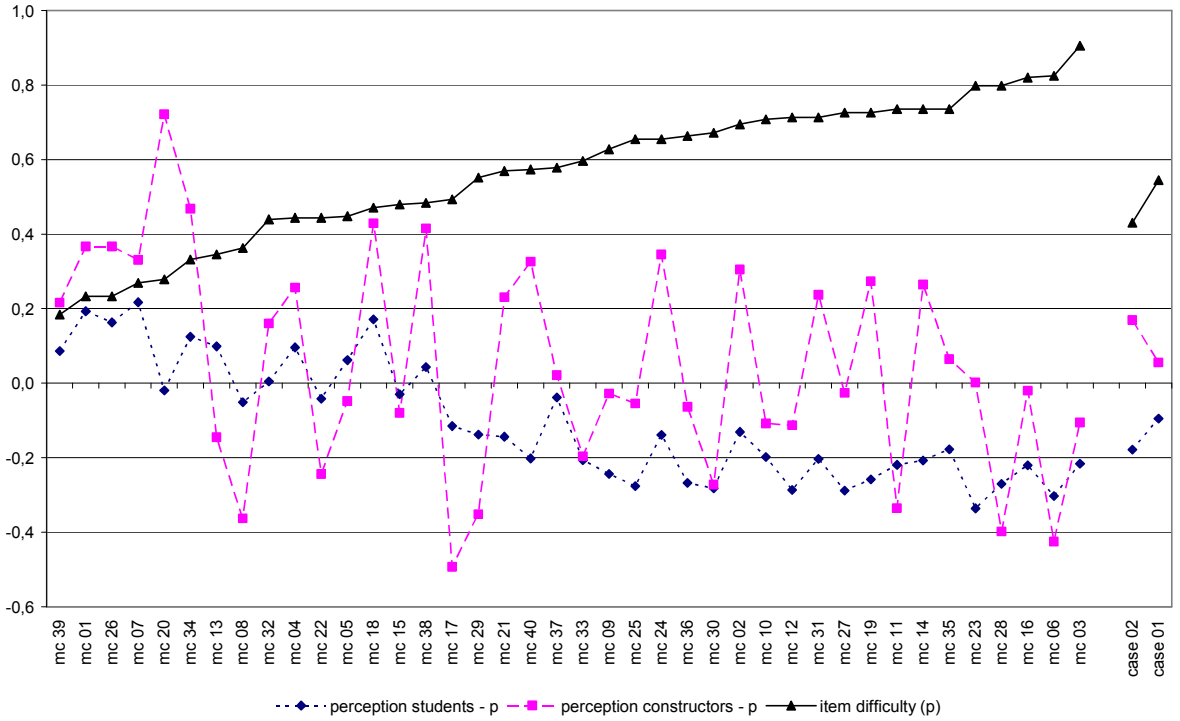

Figure 4.1. Differences in perceptions of item difficulties in order of the actual *p* value for assessment A

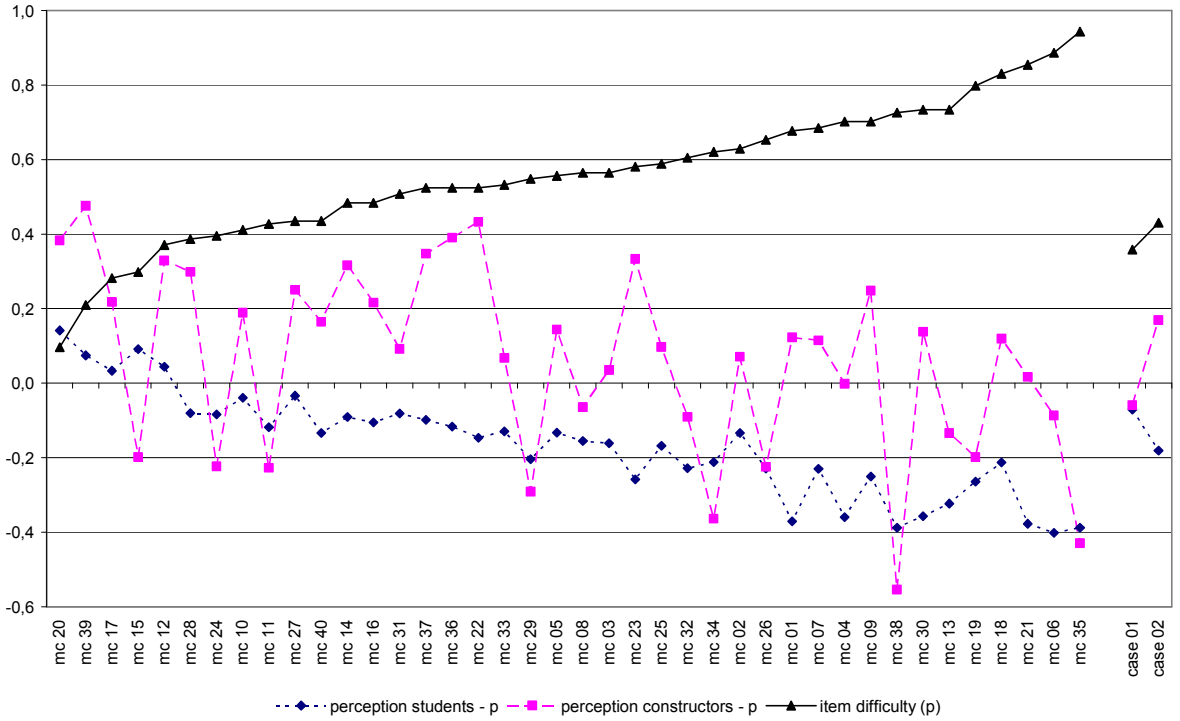

Figure 4.2. Differences in perceptions of item difficulties in order of the actual *p* value for assessment B

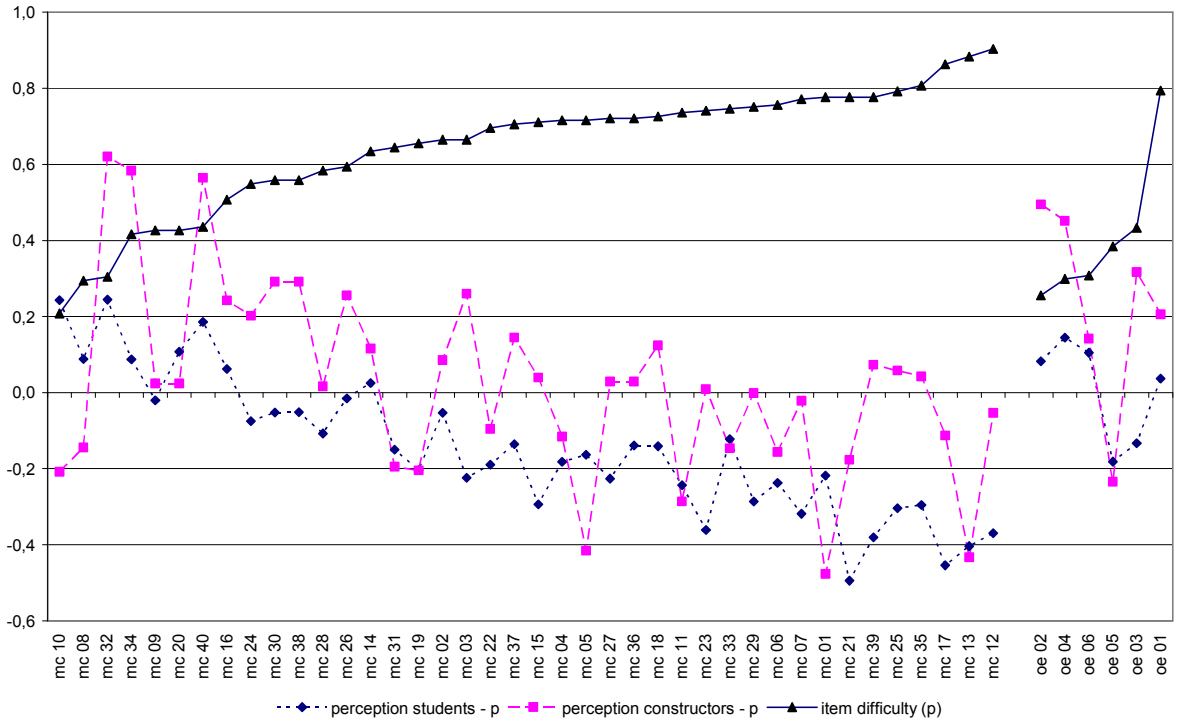

Figure 4.3. Differences in perceptions of item difficulties in order of the actual *p* value for assessment C

The results presented in Table 4.3 show that the highest scoring students (score groups 4 and 5) on assessments A and B underestimated their performance on the assessments the most. The mean difference between the students' perceptions of the item difficulties and the actual item difficulties is the highest for these two groups. In the case of assessment C, it was particularly the average scoring students (score groups 3 and 4) who underestimated their own performances. For all three assessments, the lowest scoring students (score groups 1 and 2) were able to estimate their performances quite accurately.

The teachers' estimations of the assessment difficulty for assessments A and B were .62 and .74 for assessment C. For all three assessments these estimations are close to the average actual *p* value of score group 4.

Focus group interviews

Three focus group interviews were held to gain more insight into the students' perceptions of the item difficulties. The students reported finding an assessment item more difficult when the subject matter was difficult or the item required them to make a transfer from one subject area to another. Assessment items which were detailed or less concrete, and where a choice had to be made between two plausible alternatives, were also perceived as difficult.

According to the students, the presentation of an assessment item was important too. When they had to think about the phrasing of the question, the item contained multiple propositions or contained a case study, the assessment item was perceived to be more difficult. Assessment items which contained brief and to the point questions and alternatives were seen as relatively easy. The interviews also showed a considerable influence of the educational programme on the learning behaviour of the students. Subject matters that were not discussed or practiced in the educational programme were seen as less important and therefore students did not give much attention to these topics while studying for the assessment. These items were automatically rated as difficult by the students. Furthermore, some students actually did not expect to find a case study in a multiple-choice question and as a consequence they did not prepare themselves properly to solve these questions. These questions were also seen as more difficult by these students.

In addition, some students reported that they tried to compare the assessment items with questions discussed in the educational programme or with questions from old assessments or exercise material. Assessment items containing a case study were mostly solved by recognition, instead of by problem solving strategies. Students did not retrieve the facts from the case study they were given, but tried to find similarities (e.g. "Does this case study look similar to that case study?"; "Which alternative looks similar to the solution at that time?"). Students did not read those assessment items carefully, but turned quickly to recognition without being critical regarding the alternatives and possibly misperceived the difficulty of the question. The students felt they had to use this kind of strategy because they did not have enough time.

Table 4.3. The accuracy of students' estimates for assessments A, B and C related to their performances by means of score groups

	Assessment A				Assessment B				Assessment C			
	students' estimation	actual *p* value	Difference*	% passed	students' estimation	actual *p* value	Difference*	% passed	students' estimation	actual *p* value	Difference*	% passed
Score group 1 (lowest)	.46	.40	.06	0.0	.39	.40	-.01	0.0	.43	.44	-.01	0.0
Score group 2	.46	.50	-.04	0.0	.38	.48	-.10	0.0	.52	.59	-.06	0.0
Score group 3	.47	.57	-.10	0.0	.43	.57	-.14	0.0	.51	.69	-.18	1.5
Score group 4	.39	.62	-.23	85.3	.41	.63	-.22	76.0	.59	.76	-.12	100.0
Score group 5 (highest)	.50	.73	-.23	100.0	.41	.71	-.29	100.0	.61	.89	-.10	100.0
Total	.46	.57	-.11	36.4	.39	.55	-.17	34.7	.54	.68	-.14	40.1

*Difference = students' estimation - actual *p* value

The mean of the teachers' estimates of the item difficulty level for assessments A and B is .62, and for assessment C is .74

Conclusion and discussion

This third part of our contribution examined the item difficulty of teacher-made assessments in higher education. In particular, we examined whether teachers and students have an appropriate perception of the difficulties of items.

The implications of teachers' estimations of difficulty

Our findings regarding the teachers' perceptions of the difficulties of items showed that the teachers' estimations of the difficulty of the complete assessments are appropriate. Item constructors, assessment composers and reviewers were, however, able to estimate the difficulty level correctly for only a small proportion of the assessment items. For the three assessments, about one third of the items was estimated accurately (defined as within .10 of the actual item difficulty). Hence, one could raise questions about the reliability and validity of the use of performance standards and standard setting procedures.

For most items the students' performances were overestimated by the teachers; this means that most items were more difficult for students than expected by the teachers. This finding is in line with the studies of Shepard (1995), Impara and Plake (1998) and Goodwin (1999) in which the competence of teachers to estimate the item difficulty level of the students was questioned. Judges in these studies, as in our study, tended to overestimate the performances on assessment for the total group of students. A possible explanation for this overestimation is the expertise of the teachers. According to Goodwin (1999), judges are typically experts in their fields. Because they might know too much, they cannot put themselves in the place of students adequately. Also, their expectations of the examinees are possibly too high. They may have difficulty in differing between the proportion of examinees who should have answered an item correctly and who could have answered an item correctly. The teachers in our study gave the same explanation. Their expectations were probably too high and, as a consequence, it was difficult to envision the skills and competences of the students. Another explanation is that teachers did not expect the educational programme to have a great influence on the learning behaviour of the students, such as that reported by the students in the focus group interviews.

In the studies of Mattar (2000), Plake and Impara (2001) and Bateman and Griffin (2003) teachers were able to estimate item difficulties accurately. In contrast with the teachers participating in our study, these teachers had all received training. In our study, there was no exchange of ideas between the item constructors, assessment composers and reviewers. In several studies the importance of training and practice is underlined (e.g. Mills, Melican & Ahluwalia, 1991; Plake & Impara, 2001). More discussion and training could have contributed to more accurate estimations of the item difficulties.

On the other hand, the findings can be interpreted as follows: it seems that the item constructors and assessment composers focussed on the high scoring students during the construction process of the assessment. This could explain the low degree of accuracy of teachers in estimating the difficulty levels of assessment items too. It implies that, in the case of teacher-made assessments in higher

education, to prevent disappointing student outcomes, item constructors and assessment composers together have to envision and describe the average student in terms of expectations first. They should then focus on the assessment construction process, keeping this average student in mind. Only in this way higher education can define the academic standards and ensure all participants these standards are assessed appropriately.

The implication of students' estimations

We also investigated whether students had correct perceptions of the difficulty levels of the assessment items. The number of items which were estimated accurately by the students did not differ much from the estimations of the item constructors, assessment composers and reviewers. However, in contrast to the teachers, students underestimated their own performances. This result is in line with earlier research (Dochy, 1992). Strangely, they mainly underestimated their own performances on the easiest items. Certainly, students may have different perceptions of the concept of item difficulties than those of teachers. In the interviews, students reported that the presentation of an assessment item was important in their estimation of its difficulty level. Teachers, on the other hand, did not see that as an important factor in their determination of the difficulty level of an assessment item. Therefore item constructors have to pay more attention to the presentation of the assessment items. It is also possible that students hoped that, if they indicated the assessment items as difficult ones, teachers would grade the assessment less severely. Or, are students simply unassuming about their own abilities?

Finally we examined the relationships between the students' perceptions of the difficulty levels of the assessment items and their performances on the assessments. Students tended to underestimate their performances. The students who performed well on the assessments underestimated their performances the most. This finding is in line with earlier research which stated that the more prior knowledge a student has, the more he tends to underestimate his performance (Dochy, Segers, & Buehl, 1999).

In assessments A and B, it was the highest scoring students who particularly underestimated their performances. In assessment C, it was mainly the average scoring students who underestimated their performances. Perhaps this difference can be explained by the different purposes of the three assessments. The main purpose of assessments A and B was to apply information, concepts, and principles in new situations, whereas assessment C focussed on the student's ability to recall information and understand basic concepts and principles. In the teachers' perceptions, assessment C was less difficult than assessments A and B. Future research might be interesting in order to support this finding in the different settings.

Additionally, the focus group interviews indicated a possible misperception of the item difficulty caused by a misperception of the cognitive processes students had to use in solving the items: some students tried to solve the items using incorrect strategies. For these students more practice in answering different

assessment items is needed. Also feedback from teachers or fellow students on the given answers is necessary in order to create a correct perception of the (difficulty of the) assessment items. The influences of the perception of the assessment, and the learning strategies that were used, on the students' perceptions of the item difficulties should also be subject of research.

References

Abbot, M. (2003, May). *Standard setting for complex performance assessments: A critical examination of the analytic judgment method*. Paper presented at the Annual Congress of the Canadian Society for the Study of Education, Halifax, Nova Scotia.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1985). *Standards for educational and psychological testing*. Washington DC: American Psychological Association.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey: Brooks/Cole.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (pp. 508-600). Washington: American Council on Education.

Barefoot, B. (2004). Foundations of Excellence: A New Model for First-Year Assessment, *Assessment update*, *16* (2), 5-7.

Bateman, A., & Griffin, P. (2003, November - December). *The appropriateness of professional judgement to determine performance rubrics in a graded competency based assessment framework*. Paper presented at the AARE/NZARE Conference, Auckland, New Zealand. Retrieved February 2, 2005 from http://www.aare.edu.au/03pap/

Bereby-Meijer, Y., Meijer, J., & Flascher, O.M. (2002). Prospect Theory Analysis of Guessing in Multiple Choice Tests. *Journal of Behavioral Decision Making*, *15*, 313-327.

Biggs, J. (2003). *Teaching for Quality Learning at University* (2nd ed.). Buckingham: SRHE and Open University Press.

Bloom, B.S. (Ed.). (1956). *Taxonomy of educational objectives. Handbook I: The cognitive domain.* New York: McKay.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, *17* (1), 59-88.

Brown, S., & Glasner, A. (1999). *Assessment Matters in Higher Education*. Buckingham: SRHE and Open University Press.

Chang, L. (1999). Judgmental item analysis of the Nedelsky and Angoff standard-setting methods. *Applied Measurement in Education*, *12* (2), 151-165.

Chinn, R. N., & Hertz, N. R. (2002). Alternative approaches to standard setting for licensing and certification examinations. *Applied Measurement in Education*, *15* (1), 1-14.

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

Dochy, F., Segers, M., & Buehl, M. (1999). The Relation Between Assessment Practices and Outcomes of Studies: The Case of Research on Prior Knowledge. *Review of Educational Research*, *69* (2), 147-188.

Ebel, R. L. (1972). *Essentials of educational measurement*. Englewood Cliffs: Prentice-Hall.

Fehrmann, M. L., Woehr, D. J., & Arthur, W.(1991). The Angoff cutoff score method: The impact of frame-of-reference rater training. *Educational and Psychological Measurement*, *51*, 857-872.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R.L. Linn (Ed.), *Educational Measurement* (pp. 105-146). Washington DC: American Council on Education.

Gibbs, G. (1999). Using Assessment Strategically to Change the Way Students Learn. In S. Brown & A. Glasner (Eds.), *Assessment matters in higher education* (pp. 41-53). Buckingham: SRHE and Open University Press.

Goodwin, L. D. (1996). Focus on quantitative methods: Determining cut-off scores. *Research in Nursing & Health*, *19*, 249-256.

Goodwin, L. D. (1999). Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline examinees. *Applied Measurement in Education*, *12* (1), 13-28.

Hurtz, G. M., & Auerbach, M. A. (2003). A Meta-Analysis of the Effects of Modifications to the Angoff Method on Cutoff Scores and Judgment Consensus. *Educational and Psychological Measurement*, *63* (4), 584-601.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, *35*, 69-81.

Irwin, P. M., Plake, B. S., & Impara, J. C. (2000). *Validity of item performance estimates from an Angoff standard setting study*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. Retrieved February 11, 2005, from http://www.unl.edu/BIACO/AERA/

Leathwood, C. (2005). Assessment policy and practice in higher education: purpose, standards and equity. *Assessment & Evaluation in Higher Education*, *30* (3), 307–324.

Lee, F. L., & Heyworth, R. M. (2000). Problem complexity: A measure of problem difficulty in algebra for use in cai systems. *Education Journal*, *28* (1). Retrieved December 15, 2004 from http://www.fed.cuhk.edu.hk/fllee/Papers/JournPa/

Mattar, J. D. (2000). Investigation of the validity of the Angoff standard setting procedure for multiple-choice items. Unpublished doctoral dissertation. Amherst: University of Massachusetts.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology*. Fort Worth: Holt, Rinehart & Winston.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (pp.13-103). New York: Macmillan.

Mills, C.N., Melican, G.J., & Ahluwalia, N.T. (1991). Defining minimal competence. *Educational Measurement: Issues and Practice*, *10* (2), 7-10.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, *14*, 3-19.

Plake, B. S., & Impara, J. C. (2001). Ability of panellists to estimate item performance for a target group of candidates: An issue in judgemental standard setting. *Educational assessment*, *7* (2), 87-97.

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

Shepard, L. A. (1995). *Implications for standard setting of the National Academy of Education Evaluation of National Assessment of Educational Progress Achievement Levels, Proceedings from the Joint Conference on Standard Setting for Large-Scale Assessments*. Washington: National Assessment Governing Board and National Center for Education Statistics.

Stanley, J. C. (1971). Reliability. In R.L. Thorndike (Ed.), *Educational Measurement* (pp. 356-442) Washington DC: American Council on Education.

Stowell, M. (2004). Equity, justice and standards: assessment decision making in higher education, *Assessment & Evaluation in Higher Education*, *29* (4), 495-510.

Verhoeven, B. H., Verwijnen, G. M., Muijtjens, A. M. M., Scherpbier, A. J. J. A., & Van der Vleuten, C. P. M. (2002). Panel expertise for an Angoff standard setting procedure in progress testing: Item writers compared to recently graduated students. *Medical Education*, *36*, 860-867.

Watering, G. van de, & Claessens, S. (2003, August). *The discrepancy between tutors' perceptions of student performance and achievement results*. Paper presented at the 10th European Association for Research on Learning and Instruction conference, Padova, Italy.

Webb, N. M., & Farivar, S. (1999). Developing productive group interaction in middle school Mathematics. *American Educational Research Journal*, *31* (2), 369-395.

Chapter 5

The discrepancy between teachers' perceptions of students' performances and students' actual achievements[5]

Abstract

The purpose of this research was to investigate to what extent an assessment, which is a combination of essay questions and multiple-choice questions, discriminates between less competent and more competent students, based on the teachers' perceptions of the students' competence levels. During the study, teachers observed their students during a nine week course at the end of a first-year problem-based learning setting and classified the students at the end of the course into four distinct groups perceived as: (1) barely competent; (2) moderately competent; (3) highly competent; and (4) moderately or highly competent students with high anxiety. Results show that the multiple-choice questions and the essay questions show appropriate differences between these groups of students. The performances on the multiple-choice questions and the essay questions show the same trend, i.e. low performance for the barely competent students, moderate performance for the moderately competent and anxious students and high performance for the highly competent students. To gain an insight into the stability of the students' competency over time, the assessment outcomes from students' previous courses were investigated. The results show some stability; however the anxious students performed significantly better on the essay questions on the assessment of their first course when compared to the following assessments. In addition, the highly competent students did not distinguish themselves from the other students on the essay assessment of their first course, although they did this in the following assessments. In this contribution, further explanations for these findings are discussed.

---

[5] Based on: Van de Watering, G,. Claessens, S., Dochy, F., & van der Rijt, J. The discrepancy between teachers' perceptions of students' performance and students actual achievement. *Submitted to Learning and Instruction.*

Introduction

Several studies have researched populations of students, either split up into different groups or as a whole, in order to find differences between students in learning and assessment performances, and to search for explanations for these differences. In general it can be said that differences in cognitive skills between students is an important predictor with regard to differences in student performance. Results are also achieved with previous education and prior knowledge, often regarded as predicting factors with regard to student performance. Nevertheless, there are variables (both intervening and moderating) that can influence this relationship, such as social or contextual variables (van Laar & Sidanius, 2001; Woods, 1979; Yesilcay & Akman, 1996), study skills, study approaches and study strategies (Trigwell & Prosser, 1991; Vermunt, 1996; Watkins, 2001), motivational aspects (Biggs, 1993; Conti, 2000; Wolf, Smith & Birnbaum, 1995) and test anxiety (Boekaerts, 1988; Pekrun, 1988). Assessment formats and item formats, the subject matter and the students' perceptions of the assessment may also influence, direct or indirectly, students' performances (Birenbaum, 1997; Lundeberg & Fox, 1991; Scouller, 1998; Scouller & Prosser 1994; Thomson & Falchikov; 1998). In a nutshell, as well as the cognitive capacity of the students, there are different variables that may play a role in students' learning and assessment performances.

In terms of competences, highly competent students can be considered as cognitively strong students, with appropriate skills and attitudes towards their learning and assessment environment. In more detail, with reference to a problem-based learning setting, competent students are those who can quickly understand the contents, are able to provide coherent explanations, are able to generate a plan for a solution, implement problem solving strategies that reflect relevant goals independently, are able to monitor their actions and are flexible enough to adjust their approach (Baxter, Elder & Glaser, 1996; Cao & Xu, 2005). At the other end of the scale, the barely competent student can be defined as cognitively weak, with inappropriate skills and attitudes. In most cases the competence of students is related to assessment performances. Barely competent students are, for example, described as those performing in the bottom quartile of a class. Highly competent students are those performing in the top quartile (Kennedy, Lawton & Plumlee, 2002).

An aspect of assessment reliability is the ability of the assessment to discriminate in an appropriate way between barely and highly competent students, between low achieving and high achieving students. It is especially the case that where the assessment is a measurement of students' learning outcomes and checking whether the students have mastered the objectives, the assessment should be able to identify clear and distinct differences in students' performances. In most fields in higher education, despite the fact that new learning environments such as problem-based learning create a need for new assessment formats to promote appropriate learning approaches, cognitive objectives are mostly assessed by means of essay questions or open ended questions (Driessen, van der Vleuten & van Berkel, 1999). These questions give the assessors more freedom in assessing a broad range of cognitive processes, especially in case of assessing higher order

thinking skills. It is, however, generally acknowledged that the validity and the scoring reliability of such assessments can be poor (Ackerman & Smith, 1988). Multiple-choice questions are far more objective and therefore more reliable. In the debate regarding the equivalence of multiple-choice questions and essay questions, a lot is said about the advantages and disadvantages of both formats. For example, multiple-choice questions are often associated with assessing the reproduction of knowledge and promoting surface approaches to learning. But despite the fact that multiple-choice questions are regarded as an indirect measurement and cannot measure typical skills such as writing, creating and evaluating, it can be concluded from several studies that both question formats are actually measuring the same construct (Ackerman & Smith 1988; Bennet, Rock & Wang, 1991; Driessen, van der Vleuten & van Berkel, 1999; Ward, 1982). When discussing the reliability of the measurement, these studies suggest the use of both formats.

Teachers' perceptions of student achievement and learning outcomes

Teachers can develop expectations about their students by observing them in classroom interactions, by asking questions about their study progress, and by evaluating their assignment outcomes. It can be questioned whether these expectations of the observed learning outcomes are reflected in the students' assessment performances, and to what extent these expectations are met.

Most studies about teachers' perceptions of student achievement or performances are conducted in the field of social perceptions in kindergarten, pre school or primary school. These studies are concerned with race, gender, socioeconomic status and other demographic characteristics and have their origins in studies about the role of the self-fulfilling prophecy, or the Pygmalion effect, in the classroom by Merton (1948) and Rosenthal and Jacobson (1968). The tendency for teacher expectancies of students' performances to produce results consistent with these expectancies is explained by means of this effect by several researchers (Alvidrez & Weinstein, 1999; Madon, Jussim & Eccles, 1997; Trouilloud, Sarrazin, Martinek & Guillet, 2002). However, most studies in naturalistic learning settings showed that the influence of teachers' expectancy on student performance is limited and only takes place in specific circumstances. According to Jussim (1991) there are two alternative hypotheses that can explain the relationship between teacher expectations and student behaviour: (1) there are perceptual biases and (2) the teachers are accurate. Perceptual biases occur when teachers evaluate or assess their students themselves, with certain expectations. In the case of independent assessments such as standardised tests, or in cases where the assessment construction process separates teaching from assessing, perceptual biases are not likely. Because most research in naturalistic learning settings did not explain differences in teacher expectations and student performances by the Pygmalion effect or perceptual biases, the hypothesis that teachers are capable of predicting student performance accurately is probably correct. In a review by Hoge and Coladarci (1989) for example, strong correlations were found between teachers' judgements of academic achievement and the outcomes on standardised tests. Research by Alvidrez and Weinstein (1999), Bennett, Gottesman, Rock and Cerullo (1993) and Sweet, Guthrie and Ng (1998) support the value of teacher judgements

as indicators of school achievement: teacher perceptions seem to be related to cognitive ability, to classroom behaviour and characteristics such as attentiveness, compliance, conformity, effort, independency, motivation and neatness.

One study set in higher education, concerning the relationship between teacher judgements and student performances, is worth mentioning. The aim of this study by Valle et al. (1999) was to develop an instrument to assess student performances during tutorial settings in a problem-based learning environment. The result was a questionnaire which made it possible for teachers to observe and monitor students on the fundamental components of problem-based learning. The components were identified by means of exploratory factor analysis as 'independent study' (concerning students' initiative, motivation, studying and achieving the learning objectives and tasks), 'group interaction' (concerning the students' abilities to function in the group), 'reasoning skills' (concerning the students' abilities to analyse cases, formulate hypothesis and clarify concepts), and 'active participation' (concerning students' interaction with the group). High values for Cronbach reliability coefficients of the different components and inter-item correlations of each component mean that this approach to link student performances by means of teacher observations with assessment is valid and reliable.

The context of the present study

The present study took place in the first year of a law school using problem-based learning. During a course, teachers observed the students and determined each student's competence level. These observed competence levels are compared to the assessment results at the end of the course. The purpose of this study is to investigate to what extent the assessment discriminate between less and more competent students, based on the teachers' perceptions. More specifically, it investigates whether a rather traditional assessment, consisting of multiple-choice questions and essay questions, discriminates between students in a similar way to that done by teachers. The first research question can be formulated as follows: Do multiple-choice questions and essay questions discriminate between different groups of students, identified on the basis of characteristics observed by teachers? Whether the multiple-choice questions and essay questions discriminate between the different groups of students in the same way is questioned in the second research question: Is there a difference between the different groups of students with regard to the results on the different parts of the assessment?

Since the study was conducted in the last period of the academic year and the students were classified in this period it poses the question of whether the results for this period can also be discovered for previous periods. The third question is therefore formulated as follows: Are the observations with regard to the last period also valid for previous assessments?

Method

Participants

A total of 107 students, who were working in 9 study groups, were divided by two teachers into 4 types of students (these types are described below) at the end of the fourth period of the first year, on the basis of their observations. The population consisted of 56 female and 51 male students.

Independent variables

Teachers can acquire an accurate image of the differences between students in a study group by observing them during group interaction and the reporting on the students' learning goals, by looking at students' preparation and by asking students about their study progress during or after study group meetings. Students show effort and motivation by their level of preparation for a study group: they read the prescribed material more or less, they attain the learning goals, or they consult other sources. Little participation in the preparation of learning goals or the group discussions may be indicative of fear of failure. Trouble with keeping up with difficult subjects may say something about the cognitive capacity of a student. The level to which students are able to identify relevant information for the tasks, are able to keep a view of the bigger picture, are able to correctly identify learning goals, are able to participate in a useful manner in discussions, are able to refer to sources, and are interested in solutions or problem solving strategies of other students, is indicative of study attitudes and study methods.

By means of variations in the above named variables students can be divided by teacher observations into the following student types:

> Type I student: This type is the barely competent student. They are not at all, or not sufficiently, prepared for the study group meetings. Participation in group discussions is minimal and contributions are in general not very useful and contain little useful information. They are only interested in that what will be asked in the assessment. These students participate in the group meeting mainly for the social benefits (contacts with other students). In practice, one distinguishes two sets of Type I students: those with little motivation and those with little capacity.
>
> Type II student: This type of student is the moderately competent student and can be characterized as the average student. Preparation for group meetings is usually sufficient. These students need extra teaching and for the teacher it is easily predictable which, somewhat more difficult, subject requires this extra effort. They have sufficient confidence but encounter problems in switching from one subject to another. They do not pressure themselves in studying the required materials.

> Type III student: This student is the highly competent, cognitively very strong, student. These students are able to keep a view of the bigger picture, can distinguish between the main subjects and subjects of lesser importance and check themselves whether or not they have understood the subjects. These students consult sources that go beyond the prescribed materials. They feel at home in the study group and takes responsibility in guiding the discussion in the study group. They enjoy explaining subjects to their fellow students, or exchanging information and sources. They are relaxed in studying the required material. Teachers characterize these students as adaptive, flexible and cognitively strong students who have made a conscious choice for university and this specific study.

In the study group there are also students who are moderately or highly competent but whose performance in the study group is characterized by insecurity and forced behaviour. This type of student can be characterized as follows:

> Type IV student: This type of student prepares for study group meetings and has written out most of the preparation. During discussions, however, they try to maintain a low profile. But when these students are encouraged to do so, they can get to the core of the problem quickly and can contribute in a useful manner. These students are extremely interested in the solutions of other students. They gather all the information that is discussed in the study group, even if this information is not directly relevant for the core of the subject being dealt with. This type of student has the feeling that they have to study everything, which costs a lot of time, without gaining confidence or mastering the subject. They cannot easily distinguish between the core and side issues. Teachers characterize these students as cognitively capable, but not very adaptive and inflexible.

Dependent variables

The four assessments (assessment 1 in the first period, assessment 2 in the second, assessment 3 in the third, and assessment 4 in the last period) were constructed by assessment construction groups. They were each composed of a set of 40 multiple-choice questions (with four alternatives each) and one open question (questions that resemble tasks which have been dealt with in the study groups). For these assessments students could acquire 40 points for the multiple-choice part and 10 points for the essay part.

During several assessment construction meetings with groups of experts, different aspects of the assessment items were discussed in order to compose a valid and reliable assessment. The primary purposes of assessments 1 and 3 were to investigate the students' abilities to understand basic concepts and principles (comprehension, in terms of Bloom (1956); defined as the ability to grasp the meaning of material) and apply information, concepts and principles in new situations (application; refers to the ability to use learned material in new and concrete situations). The purposes of assessment 2 were to investigate the students'

abilities to recall information (knowledge; defined as the remembering of previous learned material) and apply information, concepts and principles in new situations. The principal purposes of assessment 4 were to investigate the students' abilities to recall information and to understand basic concepts and principles.

The reliability of the assessments (Cronbach's alpha coefficient) varied from .70 (assessment 3) to .78 (assessment 1).

Procedures

Students could enrol themselves for an available study group at a time of their choosing. Subsequently the resulting study groups were divided amongst the different teachers. The students attended the study group meetings and took an examination in the 9th week after the start of the course. The total teaching of the courses consisted of 16 study group meetings, 8 skills meetings and 8 lectures. As well as the assessment of the course, a skills exam was also taken. The two teachers involved divided those students that had attended at least five of the 16 group meetings into four different groups in the last week of the course in the last period.

From the total population from the survey, 19 students were classified as a type I student, 38 students as type II, 24 as type III, and 26 as type IV. From this population a total of 100 students took the assessment in the last period (assessment 4).

Analysis

Results were analysed by means of descriptive statistics for the assessment outcomes of the different parts of the assessment. Analysis of variances (ANOVAs) procedures were conducted to probe into the differences in assessment outcomes in relation to the four types of students as observed by the teachers. The research questions will be reported on one by one.

Results

Do multiple-choice questions and essay questions discriminate between different groups of students based on characteristics observed by teachers?

In order to find out whether the separate parts of the assessment discriminate between the different groups of students, the average scores on the different parts of the assessment were calculated. Table 5.1 indicates for each student type the scores on the multiple-choice part and the essay part of the assessment.

A one-way ANOVA was conducted to evaluate the relationship between the student type and the outcomes of the multiple-choice part and the essay part of the assessment. The ANOVA was significant for both assessment parts (multiple-choice part: $F(3, 96) = 11.03$, $p < .01$, $\eta^2 = .26$; essay part: $F(3, 96) = 9.52$, $p < .01$, $\eta^2 = .22$).

Table 5.1. Descriptive statistics for the different groups of students on the separate parts of assessment 4

	Multiple-choice		Essay		
Student type	*M*	*SD*	*M*	*SD*	*N*
type I	19.19	3.29	2.12	1.12	16
type II	22.33	5.40	4.38	2.47	36
type III	27.09	3.52	5.80	1.70	23
type IV	22.32	4.15	4.64	2.42	25
Total	22.92	5.04	4.41	2.39	100

Post hoc analyses to the ANOVA consisted of conducting pairwise comparisons, to find significant differences on the assessment outcomes between student types for both parts of the assessment. Type I students score significantly lower and the type III students score significantly higher on both parts of the assessment in comparison with the other student types.

The correlation between the score on the multiple-choice part and the essay part ($r = .518$; $p = .000$) can be called, taken into account previous results of studies on assessments combining essay and multiple-choice questions (see Driessen, van der Vleuten & van Berkel, 1999), strong. This means that students who have a higher score on the multiple-choice questions also tend to have a higher score on the essay part.

Is there a difference between the different groups of students with regard to the result on the different parts of the assessment?

In order to answer this question the average difficulty rate (p_{av}) was calculated for the multiple-choice part and the essay part. The average difficulty rate of the multiple-choice part was corrected for a "forced guess" score-system in order to compare the two different parts of the combination exam (p'_{av}). Table 5.2 indicates that the average difficulty rates of both parts of the assessment are comparable.

Table 5.2. Descriptive statistics on the average difficulty rate of the separate parts of assessment 4 for the different groups of students

	p' average mc		p essay		p difference		
Student type	*M*	*SD*	*M*	*SD*	*M*	*SD*	*N*
type I	.32	.11	.21	.11	.11	.14	16
type II	.43	.18	.44	.25	.01	.25	36
type III	.59	.12	.58	.17	.01	.17	23
type IV	.43	.14	.46	.24	.03	.21	25
Total	.45	.17	.44	.24	.01	.21	100

In order to determine whether the student experiences a difference in difficulty between the two parts of the assessment, the difference between the two average difficulty rates is calculated first (p' multiple-choice - p essay) and subsequently tested by means of the ANOVA procedure. In case of type I students, in comparison with the other student types, a difference in performance between the multiple-choice questions and the essay questions was found. This difference was not significant, however. It can be concluded that, generally speaking, students who are characterized by teachers as barely competent (type I), moderately competent (type II), highly competent (type III) and those whose performances are characterized by

insecurity and forced behaviour (type IV) perform in the same way on both parts of the assessment.

Are the observations with regard to the last period also valid for previous assessments?

In order to find out whether the discovered results are period specific or period independent, the results of the students from the survey on the different parts of the assessments were set out and compared. The previous questions are, therefore, asked again on the basis of previous exams.

Table 5.3 indicates the students' performances on the multiple-choice part of the previous assessments:

Table 5.3. Descriptive statistics for the different groups of students on the multiple-choice part of assessment 1, 2 and 3

	Assessment 1			Assessment 2			Assessment 3		
Student type	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*
type I	25.24	3.67	17	18.65	3.95	17	20.53	3.78	15
type II	27.71	4.63	31	20.97	3.75	35	21.68	3.65	31
type III	29.24	4.62	21	24.83	4.89	23	25.45	3.25	22
type IV	26.70	4.05	23	20.76	3.98	25	21.67	3.47	21
Total	27.35	4.46	92	21.41	4.56	100	22.42	3.91	89

ANOVAs and post hoc analyses to the ANOVAs were conducted to evaluate the relationships between the student types and the outcomes on the multiple-choice part for all three previous assessments. The ANOVA was significant for all assessments (assessment 1: $F(3, 88) = 2.94$, $p < .05$, $\eta^2 = .09$; assessment 2: $F(3, 96) = 6.92$, $p < .01$, $\eta^2 = .18$; assessment 3: $F(3, 85) = 7.61$, $p < .01$, $\eta^2 = .21$).

Post hoc analyses reveal that type III students score significantly higher on the multiple-choice part of the assessment on all previous assessments in comparison with the other student types for assessment 2 and 3. For assessment 1 the type III students only perform significantly better than the type I students. No significant differences between type I, II and IV students were found for this part of the assessment.

On the essay part of the previous assessments the students performed as is indicated in Table 5.4. The ANOVA was significant for all assessments (assessment 1: $F(3, 88) = 2.95$, $p < .05$, $\eta^2 = .09$; assessment 2: $F(3, 96) = 6.92$, $p < .01$, $\eta^2 = .14$; assessment 3: $F(3, 85) = 5.65$, $p < .01$, $\eta^2 = .17$). For assessment 1, post hoc analysis reveals a significant difference only between type IV students and type I students. In the case of assessments 2 and 3, type III students perform significantly better on the essay part on all previous assessments in comparison with the other student types. Furthermore, the analysis reveals significant differences between type I students and type II students on assessments 2 and 3.

Table 5.4. Descriptive statistics for the different groups of students on the essay part of assessment 1, 2 and 3

	Assessment 1			Assessment 2			Assessment 3		
Student type	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*
type I	3.44	1.70	17	2.56	1.37	17	1.73	1.46	15
type II	4.39	1.77	31	3.60	1.74	35	2.84	1.38	31
type III	4.74	1.96	21	4.50	1.66	23	3.84	1.58	22
type IV	5.15	1.97	23	3.22	1.47	25	2.52	1.92	21
Total	4.48	1.91	92	3.54	1.70	100	2.83	1.70	89

To answer the question of whether there is a difference between the different groups of students with regard to their results on the different parts of the previous assessments, the average (corrected) difficulty rates (p_{av}) were calculated for the multiple-choice part and the essay part of these assessments. Figure 5.1 is a graphical representations of the difficulty rates of assessment 4 and the previous assessments. From Figure 5.1 it can be concluded that, with regard to the difficulty of the separate parts of the assessment, the different groups of students perform in the same way on the previous assessments as on the most recent assessment (assessment 4), except for assessment 1. In the case of assessment 1, type IV students seem to perform relatively better on the essay questions than on the multiple-choice questions in comparison with the other student types. This was tested and confirmed by means of the ANOVA procedure ($F(3, 88) = 2.74$, $p < .05$, $\eta^2 = .08$).

Discussion

From this study involving first year students in a law faculty, it becomes clear that students who are characterized as highly competent by the teachers (type III students) performed better on both parts (multiple-choice questions and essay questions) of the assessment, in comparison with the other student types. The barely competent students (type I students) performed worse on both parts of the assessment. Not only has it become clear that both parts of the assessment lead to the same conclusion, it has also become clear that the perceptions of teachers of the performance of students (as operationalised in the different student types) corresponds with the actual performance of those students. It seems that Jussim (1991) is right to conclude that teachers are accurate in predicting student performance. The results and conclusions from studies by Alvidrez and Weinstein (1999), Bennet, Gottesman, Rock and Cerullo (1993), Hoge and Coladarci (1989), and Sweet, Guthrie and Ng (1998) are confirmed by our results. If we investigate further into the differences between these student groups relating to their performance on the two parts of the assessment, we can conclude that students perform on the multiple-choice questions in the same way as on the essay questions.

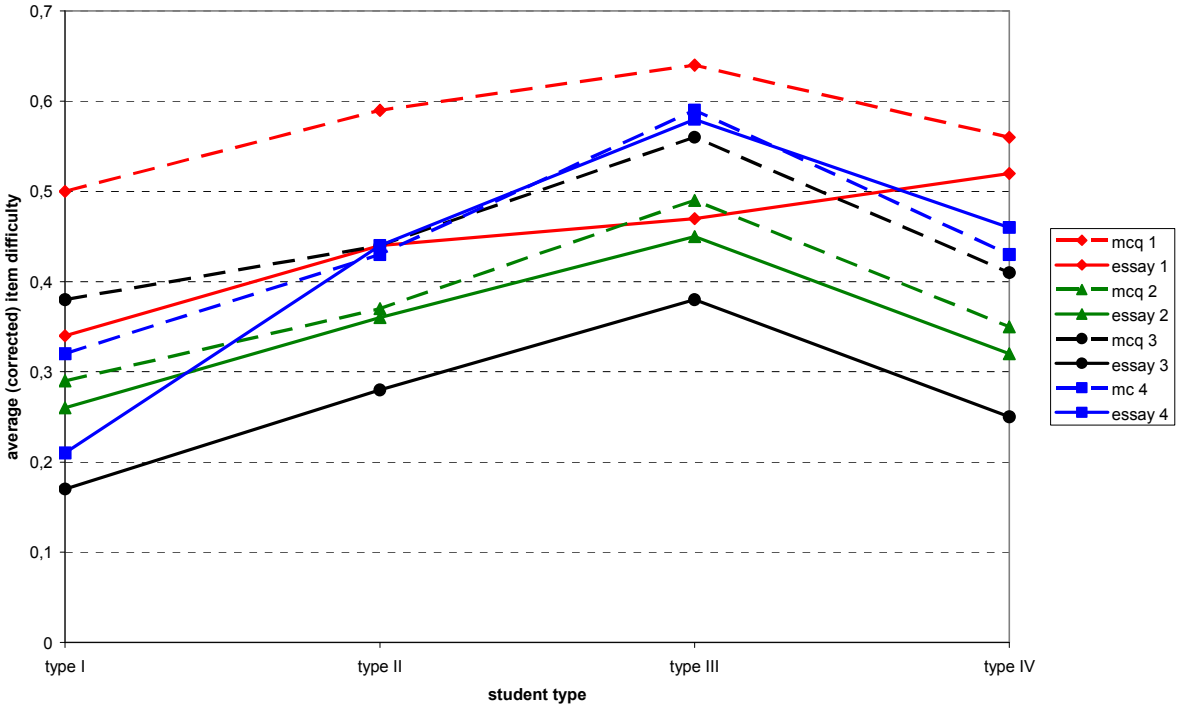

Figure 5.1. Average difficulty level of the assessment parts of the 4 assessments

Since the study was conducted in the last period of the academic year and the students were classified in this period, this poses the question of whether the results for this period can also be discovered for previous periods. It could well be that the results of the last period are specific for that period, because the students studied a different subject, in a different group, with a different teacher, compared to any of the previous periods. In addition it is not all that easy to compare the different assessments since it could well be that, based on the different characteristics of the different subjects, assessments differ in difficulty or in the division between knowledge (reproduction based questions and comprehension based questions) and skills (application based questions). Assessment 1 discriminates differently between the groups identified in the fourth period than the other three assessments. On the one hand it is possible that students were still looking for a suitable study method or were not used to answering essay questions (in a manner satisfactory for the assessors). On the other hand it could well be that assessment effects play a role in this phenomenon. A possible explanation for this result may lie in the so-called pre-assessment effects (Gielen, Dochy & Dierick, 2003). With this term the actual influence of the assessment on the study methods of the student, in particular the student's preparation for that assessment, is indicated. Type I students presumably view the assessment mainly as a multiple-choice assessment since the majority (with the least effort) of the points awarded can be gained in the multiple-choice part of the assessment. According to Boud (1990), students focus on those subjects and learning levels that are assessed and earn them, according to their perceptions, points. Obviously, there is a discrepancy between what is actually tested and what they expect to be tested (Broekkamp, 2002). From Broekkamp's study into the preparation of students in secondary education for their assessments that give access to university, it can be concluded that students mainly use reproductive strategies. According to Broekkamp, these strategies lead to reasonable results and the insight that is required in this stage is not of such a magnitude that it requires more thorough study approaches. If we combine the conclusions from our study with the conclusions reached by Birenbaum and Feldman (1998), students with a surface study approach tended to prefer multiple-choice formats, and Scouller (1998), students are more easily enticed to use a surface approach to learning if an assessment merely consists of multiple-choice questions, we can conclude that the type I student will have a preference for multiple-choice questions and will employ a surface study approach. The surface strategy will lead to this preference for a surface approach. This is a cyclic process which type I students, reflecting on their previous assessment performances, will not escape. It is to be believed that this type of student prepares for an assessment by collecting and practicing old assessments and by participating in training provided by student clubs in order to improve their performance on this part of the assessment.

For the type III students it could also well be that, on the basis of their previous experience and results in secondary education, they focus themselves on the multiple-choice part of the assessment and employ a surface study approach at first. As a consequence the type III student is less identifiable from the results of the essay part of the first assessment. Only after the first assessment do they realise that this form of assessment requires more than reproduction and as a consequence they

change their study approach, which leads to better results on both the multiple-choice and the essay parts of the assessment. In this case one can identify a positive influence of the pre-assessment effect.

In the case of the type IV students, there is a change in the students' perceptions of the assessment, which similarly influences the result of the assessment. After their experience with the first assessment, where this type of student achieves a better result on the essay questions than the other types, the student changes his perception of the assessment. Although they gain high scores on the essay questions, they are probably disappointed by the result of the assessment as a whole to such a degree that they alter the emphasis of their preparation to prepare more for multiple-choice questions. This occurrence can also be supported by the views of Boud (1990): students are prepared to change their study approaches if they believe their assessment results will improve. It seems that type IV students change their study approaches to one that is more directed towards reproduction and therefore to a surface approach. In addition to that they apply inappropriate study methods (not distinguishing between core matters and more peripheral matters). This could be an explanation for the fact that their assessment performances drop in the following exams. This occurrence could be identified as a negative influence of the pre-assessment effects on study behaviour.

Further study could be aimed at how different groups of students perceive the link between teaching and assessment. Recently the importance of extensive integration of teaching and assessment, and improving the assessment culture by means of blended assessment, has been pointed out (Dochy, 2005; Dochy, Gijbels & Segers, in press). Whether the different groups of students perceive this integration differently could also provide interesting information for this field of study.

References

Ackerman, T.A., & Smith, P.L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement*, *12* (2), 117-128.

Alvidrez, J., & Weinstein, R.S. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology*, *91* (4), 731-746

Baxter, G. P., Elder, A. D., & Glaser, R. (1996). *Assessment in instruction in the science classroom*. (CSE Tech. Rep. No. 418). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Bennet, R.E., Rock, D.A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement*, *28* (1), 77-92.

Bennet, R.E., Gottesman, R.L., Rock, D.A., & Cerullo, F. (1993). Influence of behaviour perceptions and gender on teachers' judgement of students' academic skills. *Journal of educational psychology*, *85* (2), 347-356.

Biggs, J.B. (1993). What do inventories of students' learning processes really measure? A theoretical review and clarification. *British Journal of Educational Psychology*, *63*, 1-17.

Birenbaum, M. (1997). Assessment preferences and their relationship to learning strategies and orientations. *Higher education*, *33*, 71-84.

Birenbaum, M., & Feldman, R.A. (1998). Relationships between learning patterns and attitudes towards two assessment formats. *Educational Research*, *40*, 90-98.

Bloom, B.S. (Ed.). (1956). *Taxonomy of educational objectives. Handbook I: The cognitive domain.* New York: McKay.

Boekaerts, M. (1988). Motivated learning: bias in appraisals. *Studies in Educational Evaluation*, *12*, 267-280.

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

Broekkamp, H. (2002). *Task demands and test expectations* [Dissertation]. Amsterdam: University of Amsterdam.

Cao, L., & Xu, P. (2005). *Activity patterns of pair programming*. Proceedings of the 38th Hawaii international conference on system sciences.

Conti, R. (2000). College goals: Do self-determined and carefully considered goals predict intrinsic motivation, academic performance, and adjustment during the first semester? *Social Psychology of Education*, *4*, 189-2111

Dochy, F. (2005, August). *Learning lasting for life and assessment: How far did we progress?* Presidential address at the 20th conference of the EARLI, Nicosia, Cyprus. Retrieved November 11, 2005, from http://www.earli.org/conferences/previous_conferences/earli_2005/presidential_adress

Dochy, F., Gijbels, D., & Segers, M. (in press). Learning and the emerging new assessment culture. In L. Verschaffel, F. Dochy, M. Boekaerts, & S. Vosniadou (Eds.), *Instructional psychology: Past, present, and future trends. Sixteen essays in honour of Erik De Corte*. Oxford: Elsevier Science (Pergamon).

Driessen, E., Vleuten, C. van der, & Berkel, H. van (1999). Beyond the multiple-choice v. essay questions controversy. *The Law Teacher*, *33*, 159-171.

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.), *Optimising New Modes of Assessment: In Search of Qualities and Standards* (pp. 37-54). Dordrecht: Kluwer Academic Publishers.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research*, *59*, 297–313.

Jussim, L. (1991). Social perception and social reality: A reflection-construction model. *Psychological Review*, *98*, 54–73.

Kennedy, E.J., Lawton, L., & Plumlee, E.L., (2002). Blissful ignorance: The problem of unrecognized incompetence and academic performance. *Journal of Marketing Education*, *24* (3), 243-252

Laar, C. van, & Sidanius, J. (2001). Social status and the academic achievement gap: A social domain perspective. *Social Psychology of Education*, *4*, 235-258.

Lundeberg, M.A., & Fox, P.W. (1991). Do laboratory findings on test expectancy generalize to classroom outcomes? *Review of Educational Research*, *61*, 94-106.

Madon, S., Jussim. L., & Eccles, J. (1997). In search of the powerful self-fulfilling prophecy. *Journal of Personality and Social Psychology*, *72* (4), 791-809

Merton, R.K. (1948). The self-fulfilling prophecy. *Antioch Review*, *8*, 193-210.

Pekrun, R. (1988). Anxiety and motivation in achievement setting: towards a system-theoretical approach. *Studies in Educational Evaluation*, *12*, 307-323.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectations and pupils' intellectual development*. New York: Holt, Reinhart & Winston.

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

Scouller, K.M., & Prosser, M. (1994). Students' experiences in studying for multiple choice question examinations. *Studies in Higher Education*, *19*, 267-269.

Sweet, A.P., Guthrie, J.T., & Ng, M.M. (1998). Teacher perceptions and student reading motivation. *Journal of Educational Psychology*, *90* (2), 210-223.

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*, 379-390.

Trigwell, K., & Prosser, M. (1991). Relating learning approaches, perceptions of context and learning outcomes. *Higher Education*, *22*, 252-266.

Trouilloud, D.O., Sarrazin, P.G., Martinek T.J., & Guillet, E. (2002). The influence of teacher expectations on student achievement in physical education classes: Pygmalion revisited. *European Journal of Social Psychology*, *32*, 591–607.

Valle, R., Petra, I., Martinez-González, A., Rojas-Ramirez, Morales-Lopez, J., & Piña-Garza, J.A. (1999). Assessment of student performance in problem-based learning tutorial sessions. *Medical Education*, *33*, 818-822

Vermunt, J.D.H.M. (1996). Metacognitive, cognitive and affective aspects of learning styles and strategies: A phenomenographic analysis. *Higher Education*, *31*, 25-50

Ward, W.C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude test. *Applied Psychological Measurement*, *6*, 1-11.

Watkins, D. (2001). Correlates of approaches to learning: A cross-cultural meta-analysis. In R. J. Sternberg & L. Zhang (Eds.), *Perspectives on thinking and cognitive styles* (pp. 165-195). Mahwah, NJ: Erlbaum.

Wolf, L.F., Smith, J.K., & Birnbaum, M.E. (1995). Consequence of performance, test motivation, and mentally taxing items. *Applied measurement in education*, *8* (4), 341-351

Woods, P. (1979). *The divided school.* London: Routledge and Kegan.

Yesilcay, Y., & Akman, K.I. (1996). A statistical model for the early detection of achievers and non-achievers in a university. *Studies in Educational Evaluation*, *22*, 59-77.

Chapter 6

Integrating assessment-tasks in a problem-based learning environment[6]

Abstract

The purpose of this study was to get more insight in the effects of written assessment-tasks integrated in a problem-based learning environment. Both the influence on students' performances and students' perceptions were investigated. Students' final exam results were used to find out whether students who make the assessment-tasks do better than students who do not. Answers from questionnaires and semi-structured interviews were used to discover the most important concerns in students' and teachers' perceptions of the assessment-tasks. The results indicate that making the assessment-tasks had positive influence on the students' overall performance. From the questionnaires and interviews it appears that both the students and the teachers see the benefits of the assessment-tasks. It is concluded that small steps in the change of the assessment system can result in relatively big changes in students' learning and results.

---

[6] Based on: Gijbels, D., van de Watering, G., & Dochy, F. (2005). Integrating assessment-tasks in a problem-based learning environment. *Assessment and Evaluation in Higher Education*, *30* (1), 73-86.

Introduction

The implementation of powerful learning environments in line with constructivist learning theories point to the necessity of reconceptualising the nature of assessment. It is generally believed and shown that assessment has an important impact on instruction and learning (Gibbs, 1999; Scouller, 1998). The alignment between the learning environments' objectives and the assessment is a 'magic bullet' in improving learning (Cohen, 1987). The direct and indirect impacts of assessment may be either positive or negative (Crooks, 1988). The main purpose is to make the assessment congruent with the instruction and align the assessment to what students should be learning (Biggs, 2003). Faced with such powerful contexts, assessment should be used strategically, designed to have educationally sound and positive influences. The traditionally view that the assessment of students' achievement is separate from instruction and only comes at the end of the learning process, is no longer tenable. As assessment, learning and instruction become more and more integrated, there is a strong support for representing assessment as a tool for learning (Dochy & McDowell, 1997; Sambell, McDowell & Brown, 1997). In this respect, Birenbaum (1996) has made a useful distinction between two cultures in the measurement of achievement. In the traditional so-called testing culture, instruction and testing are considered to be separate activities. The assessment culture is a consequence of the need to make learning and instruction more in congruence with assessment (Segers, Dochy & Cascallar, 2003).

Bridging the gap between new developments in the assessment culture and the daily educational and assessment practice faces a number of difficulties (Black & William, 1998). For one, assessment is still seen as separate from learning and instruction by many teachers (Torrance & Pryor, 1995). In order to bridge this gap Gibbs and Simpson (2003) present a framework based on theory and research on strategic changes in assessment. The framework can be used to identify the potential for improving student learning by making principle changes to assessment. The framework consists of 5 dimensions, under which assessment supports student learning. First, the design of the assessment can be used to influence the quantity and distribution of student effort. This is the case when the assessed tasks capture sufficient study time and effort and distribute this effort evenly across the topics and weeks. Second, the quality and the level of the students' effort can be influenced by assessment. When the tasks engage students in productive learning activities and communicates clear and high expectations to the students, assessment supports student learning. The three last dimensions concern the importance of feedback in the design of assessment as a tool for learning. The third dimension stresses the quantity and timing of the feedback. Feedback should be provided quickly enough to be useful to students and should be given both often enough and in enough detail. Fourth, the quality of feedback is important. Feedback should focus on learning, be understandable for students and linked to the purpose of the tasks and the criteria. Finally, students' response to feedback should be taken into consideration. Feedback should be received by and attended to the students and students should act upon the feedback in order to improve their tasks or their learning.

However, the integration of assessment, learning and instruction remains a challenge for most teachers. A progressive step in the desired direction would be to integrate teacher made written assessment-tasks in the learning process (Struyf, Vandenberghe & Lens, 2001). The central question in this study is whether integrating several well-designed teacher made written assessment-tasks in the learning environment can result in improvements in the overall-student performance? It is expected that students who make such assessment-tasks during a course will have higher grades on their final exam compared to students who do not participate in such assessment-tasks.

From empirical studies regarding the effects of integrated learning-assessment environments it is known that these environments do not always demonstrate the expected learning outcomes (Segers, 1996). Recent research shows that the way the learning environment is perceived by the students, rather than the factual curriculum, affects to a large extent how students cope with the learning environment and consequently their learning results (Segers & Dochy, 2001). It follows that educational interventions will be ineffective unless they modify students' perceptions (Lawness & Richardson, 2002). This means that investigating the way the learning environment is perceived by the students seems to be crucial for interpreting their learning outcomes (Segers, Dochy & Cascallar, 2003). Therefore, students' and teachers' perceptions are also taken into account in this study.

In the present study, the case of written assessment-tasks integrated in the learning-environment of a European Law School, using problem-based learning is analyzed. First, the characteristics of the problem-based learning environment and its assessment will be summarized. Second, the effects of the assessment-tasks on students' performance will be presented. Finally, students' and tutors' perceptions of the learning-assessment environment will be taken into consideration in order to get more insight into the effects.

Problem-based learning

Although originally developed for medical training in Canada, the orthodox version of problem-based learning (PBL) has been modified and applied globally in many disciplines (Gijselaers, 1995). Problem-based learning is at present receiving more and more attention in various programs of higher education. In the literature, PBL has been defined and described in different ways. In spite of the many variations of PBL that have evolved, a basic definition is needed to which other educational methods can be compared. Based on the original method as developed in McMasters Univerisity, Barrows (1996) developed six core characteristics of PBL. The first characteristic is that learning needs to be student-centered. Second, learning has to occur in small student groups under the guidance of a tutor. The third characteristic refers to the tutor as a facilitator or guide. Fourth, authentic problems are primarily encountered in the learning sequence, before any preparation or study has occurred. Fifth, the problems encountered are used as a tool to achieve the required knowledge and the problem-solving skills necessary to

eventually solve the problem. Finally, new information needs to be acquired through self-directed learning.

In the law school, in general students work in small tutorial groups (12-18 students) and meet twice a week under the supervision of a teacher (tutor) but chaired by a student-member of the group. Each session, students are confronted with a range of tasks or problems, which they analyze and try to solve by formulating 'learning goals' for their self-study. In the next session students report their findings on the basis of the materials they looked for, found and studied and start with analyzing new problems. Besides this, students are enrolled on a weekly basis in somewhat larger 'practical groups' (24-36 students) and have one lecture a week. For a more extensive description of the legal curriculum and the problem-based approach in the law school, see Moust (1998) or Pletinckx and Segers (2001).

Assessment and PBL

A wide range of assessment methods has been used to assess students learning in PBL. Ranging from traditional multiple-choice exams over essay exams to new modes of assessment such as case-based assessment, self- and peer assessment, performance-based assessment and portfolio assessment. Recently, many educators and researchers have advocated new modes of assessment in order to be congruent with the educational goals and instructional principles of PBL (Segers, Dochy & Cascallar, 2003). It is generally recognized that a seventh characteristic should be added to the six core characteristics of Barrows (1996). Essential for PBL is that students learn by analyzing and solving representative problems; consequently, a valid assessment system evaluates students' problem-solving competencies in an assessment-environment that is congruent with the PBL environment. This means, the assessment in PBL should take into account both the organization of the knowledge base, as the students' problem solving skills (Segers, et al., 2003).

Recently, a meta-analysis of the effects of PBL (compared to more traditional educational methods) included the method of assessment as a moderator variable, suggesting that the more an instrument is capable of evaluating the students' competence in knowledge application, the larger the ascertained effect of PBL would be (Dochy, Segers, Van den Bossche & Gijbels, 2003). A further exploration of the effect of what is measured with the assessment on the effects of PBL (Gijbels, Dochy, Van den Bossche & Segers, 2003) showed that there is a difference in the reported effects of PBL between the different measurement-levels used in the study. As expected, the effect of PBL is larger compared to conventional education when the assessment method is focusing at 'the understanding of principles that link concepts'. Contrary to studies suggesting that the effects of PBL are larger when the more complex levels of the knowledge structure are being assessed, the effect size for 'application' (linking of concepts and principles to application conditions and procedures) was not statistically significant. These results implicate a challenge for PBL to pay more attention to 'application' in both the teaching and learning environment as the assessment.

In the law faculty, for each course, a table of specification using Bloom's taxonomy (1956) is created in order to guarantee each subject matter is assessed on

the desired level. Generally, assessment takes place immediately after each course by means of multiple-choice and/or essay questions. For more information about the assessment system in the law school, see Driessen, Van der Vleuten and Van Berkel (1999) or Driessen and Van der Vleuten (2000).

Research questions

In order to get more insight in the effects of written assessment-tasks integrated in the learning environment on students' performance two research questions are formulated. First, do students who make the assessment-tasks do better in their final exam compared to students who do not? Second, what are the most important concerns in students' and teachers' perceptions of the assessment-tasks?

Method

Participants

A total of 205 students, following a course on public law in the second year of their law study, participated. Out of these 205 students, 164 students completed all six assessment-tasks. These students will be considered as the participants in the 'assessment-task' condition. The remaining 41 students didn't complete the six assessment-tasks and are as a consequence considered as the participants in the 'no assessment-task' condition. A total of 10 staff-members involved in the course participated in the study by completing the questionnaires and being interviewed.

Procedure

The research was carried out within the context of a compulsory second year law course. The study was carried out to evaluate the introduction of assessment-tasks in the faculty. A total of six assessment-tasks were distributed over different topics and weeks in the course. Assessment in the course was twofold. The final exam consisted of 40 multiple-choice questions and took place at the end of the course. During the course, students had the opportunity to complete the six assessment-tasks on a voluntary basis, which could result in an extra 'bonus point', added to the score of the final exam. Both the assessment-tasks and the multiple-choice questions asked several cognitive activities from the students in line with the instructional goals of problem-based learning.

The design of the assessment-tasks was to a great extent in line with the framework presented by Gibbs and Simpson (2003). Students were stimulated to produce qualitative learning activities by giving the one extra 'bonus-point' only if all six assessment-tasks showed to be of sufficient quality and effort. The feedback students got from their tutor or from the plenary discussion in the tutorial group could help them to make their next assessment-task better and to get a better understanding of the learning materials to be studied in order to pass the final exam at the end of the course.

Two methods for collecting data were employed: a quantitative and a qualitative approach. The quantitative approach will be presented first. Variables included are students' prior assessment scores, six assessment-tasks, a multiple-choice final exam and closed questions for both staff members and students.

Prior assessment scores

Since differences in performances between students making and succeeding all the assessment-tasks and students not doing so might be due to differences in prior academic achievement rather than in making or not making the assessment-tasks, the average scores of the students' previous academic exams was taken into account. Prior academic achievement is shown to be a good predictor of performance (House, Hurst & Keely, 1996, Curall & Kirk, 1986). As stated by Young (1993) "*the one aspect of student performance on which there is general consensus on importance is academic achievement, typically measured by the grade point average*". In this case the grade point average (GPA) consisted of the scores students earned in the six previous final exams they took.

Assessment-tasks

With each of the six assessment-tasks, the students were asked to write an essay in which they had to go through a process of evaluation, synthesis, analysis or understanding of the study material or had to write down in detail the solution of a problem as presented in a case (see Figure 6.1). In order to succeed, each of the six assessment-tasks had to be of sufficient effort and correctness. The tasks were discussed in the same way as the other problem tasks in the course after the students handed in their assessment-task. In this way the students got some plenary feedback. One week later, the students received back their assessment-task from the tutor, with the necessary feedback.

Case: Assessment-task
Neighbour Smith from the previous case also received an announcement of the given (limited) construction permit to Miss Jones on the 3rd of February 2003. He fears to have not that much light in his neighbouring garden, if the warehouse for flowers arises. He wonders whether he can stop the start of the building, although he is not involved in the ongoing procedure between the Miss Jones and the local authority. Moreover, it shouldn't be going to cost a lot, since he's on social security. On the other hand, he is so untrained, that he could use legal aid.

Give your advice to neighbour Smith, write the necessary documents (maximum 2 A4).

Figure 6.1. Example of an assessment-task

Final exam

The objectives of the final exam are threefold and derived from Bloom's taxonomy (Bloom, 1956): To investigate the student's ability to recall information (in terms of Bloom: knowledge; defined as the remembering of previous learned material), to understand basic concepts and principles (comprehension; defined as the ability to grasp the meaning of material) and to apply information, concepts, and principles in new situations (application; refers to the ability to use learned material in new and concrete situations). The final exam consisted of 12 knowledge based questions (30%), 14 comprehension based multiple-choice questions (35%) and 14 application based multiple-choice questions (35%). Out of these questions, 2 knowledge based, 4 comprehension based and 4 application based questions were orientated towards the topics related to the assessment tasks. The final exam was being found valid by the course supervisor. On average students scored 26.01 out of 40 on the exam (*SD* = 5.06). The internal consistency reliability coefficient was also measured and found appropriate (Coefficient alpha = .74).

Closed questions

The students' and tutors' survey consisted of three parts. The first part of the survey (both students and tutors) consisted of four questions concerning their view of working with the assessment tasks during the course. The items (e.g. "During the course it was made possible to make (assessment) tasks embedded in the curriculum") were rated on a five point Likert scale (1 = disagree completely, 5 = agree completely). The reliability of this part, questioning whether the assessment-tasks were perceived by the students and tutors as intended by the designers of the assessment-tasks, was high (Coefficient alpha = .78). The second part of the survey (both students and tutors) was developed to measure the students' and tutors' perception of different quality aspects of the (problem-based) learning environment, especially the assessment. The items are based on the characteristics of the learning-assessment environment and on the expectations of the staff concerning students' learning activities. This part of the questionnaire consisted of 15 items on a five point Likert scale concerning mostly the motivational aspects of the assessment tasks (e.g. "The use of assessment tasks stimulates me to work systematically"). The reliability was high (Coefficient alpha = .84). The third part of the survey (only for students) concerned the time students spend on the assessment tasks and their self-study activities.

Qualitative instruments

The qualitative aspect of this study took the form of an open ended questionnaire for both staff and students and a semi-structured interview with each staff-member. For example the students were asked to reflect on the assessment tasks embedded in the curriculum and making assessment tasks as part of the

curriculum in general. The aim was to get more insight in the most important concerns in students' and teachers' perceptions of the assessment tasks.

Results

The research questions will be reported one by one after the summary of the data of the prior assessment scores of the students in both the 'assessment-task' and the 'no assessment-task' condition.

Previous performance

The difference between the previous final exams mean scores of the two assessment groups is significant ($t(70.64) = -3.36$, $p < .05$) meaning that the two groups indeed differ in academic achievement. The group of students making the assessment-tasks did significantly better on their previous final exams. In order to make both groups comparable, analyses of covariance (ANCOVA) are used as method of analysis in order to answer the first two research questions. Preliminary analysis of the data involved inspection of normality and homogeneity of variance assumptions. Normal plots, stem-and-leaf plots and the calculation of skewness and kurtosis were used to check the normality of distribution. To test the equality of group variances the Levene statistics was calculated. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate (GPA) and the dependent variable (final assessment) did not differ significantly as a function of de independent variable (assessment-task). All assumptions for the analysis were met.

Effect of participating in the assessment tasks on the final exam

The ANCOVA was significant ($F(1, 202) = 9.63$, $MSE = 10.39$, $p < .01$, partial $\eta^2 = .05$). This means that, after correction for differences in students' prior performances, students who make and succeed all six assessment-tasks perform better on the final exam. The strength of relationship between the condition (assessment tasks or no assessment task) and the result on the final exam however was small, as assessed by a partial η^2, with the condition accounting for 5% of the variance of the dependent variable, holding constant the GPA. In order to investigate whether this effect is the result of a better preparation of the students for those topics treated with the assessment tasks only, our first research question is further divided in two more detailed research questions: first, do students who have completed and succeeded all the assessment tasks perform better on those questions in their final exam that are related to the topics treated with the assessment tasks compared to students who did not completed or succeeded all the assessment tasks? Second, do students who have completed and succeeded all the assessment tasks perform better on that part of the final exam that is not related to the topics treated

with the assessment tasks compared to students who did not completed or succeeded all the assessment tasks?

Effect of participating in the assessment tasks on related final exam questions

The results of the ANCOVA show that, after correction for differences in prior performances, there is a significant difference in outcome on that part of the final exam orientated towards the topics related to the assessment tasks between the two assessment groups ($F(1, 202) = 11.18$, $MSE = 2.18$, $p < .01$, partial $\eta^2 = .05$). Students who took the assessment tasks and met the criteria of these tasks performed significantly better on the questions in the final exam related to the topics treated in the assessment-tasks.

Effect of participating in the assessment tasks on non-related final exam questions

The results of the ANCOVA indicate that, after correction for differences in prior performances, there is also a significant difference in the outcome for that part of the exam not orientated towards the topics related to the assessment tasks ($F(1, 202) = 3.97$, $MSE = 6.46$, $p < .05$, partial $\eta^2 = .02$). Apparently, working with assessment tasks has a positive influence on the performance not only of the related topics but also of the non- related topics. Although the statistical significance is smaller than with the related final exam question (significant at the .05 versus the .01 level) and the partial η^2 is relatively small (meaning that participating in the assessment tasks accounts for only 2 percent of the variance of the final exam if the prior performances of the students are hold constant) there is statistical significant evidence that working on the assessment tasks had a positive influence on the performance not only on the specific questions but also on the whole final exam performance.

Students' and teachers' perceptions of the learning-assessment environment

In order to search for explanations for the results of these studies, students' and tutors' perceptions of the learning-assessment environment are investigated exploratively. This is also used as the input for recommendations for the improvement of educational practice. First, we will report the results of the closed questions.

Closed questions

In the tutors' view, the assessment tasks were embedded very well in the curriculum. All students, who took the assessment tasks, shared this point of view and were also positive about the way it was embedded. Both students and tutors share the opinion that by taking the assessment tasks more insight was gained into the extent of understanding the course material. There was, however, no significant

correlation found between the assessment score and the perception of students on working with the assessment tasks during the course.

Students and tutors thought the assessment tasks were meaningful because they stimulated some desired learning activities. However, their opinions differed regarding the extent to which the learning activities were stimulating desired learning activities. Students were positive about the overall stimulus going out from the assessment tasks. They spend more time on the study material and in their view they worked more independent and more systematically as a result of the assessment tasks. Students found themselves more capable in the case of problem solving processes, critical thinking and reasoning. In a nutshell, according to the students, it was because of the assessment tasks they mastered the study material better, despite the fact that making these tasks was at the expense of the preparation of the group sessions. A small negative, but significant correlation was found between the two items asking whether making the assessment tasks was at the expense of the preparation of the tutorial and practical group sessions and the final exam score (respectively: $r = -.26$, $p < .01$ and $r = -.27$, $p < .01$). This suggests that students with a better time management perform slightly better on the final exam. Students preferred to work with assessment tasks in the future, highly motivated by the prospectus of the bonus point. In the view of the tutors the use of assessment tasks especially (or only) stimulated independent activities and reasoning. In contrast with the opinion of the students, the tutors did not think the students generally mastered the study material better.

Students who took the assessment tasks and met the criteria reported to have spend substantially more hours on self-study per week for the tutorial groups ($M = 18.44$, $SD = 9.60$) than the students who did not take or turn in the assessment tasks successfully ($M = 13.04$, $SD = 8.60$). However, there was no significant correlation between the amount of reported self-study hours and students' score on their final exam.

Qualitative data results

The open-ended questions and the interviews were found to have many features in common. Five important issues emerged: the learning effect, time management, the bonus point system, the usefulness of the assessment tasks, and the feedback.

The learning effect

The learning effect of the assessment tasks was very obvious for both students and tutors. As students stated: "*working with the tasks was very stimulating to work more intensive with the learning material*". Consequently, the students were, according to the more experienced tutors, better prepared for the tutorial groups. This resulted according to the tutors in better, more in-depth discussions. Because of this effect, tutors argued for the use of assessment tasks strategically (also taking their time schedule into account, see feedback) in case of difficult tenets.

Feedback

Tutors reported that assessing a piece of work and giving appropriate feedback was only possible by checking it thoroughly. Further more they thought it was important for students to receive sound and detailed feedback. They found it inappropriate not to check students' work thoroughly and only looking for evidence of sufficient effort, as was proposed at first. However, for many tutors this resulted in a larger than planned amount of time for the tutors to evaluate the tasks. The presence of a well-constructed correction model was, according to the tutors, very helpful and time reducing to give their feedback. Also students reported to find the feedback useful.

Time management

Students spend more time on the learning material related to the assessment tasks. Though they spend more time on these tasks, they did not spend much more time overall. Taking the assessment tasks was mostly at the expense of students' preparation of other tutorial or practical group session as stated by the students themselves and as observed by the tutors. Because of the assessment tasks tutors spend more time evaluating the tasks, process the results and manage the data. Consequently, for some tutors this meant they were less prepared for the tutorial group sessions and for other tutors this meant they worked overtime.

Bonus point system

The students were very positive about the bonus point system. Nevertheless, they suggested a major change in the system. Their main consideration was about the way the bonus point was awarded. They thought that awarding a part of a bonus point for each assessment tasks is a more fair system instead of awarding a single bonus point for the total 6 assessment tasks. In this way students in circumstances beyond their control wouldn't be excluded for a bonus point. The tutors had two major considerations about the system. Their first consideration was about the weight of the bonus point. Their second consideration was about the amount of tasks. Several tutors suggested decreasing the amount of assessment tasks (three instead of six). In the eyes of the tutors the main benefit of this operation is that it reduces the amount of hours correcting/evaluating the assessment tasks and so this won't be at the expense of the preparation for the tutorial group sessions.

The usefulness of the assessment tasks

For students the criteria of the assessment tasks were not all stated clear and concrete enough. At first, most students didn't know exactly what was to be expected. Frequently students asked questions about its form and the content criteria. The tutors' additional explanations and guiding had to set this problem right. Other indicated problems were more structural like problems with the timing of the assessment tasks: the spacing of the 6 assessment-tasks over the 8 weeks sometimes was poor as was the tuning between the assessment tasks presented in the practical course book and the subject matter treated at that moment in the tutorial group. In other words, there was a need of especially constructed assessment tasks, capable of promoting long term learning effect, and containing an

amount of topics handled in the previous settings. Although the present assessment-tasks meet these requirements to a large extent, improvements could be made.

Conclusion and discussion

By introducing teacher made assessment-tasks in a problem-based learning course, the PBL learning environment becomes more in alignment itself. The five dimensions in the framework under which assessment supports student learning (Gibbs and Simpson, 2003) seem to be recognized by students and teachers in the assessment-tasks. When corrected for differences in prior academic performance, students who did make all the assessment-tasks successfully during the course, performed better on their final exam compared to students who did not make nor succeed the assessment-tasks. The fact that students did not only perform better on those parts of the final exam which were related to the assessment-tasks, but also on non-related questions indicates that the introduction of the assessment-task helped students to address more appropriate student learning activities, going beyond the six assessment-tasks and its content. The way in which the assessment tasks seem to have influenced the general learning approach of the students should be subject of further research.

From the closed questions it is clear that, in general, both students and tutors were happy about the way the assessment-tasks were embedded in the curriculum. Students reported to study more, more critically and more systematically as a result of the assessment tasks. It is suggested that students with a better time-management perform slightly better on their final exam.

Also the qualitative data seems to suggest that both the students and the teachers see the benefits of the assessment-tasks. Students are driven by the extra bonus-point to complete the assessment-tasks in the first place. Tutors are concerned about the feasibility of the whole project. As the students, they perceive a time-management problem: how to manage to give feedback on all the assessment-tasks. The presence of a correction model is helpful, but in the future the use of self- and peer-assessment can be considered to the benefit of the learning of the students and the time management of the tutors.

Several remarks can be made on this study and the implementation of the assessment-tasks. First of all, the students were free to participate in the study. One could assume that only the better students would be willing to spend extra time on the assessment-tasks in order to gain an extra bonus-point. Although the groups were corrected for their prior academic achievements, this study was far from a randomized trial. Other factors, such as students' motivation, could have had an influence on students' choice to participate in the assessment-tasks. Another possible moderating variable that has not been taken into account is teacher time: students participating in the assessment-tasks received more teacher time by means of the personal feedback on each assessment-task. Finally, the perspective of the bonus-point did motivate the students to participate in the assessment-tasks, but also hindered them seeing the assessment-tasks as a tool for learning.

In spite of its limitations, this study is in line with and supports the framework Gibbs and Simpson (2003) have described and suggests that small steps in the

change of the assessment system can result in relatively big changes in students' learning and results.

Acknowledgement

The authors are grateful to Catherine De Rijdt, Sabine Dierick, Katrien Lauwaert and Joëlle Pletinckx for their help with collecting the data.

References

Barrows, H.S. (1996). Problem-based learning in medicine and beyond. In L. Wilkerson & W.H. Gijselaers (Eds.), *Bringing problem-based learning to higher education: Theory and Practice.* New directions for teaching and learning, No. 68 (pp. 3-13). San Francisco, CA: Jossey-Bass.

Biggs, J. (2003). *Teaching for Quality Learning at University* (2nd ed.). Buckingham: SRHE and Open University Press.

Birenbaum, M. (1996). Assessment 2000: towards a pluralistic approach to assessment. In M. Birenbaum & F. Dochy (Eds.), *Alternatives in Assessment of Achievement, Learning Processes and Prior Knowledge* (pp. 3-30). Boston: Kluwer Academic.

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5* (1), 7-74.

Bloom, B.S. (Ed.). (1956). *Taxonomy of educational objectives. Handbook I: The cognitive domain.* New York: McKay.

Cohen, S.A. (1987). Instructional alignment: searching for a magic bullet. *Educational Researcher*, *16* (8), 16-20.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, *58* (4), 438-481.

Curall, S.C., & Kirk, R.E. (1986) Predicting success in intensive foreign language courses. *Modern Language Journal*, *70* (2), 107-113.

Dochy, F., & McDowell, L. (1997). Assessment as a tool for learning. *Studies in Educational Evaluation*, *23* (4), 279-298.

Dochy, F., Segers, M., Van den Bossche, P., & Gijbels, D. (2003). Effects of Problem-based Learning: A Meta-analysis. *Learning and Instruction*, *5* (13), 533-568.

Driessen, E., & Van der Vleuten, C. (2000). Matching student assessment to problem-based learning: lessons from experience in a law faculty. *Studies in Continuing Education*, *22* (2), 235-248.

Driessen, E., Van der Vleuten, C., & Van Berkel, H. (1999). Beyond the multiple-choice v. essay questions controversy: combining the best of both worlds. *The Law Teacher*, *33* (2), 159-171.

Gibbs, G., & Simpson, C. (2003). Does your assessment support your students' learning? *Learning and Teaching in Higher Education*, *1* (1), Retrieved December 5, 2005, from http://www.open.ac.uk/science/fdtl/documents/

Gijbels, D., Dochy, F., Van den Bossche, P., & Segers, M. (2003, August) *The relation between assessment practices and outcomes of studies: The case of problem-based learning***,** Paper presented at the 10th Biennial conference of the European Association for Research on Learning and Instruction, Padua, Italy.

Gijselaers, W. (1995). Perspectives on problem-based learning. In W. Gijselaers, D. Tempelaar, P. Keizer, J. Blommaert, E. Bernard & H. Kasper (Eds.), *Educational Innovation in Economics and Business Administration: the Case of Problem-based Learning* (39-52). Norwell: Kluwer Academic Publishers.

House, J.D., & Hurst, R.S., & Keely, E.J. (1996). Relationship between learner attitudes, prior achievement, and performance in a general education course: a multi-institutional study. *International Journal of Instructional Media*, *23* (3), 157-271.

Lawness, C.J., & Richardson, J.T.E. (2002). Approaches to studying and perceptions of academic quality in distance education. *Higher Education*, *44*, 257-282.

Moust, J.H.C. (1998) The problem-based education approach at the Maastricht Law School. *The Law Teacher*, *32*, 5-37.

Pletinckx, J., & M. Segers (2001). Programme evaluation as an instrument for quality assurance in a student-oriented educational system. *Studies in Educational Evaluation*, *27*, 355-372.

Sambell, L., McDowell, L., & Brow, S. (1997). "But is it fair?": An exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation*, *23* (4), 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. (1996). Assessment in a problem-based economics curriculum. In M. Birenbaum & F. Dochy (Eds.), *Alternatives in Assessment of Achievements, Learning Processes and Prior Learning* (pp. 201-226). Boston: Kluwer Academic Press.

Segers, M., & Dochy, F. (2001). New assessment forms in problem-based learning: the value-added of the students' perspective. *Studies in Higher Education*, *26* (3), 327-343.

Segers, M., Dochy, F., & Cascallar, E. (2003). The era of assessment engineering: changing perspectives on teaching and learning and the role of new modes of assessment. In M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimising New Modes of Assessment: In Search of Qualities and Standards* (pp. 1-12). Dordrecht: Kluwer Academic Publishers.

Torrance, H., & Pryor, J. (1995, April) *Making sense of 'formative assessment': Investigating the integration of assessment with teaching and learning,* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Young, J.W. (1993) Grade Adjustment Methods. *Review of Educational Research*, *63* (2), 151-165.

Chapter 7

Conclusion and discussion

The main issue of this dissertation was concerned with assessment and its outcomes. It is believed that assessment has an important impact on learning and can be a powerful means to focus on learning. In order to have a positive effect from the assessment, it is important to align the learning environment and the assessment and consequently enhance appropriate learning activities. However, a fit between the learning environment and the assessment alone is not effective enough. From several studies it is known that preferences and perceptions have its impact on how students cope with the learning environment, the assessment and the learning outcomes.

The studies in this dissertation took place in a constructivist learning environment with the characteristics of a problem-based learning, as designed at the faculty of law of the University of Maastricht. Assessment in a problem-based learning environment can be problematic with regards to the previously mentioned alignment. Though the assessments used in the studies were rather traditional (a combination of multiple-choice questions and essay questions) the assessments were considered valid and (partly aligned) because a part of the assessments focussed on assessing the students ability to apply knowledge in new situations, to solve problems by choosing the right answer (in case of multiple-choice questions) or by writing down the solution and its motivation (in case of essay questions).

Figure 7.1 represents schematically the relations between the different concepts in this dissertation and the chapters in which the concepts and its relationships were object of study. The study in the chapter 1 was concerned about the students' perception of their learning environment. In chapter 2 the focus lies on the students' approaches to learning in the problem-based learning environment and on the assessment outcomes using a rather conventional assessment format. Chapter 3, 4 and 5 focus on students' and teachers perceptions of assessment, item difficulty (and performance setting) and learning outcomes. Chapter 6 focuses on the learning outcomes and teachers' and students' perceptions of the adaptation of the assessment system towards more constructivist alignment. In this chapter the results of each study will be summarised and the conclusions will be discussed in three separate sections: 'Students' perceptions of their learning environment and their approaches to learning', 'the role of preferences and perceptions of assessment', and 'Adapting the assessment towards a more constructivist alignment'.

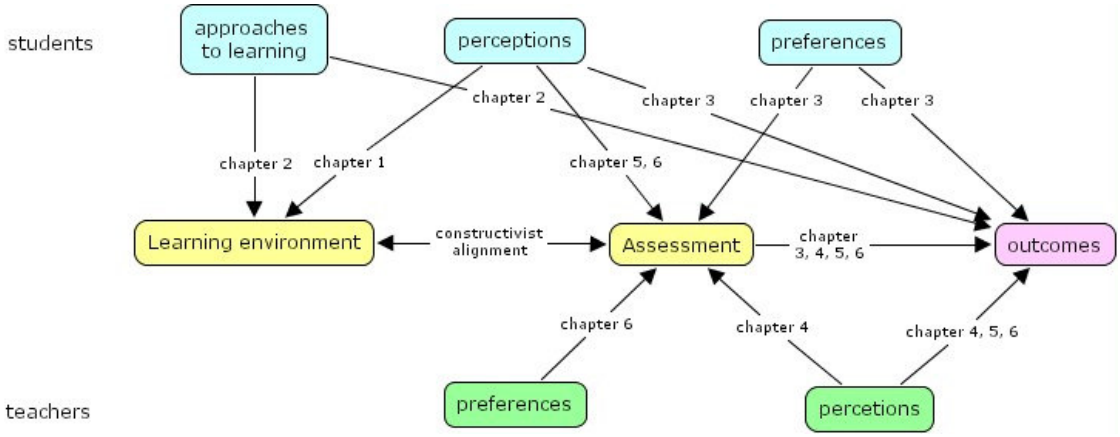

Figure 7.1. Schematic representation of the objects of study in the dissertation

Students' perceptions of their learning environment and their approaches to learning

Students' perceptions of the learning environment are seen as a powerful factor in the way that students cope with that learning environment. Learning environments based on constructivism have the potential to improve the educational outcomes, but are less effective if educational interventions don't succeed in modifying students' perceptions in the intended way. The study in the chapter 1 was concerned with the students' perception of their learning environment. In chapter 2 the students' approaches to learning in the problem-based learning environment as described in the first chapter were measured.

Students in a problem-based learning environment and in a conventional lecture-based environment were asked, using the questionnaire of Tenenbaum, Naidu, Jegende and Austin (2001; consisting of seven key factors of constructivist learning environments), how they experienced the educational practice in order to probe into students' perceptions of their learning environments. The main aim of this chapter was to investigate whether students in a problem-based learning environment perceive the learning environment to be more constructivist compared to the students' perceptions of a conventional lecture-based environment. Results showed that, in case of four key factors, students participating in problem-based learning perceived their learning environment to be more constructivist. These key factors are (in order of difference in perception between the two learning environments): 'conceptual conflicts and dilemmas', 'sharing ideas with others', 'meeting students needs' and 'arguments, discussions, debates'. The main characteristics of the first two factors (confronting students with conceptual conflicts and learning is a cooperative process) determine the strength of problem-based learning (PBL) in incorporating constructivist principles. The main characteristic of the factor 'arguments, discussions and debates' is learning is an

active and cumulative construction of knowledge. Though the difference in perception was significant, the factor 'meeting students' needs' was only moderately present in the problem-based learning environment. Meeting students' needs is about issues such as students' satisfaction with their learning outcomes, the possibility for students to pursue personal goals, to benefit from their own learning difficulties, and to what extent the faculty takes students needs and concerns into consideration. The moderate presence of this factor indicates that students in this learning environment only had a relatively small say in the learning process.

As a consequence one of the opportunities, which will be more in line with the constructivist principles, is to make sure that learning is more student centred, to make sure that students are in a larger extent owners of their learning process. And although the students' perceptions differed significantly on these factors and effect sizes varied from large to sufficient, the differences between the two learning environments are not 'extremely' large. For the conventional lecture-based course this means that, according to the perceptions of the students, constructivist principles are also partly incorporated. No differences were found for the factors 'making meaning, real life examples', 'motivation towards reflections and concept investigation' and 'materials and measures targeted toward solutions'. For the constructivist learning environment this means that, if problem-based learning claims to be highly consistent with constructivist features, at least in the perception of the students, a lot of opportunities still remain to be taken up. This is in line with the statement made by Savin-Baden (2000) that the "*potential of problem-based learning is yet to be fully realized*" (p. 2). For problem-based learning, tutors should be aware of the importance of facilitating to create a well functioning, cooperative tutorial group that promotes meaningful knowledge construction. Our own experiences have shown that this is not fully realised in the faculty.

The revised two factor study process questionnaire (R-SPQ-2F) measures to what extent students prefer to use a deep approach to learning or a surface approach to learning (Biggs, Kember & Leung, 2001). It is also possible students do not have an approach to learning at all or like to use both approaches depending on what is most suitable in the given context. The main aim of chapter 2 was to gain more insight into the relationships between students' approaches to learning and the different components of problem-solving, according to Sugrue's taxonomy (Sugrue, 1993, 1995), that were being measured using multiple-choice questions. The students in our sample showed slightly higher scores for a deep approach than for a surface approach to learning. Biggs et al. claims that a high quality learning environment will stimulate students to use a deep approach to learning. Students preferring or using surface approaches to learning in problem-based learning are not stimulated in a sufficient way by means of the tutors, the tasks, the fellow students and the assessment. Plotting students' approaches to learning indicated that in our study only a few students used surface approaches to learning. A lot of students had high scores for deep approach to learning and low scores for surface approach. Most likely these students recognised the key factors of a constructivist learning environment and perceived the learning environment as effective to use deep approaches to learning. However, another, substantial, group of students had low scores for both deep and surface approaches. Probably these students did not

perceive the learning environment as sufficient constructivist and therefore were not encouraged by the learning environment to use a deep approach to learning. Having no approach to learning at all, means that students depend heavily on guidance from the facilitators in the learning environment (tutors and/or the fellow students). One can question if this is a typical side effect of problem-based learning for students lacking metacognitive skills and not knowing how to cope with the learning environment. Or is it the proof of a very weak student population, lacking self-regulating and metacognitive skills? Nevertheless, this means these students have to be detected in an early stage to prevent them from struggling throughout their study. In order to promote appropriate study behaviour these students have to discover and develop appropriate study skills and strategies and learn to apply it. Whether tutors were able to detect inappropriate student behaviour by means of observations in the tutorial group was one of the research questions in chapter 6.

Looking at the relationship between students' approaches to learning and the performance on the different components of problem-solving being measured within the assessment we did not find any significant results. The assessment was not capable of differentiating between students using a different approach to learning. Several conclusions were drawn upon this result. Firstly, it is possible that Sugrue's model of the cognitive components of problem solving is not suitable for the domain of law. As Sugrue (1995) already remarked herself the model lends itself extremely well to domains such as science, mathematics, economics and geography, but other domains such as history might be more complicated to use the model. Law might be another domain in which the use of her taxonomy is more complicated. Secondly, may be it is not possible to identify relationships between study approach and study outcomes in terms of components of problem solving. According to Haladyna (2004) techniques in exposing subscales reflecting cognitive processes are not satisfactory: Item constructors may agree about the intended type of cognitive process the item is supposed to measure, it may measure an entirely different cognitive process simply because a student can experience the item in a different way than the item constructors and other students. But, the most important conclusion is the notion that the assessment method and question format has probably more influence on the way students study for, and respond to, the assessment questions (Minbashian, Huon & Bird, 2004). The perception of the assessment plays a much bigger role than the perception of the learning environment only as was reported by Segers, Dochy and Cascallar (2003), Scouller (1998) and Struyven, Dochy and Janssen (2003). This notion about the importance of perceptions of assessment permits more explorative studies investigating the influence of perception of assessment on study approaches, learning outcomes and performances in different contexts to be conducted as it was done and described in the chapters 3 through 6 in this dissertation.

The role of preferences and perceptions of assessment

The chapters 3 through 6 centred on the perception of assessment. As pointed out in chapter 2, finding no direct significant relationship between study approaches and assessment outcomes, it is likely that the assessment and question format

influences the way students study. The perception of the assessment system plays a role, probably a bigger role than the perception of the learning environment only, in the way students behave. As a consequence, these perceptions influence their outcomes. Chapter 3 had two purposes. The first purpose was to gain more insight into students' actual assessment preferences and perceptions of assessment. The second purpose was to explore the effects of these preferences and perceptions on the students' assessment outcomes. The assessment in this study was very traditional: a combination of multiple-choice and essay questions, measuring a variety of knowledge and cognitive skills. In chapter 6 the modification of the traditional assessment by the implementation of assessment-tasks is described as to bring the assessment more in line with the learning environment. The influence of this implementation on students' performances and on students' and teachers' perceptions was investigated by means of two research questions: First, do students who make the assessment tasks do better in their final exam compared to students who do not? And second, what are the most important concerns in students' and teachers' perceptions of the assessment tasks?

Chapter 4 has two parts. The first part is a review of research on teachers' and students' perceptions on item difficulty. The second part is an empirical study of the ability of students and teachers to estimate item difficulty correctly. In chapter 5 teachers observed their students during a course in a problem-based learning setting and categorised these students at the end of the course into four distinct groups perceived as barely competent, moderately competent, highly competent, and students with high anxiety. The purpose of this research was to investigate to what extent the assessment discriminates between less competent and more competent students based on the teachers' perceptions.

Students' assessment preferences and perceptions of assessment

From different studies regarding assessment preferences it seems that students prefer assessment formats which reduce stress and anxiety and it is assumed that students will perform better on their preferred assessment formats. Though the population in our study in chapter 3 preferred traditional written assessment (rather with the use of supporting material), we did not find a positive relationship between this assessment format and the assessment outcomes. Neither did we find a relationship between assessment preferences and perceptions of assessment.

The assessments in the studies were generally, as intended by the item constructors and the assessment composers, perceived more as assessing the application of knowledge, problem solving, the drawing of conclusions, and analysing and interpreting, than assessing the reproduction of knowledge. Despite this, in only 40% of the cases there was a clear correspondence between the intended level of cognitive processes and the perceived level of processes in the assessment. From these findings it can be concluded that on the one hand it is possible for students to have a clear picture of the assessment demands. On the other hand, the presence of multiple-choice questions in the assessment could have caused the overestimating of the reproduction of knowledge. Despite of the fact the multiple-choice and essay questions in the assessments emphasise on higher levels

of cognitive processing, students seems to focus on reproduction of knowledge especially. An explanation might be that teachers do stress the reproduction of knowledge. Of course students' perceptions are based on previous experiences. The most efficient way to assess reproductive knowledge is with the help of multiple-choice questions. And often, if multiple-choice questions are used, that is exactly what multiple-choice questions are used for: to measure reproductive knowledge. As discussed by Scouller and Prosser (1994), this leads to a strong association between multiple-choice examinations and the employment of surface learning approaches, and leading to successful outcomes.

We found, though not statistically significantly, that students with matching perceptions scored slightly better on the total assessment, compared to students with misperceptions. Probably because a lot of students have a mismatching perception of the cognitive processes measured by the assessment we did not find any direct relationships between students' perceptions of assessment and their assessment results. It is also possible that students' approaches to learning moderate the relationships between students' perceptions and their assessment results. From chapter 2, we know that a substantial proportion of the student population in the described problem-based learning environment do not use appropriate approaches to learning. Qualitative data about the course and the assessment revealed some insights into how they perceived the assessment. For some students it was unimaginable to use cases and problem scenarios in a multiple-choice assessment. As a consequence, when preparing for the assessment, these students did not prepare for applying knowledge, meaning that they did not practice their problem solving skills. Students also seem to identify certain questions as being reproduction based, because the questions or the cases used in the assessment resembled the tasks or cases used in the preceding educational programme. These results imply that a lot of students need help in building up a matching perception of what is assessed by means of the assessment formats that are used. Just giving help using examples of assessment items and discussing their answers seems not to be enough for these students. This again might be a sign of a very weak student population entering this faculty. The purpose of the assessment questions and the cognitive processes to be used for answering the question correctly must also be made clear and preferably be practiced for such students (which might turn out or does actually lead to undesirable teaching to the test). This is especially true in cases where a multiple-choice format is used to measure higher levels of cognitive processes, rather than reproducing knowledge.

Certainly the relationships between study approaches and perceptions of differently implemented assessment formats in different learning environments should be further explored, in combination with the findings of Scouller and Prosser (1994), Birenbaum and Feldman (1998) and Scouller (1998) about the relationships between study approaches and perceptions of assessment. This should give more insight in the effectiveness of aligning the learning environment and the assessment system.

Students' and teachers' perception of item difficulty

In chapter 4 the students' and teachers' perceptions of the difficulty levels of the assessment items were subject of study. This research was aiming at gaining more insight into the degree to which assessments are correctly aimed at in terms of level of competence in relation to the defined learning goals. The chapter contained a review of previous research into teachers' and students' perceptions of item difficulty and an empirical study, investigating the accuracy of teachers' estimations and students' perceptions of difficulty.

The review revealed that the outcomes of previous research into teachers' perceptions of item difficulty are not consistent. Most research shows that estimating the difficulty of items or assessment standards is a difficult job. Teachers tend to overestimate the difficulty of easy items and underestimate the difficulty of difficult items. The accuracy of the estimates can be improved by the information the estimators (judges, teachers, assessment constructors) have about the target group (total group of students, borderline group students, average students) and former assessment results, defining the target group before the estimation process, the possibility of having discussions amongst the estimators about the defined target group of students and its corresponding standards during the estimation process, the amount of training in item construction, and practice with estimating. In contrast, students seem to be better estimators of item difficulty.

Our findings from the empirical study regarding the teachers' perceptions of the difficulties of items showed that the teachers' estimations of the difficulty of the complete assessments are appropriate. Item constructors, assessment composers and reviewers were, however, able to estimate the difficulty level correctly for only a small proportion of the assessment items. For the three assessments, about one third of the items was estimated accurately. Hence, one could raise questions about the reliability and validity of the use of performance standards and standard setting procedures. For most items the students' performances were overestimated by the teachers; this means that most items were more difficult for students than expected by the teachers. This finding is in line with the studies of Shepard (1995), Impara and Plake (1998) and Goodwin (1999) in which the competence of teachers to estimate the item difficulty level of the students was questioned. Judges in these studies, as in our study, tended to overestimate the performances on assessment for the total group of students. Three possible explanations were formulated to describe this finding: One possible explanation for this overestimation is the expertise of the teachers. According to Goodwin, judges are typically experts in their fields. Because they might know too much, they cannot put themselves in the place of students adequately. Also, their expectations of the examinees are possibly too high. They may have difficulty in differing between the proportion of examinees who should have answered an item correctly and who could have answered an item correctly. The teachers in our study gave the same explanation. Their expectations were probably too high and, as a consequence, it was difficult to envision the skills and competences of the students. A second explanation is that teachers did not expect the educational programme to have a great influence on the learning behaviour of the students, such as reported by the students in the focus group

interviews. More discussion and training could have contributed to more accurate estimations of the item difficulties. A third explanation of these results is that the item constructors and assessment composers seem to focus too much on the high scoring students during the construction process of the assessment. This implies that, in the case of teacher-made assessments, to prevent disappointing student outcomes, item constructors and assessment composers together have to envision and describe the average student in terms of expectations first. Subsequently they should focus on the assessment construction process, keeping this average student in mind.

The findings regarding the students' perceptions of the item difficulties showed that, in contrast to the teachers, students underestimated their own performances. Strangely, students mainly underestimated their own performances on the easiest items. A possible explanation of this finding is that students may have different perceptions of the concept of item difficulties compared to those of teachers. In the interviews, students reported that the presentation of an assessment item was important in their estimation of its difficulty level. Teachers, on the other hand, did not see that as an important factor in their determination of the difficulty level of an assessment item. A second explanation is that students hoped that, if they indicated the assessment items as difficult ones, teachers would grade the assessment less severely.

Examining the relationships between the students' perceptions of the difficulty levels of the assessment items and their performances on the assessments, results show that the students who performed well on the assessments underestimated their performances the most. This finding is in line with earlier research which stated that the more prior knowledge a student has, the more he tends to underestimate his performance (Dochy, Segers & Buehl, 1999). Additionally, the focus group interviews indicated a possible misperception of the item difficulty caused by a misperception of the cognitive processes students had to use in solving the items: as we already concluded in chapter 2 and 3 some students tried to solve the items using incorrect strategies caused by a misperception which probably also moderate the perception of the item difficulty. The influences of the perception of the assessment, and the learning strategies that were used, on the students' perceptions of the item difficulties should also be subject of further research.

Teachers' perception of learning outcomes

In chapter 5 teachers observed their students during a course in a problem-based learning setting and classified these students at the end of the course into four distinct groups perceived as barely competent, moderately competent, highly competent, and students with anxiety. The main aim of this study was to investigate to what extent different parts of the assessment, multiple-choice questions and essay questions, discriminates between less and more competent students based on the teachers' perceptions. Also the validity of the observations in the course was questioned in this study: were the results course specific or could the same findings also be found in previous courses?

Results indicate that the performances on the multiple-choice questions and the essay questions show the same trend, i.e. low performance for the barely competent students, moderate performance for the moderately competent and anxious students and high performance for the highly competent students. Students that are characterized as highly competent by the teachers performed better on both parts (multiple-choice questions and essay questions) of the assessment in comparison with the other student types. The barely competent students performed worse on both parts of the assessment. From these results it can be concluded that the separate parts of the assessment discriminate in an appropriate way between the different groups of students. Secondly, there are no real differences in performance between the two assessment parts for the different group of students. Thirdly, it has also become clear that the perception of teachers on the performance of students (as operationalised in the different student types) is in line with Jussim's observations (Jussim, 1991) and corresponds with the actual performance of those students.

The results suggest that, compared to the subsequent assessments the first assessment discriminates differently. On the one hand it could be possible that students were still looking for a suitable study method or were not used to answer essay questions in a manner satisfactory for themselves and the assessors. On the other hand it could be that assessment effects play a role in this phenomenon. A possible explanation for this result may lie in the so-called pre-assessment effects (Gielen, Dochy & Dierick, 2003). With this concept the actual influence of the assessment on the study methods of the student, in particular his preparation for that assessment is indicated. Barely competent students presumably perceive the assessment mainly as a multiple-choice assessment since the majority of the points awarded can be gained in the multiple-choice part of the assessment. According to Boud (1990) students focus on those subjects and learning levels that are assessed and earn them, according to them, points. Obviously, there is a discrepancy between what is actually tested and what they expect to be tested (Broekkamp, 2002). If we combine the conclusions from our study with the conclusions reached by Birenbaum and Feldman (1998), students with a surface study approach tended to prefer multiple-choice formats, and Scouller (1998), students are more easily enticed to use a surface approach to learning if an assessment merely consists of multiple-choice questions, we can conclude that the barely competent student will have a preference for multiple-choice questions and will employ a surface study approach. The surface strategy will lead to this preference for a surface approach. This is a cyclic process where this type of student, regarding his previous assessment performance, will not escape. Also in the case of the students categorised as 'average' or 'cognitively strong' but with characteristics of fear of failure or displaying inappropriate study methods there is a change in the perception of students to the assessment, which similarly influences the result of the assessment. After their experience with the first assessment, where this type of student achieves a better result on the essay questions than the other types, the student changes his perception of the assessment. Although they score high on the essay questions they are probably disappointed by the result of the assessment as a whole to such a degree that they replace the emphasis of their preparation to the multiple-choice questions. This occurrence can also be supported with the views of

Boud (1990): students are prepared to change their study approach if they believe their assessment results will improve. It seems that this type of student changes his study approach to an approach that is more directed towards reproduction and therefore to a surface approach. In addition to that they apply false study methods (not distinguishing between core matters and more peripheral matters). This can explain their drop in assessment performance in the following assessments. This phenomenon could be identified as a negative influence of the pre-assessment effects on study behaviour. On the other hand, the highly competent student is less identifiable in the results of the first exam. Only after the first assessment they realize that this form of assessment requires more than reproduction and as a consequence they change their study approach, which leads to better results on both the multiple-choice as the essay part of the combination exam. In this case one can identify a positive influence of the pre-assessment effect.

Adapting the assessment towards a more constructivist alignment

In chapter 6 the purpose of the study was to get more insight into the effects of written assessment-tasks integrated in a problem-based learning environment. The implementation of assessment-tasks is considered to be a necessary move, in line with constructivist learning theories, to improve learning and to align assessment more with the learning environment. To bridge the gap between learning and assessment Gibbs and Simpson (2003) presented a framework to improve learning by making principle changes to assessment. These changes to assessment focus on student effort and on feedback. First, the assessment should be used to influence the quantity and distribution of student effort. Second, the quality and the level of the students' effort should be influenced by the assessment. The third dimension stresses the quantity and timing of the feedback. Fourth, the quality of feedback is important. Fifth, students' response to feedback should be taken into consideration. Our design of the assessment-tasks was to a great extent in line with this framework. During the course students worked in a problem-based learning setting on different (problem) tasks every week. Six of these tasks were been promoted assessment-tasks. With each of the six assessment-tasks, the students were asked to write an essay in which they had to go through a process of evaluation, synthesis, analysis or understanding of the study material or had to write down in detail the solution of a problem as presented in a case. In order to succeed, each of the six assessment-tasks had to be of sufficient effort and correctness. Students were stimulated to produce qualitative learning activities by giving an extra 'bonus-point' for the final assessment only if all six assessment-tasks showed to be of sufficient quality and effort. The tasks were discussed in the same way as the other problem tasks in the course after the students handed in their assessment-task. In this way the students got some plenary feedback. One week later, the students received back their assessment-task from the teacher, with the necessary individual feedback. The feedback students got from their teacher or from the plenary discussion in the tutorial group could help them to make their next assessment-task better and to get a better understanding of the learning materials to be studied in order to pass the final

exam. The assessment-tasks emphasised on cognitive processes such as analysing, integrating, synthesising, evaluating and problem solving. With the help of the tasks these processes were assessed more extensively than with for example a final exam only. Quantitative data, by means of the assessment outcomes and questionnaires consisting of closed question, and qualitative data, by means of semi-structured interview and questionnaires consisting of open questions, were used to examine the effects of the implementation of the assessment-tasks.

The quantitative results showed that, when corrected for differences in prior academic performance, students who did make all the assessment-tasks successfully during the course, performed better on their final exam compared to students who did not make nor succeed the assessment-tasks. The fact that students did not only perform better on those parts of the final exam which were related to the assessment-tasks, but also on non-related questions indicates that the introduction of the assessment-tasks helped students to address more appropriate learning activities, going beyond the assessment-tasks and its content. Both students and teachers perceived the assessment-tasks as meaningful. According to them the tasks stimulated some desired learning activities. Students found themselves more capable in the case of problem solving processes, critical thinking and reasoning. In the view of the tutors the use of assessment-tasks especially stimulated independent activities and reasoning. Students reported to study more, more critically and more systematically as a result of the assessment tasks. It is suggested that students with a better time-management perform slightly better on their final exam.

The qualitative data suggest that both the students and the teachers see the benefits of the assessment-tasks. Working with the assessment-tasks was stimulating to work more intensively with the learning material. As a consequence the students put more effort in their work and were better prepared for participating in the tutorial groups. According to the teachers this resulted in more in-depth discussions. Although adapting the assessment using assessment-tasks led to positive results two issues related to feedback have to be taken into consideration: the criteria used and feedback on students' performances. In the beginning the students had some problems with the criteria of the assessment tasks. They were not clear and not concrete enough, so the students did not know what was expected from them. According to the framework (Gibbs & Simpson, 2003) the feedback should be linked to the purpose of the tasks and the criteria. If the criteria are not clear the feedback will not be so effective. By making the criteria clear in advance the feedback on students' performances will be more understandable for the students and can be linked to learning more easy. For an optimal effect of the assessment-tasks the teachers thought they had to give feedback on every task to the students personally. And although the students reported to find the personal feedback useful, for many teachers this led to time-management problems. The presence of a correction model was helpful, but in the future the use of self and peer assessment can be considered to the benefit of the learning of the students and the time-management of the teachers. Also co- assessment in which tutors and students formulate criteria together is an option to be considered. Despite the fact there were some limitations, the study supported the suggestion that small steps in the change

of the assessment system can result in relatively big changes in students' learning and results.

Implications for practice

Students participating in a constructivist learning environments do not always demonstrate the expected approaches to learning. As mentioned in the introduction, the effects of implementation of constructivist learning environments, such as problem-based learning, do not always demonstrate the expected outcomes (Birenbaum, 2000; Segers & Dochy, 2001). There are two reasons for this: because the assessment and the learning environment are not aligned (Biggs, 2001), and because students have incorrect perceptions of the learning environment or the assessment system (Lawness & Richardson, 2002).

From our studies in this dissertation it can be concluded that a substantial part of the student population did not recognise (all) constructive key factors. A lot of students did not experience problem-based learning as student-centred. Moreover, a substantial part of the students did not use any approach to learning at all. These students depend on the guidance of the facilitators in the learning environment. Two considerations emerge from these conclusions. Firstly, indeed the potential of problem-based learning as mentioned by Savin-Baden (2000) is not fully realised. Secondly, indeed a problem-based learning environment does not always promote a deep approach to learning. Both considerations put an emphasis on the importance of a well functioning, cooperative tutorial group that promotes meaningful knowledge construction for all students. The role of the teacher is important to facilitate students to pursue their personal learning goals and to cope with their learning difficulties. In order to achieve this the teachers have to focus on students' approaches to learning, study skills and strategies. From our study, it is also clear that teachers are able to identify students with inappropriate study methods during interactions in the tutorial groups. Assessment-tasks can facilitate teachers to identify these students and to focus students on appropriate study approaches.

According to our studies the students' perceptions of assessment do differ from teachers' perceptions of item difficulty and the cognitive processes. This means on the one hand that students do not have a clear picture of the assessment demands. On the other hand it seems that assessment composers do not have a clear picture of the average student's performance standard. To prevent disappointing student outcomes, item constructors and assessment composers together have to envision and describe the average student in terms of expectations first. They should then focus on the assessment construction process, keeping this average student in mind. For students it is necessary to build upon a matching perception of what is assessed by means of the assessment format that is used. Giving help using examples of items and discussing their answers is not enough. The purpose of the assessment questions and the students' cognitive processes to be used for answering the questions correctly must also be made clear and preferably be practised in the tutorial groups. Since this is obviously not the case, one can question the alignment

between the activities in the tutorial group and the assessment. In this case assessment-tasks can also help students to get a clear picture of the assessment demands as intended by the assessment composers.

Suggestions for further research

Though the concepts in the different studies in this dissertation are related to each other, as presented in Figure 7.1, the studies in this dissertation are loosely coupled. This means that the conclusions evoke some issues for further research:

- Further research on constructivist learning environments should focus on the engineering of an optimal mix of learning environments and take into account students' perceptions of the blend of lectures, problem- and case-based learning groups, practical work, task-oriented learning, workplace learning, online learning opportunities, etc. and their approaches to learning
- Further research should probe into the relationship between students' approaches to learning and their outcomes on and perceptions of a blend of assessment methods.
- The influences of the perception of assessment and students' learning strategies on the students' perceptions of the item difficulties should be subject of research.
- Further study could be aimed at how different groups of students perceive the link between teaching and assessment; Whether the different groups of students perceive this integration differently could also provide interesting information for this field of study.

References

Biggs, J. (2001). The reflective institution: Assuring and enhancing the quality of teaching and learning. *Higher Education*, 41, 221-238.

Biggs, J., Kember, D., & Leung D.Y.P. (2001). The revised two-factor study process questionnaire: R-SPQ-2F. *British Journal of Educational Psychology*, *71*, 133-149.

Birenbaum, M. (2000). *New Insights into Learning and Teaching and the Implications for Assessment.* Keynote address at the 2000 conference of the EARLI SIG on Assessment and Evaluation, September 13, Maastricht, The Netherlands.

Birenbaum, M., & Feldman, R.A. (1998). Relationships between learning patterns and attitudes towards two assessment formats, *Educational Research*, *40* (1), 90-97.

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

Broekkamp, H. (2002). *Task demands and test expectations* (Dissertation). Amsterdam: Universiteit van Amsterdam.

Dochy, F., Segers, M., & Buehl, M. (1999). The Relation Between Assessment Practices and Outcomes of Studies: The Case of Research on Prior Knowledge. *Review of Educational Research*, *69* (2), 147-188.

Gibbs, G., & Simpson, C. (2003). Does your assessment support your students' learning? *Learning and Teaching in Higher Education*, *1* (1), Retrieved December 5, 2005, from http://www.open.ac.uk/science/fdtl/documents/

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.), *Optimising New Modes of Assessment: In Search of Qualities and Standards* (pp. 37-54). Dordrecht: Kluwer Academic Publishers.

Goodwin, L. D. (1999). Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline examinees. *Applied Measurement in Education*, *12* (1), 13-28

Haladyna, T.M. (2004). *Developing and validating multiple-choice test items* [3rd ed.]. Mahwah, NJ: Lawrence Erlbaum Associates.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, *35*, 69-81.

Jussim, L. (1991). Social perception and social reality: A reflection-construction model. *Psychological Review*, *98*, 54–73.

Lawness, C.J., & Richardson, J.T.E. (2002). Approaches to studying and perceptions of academic quality in distance education. *Higher Education*, *44*, 257-282.

Minbashian, A., Huon, G.F., & Bird, K.D. (2004). Approaches to studying and academic performance in short-essay exams. *Higher Education*, *47* (2), 161-176.

Savin-Baden, M. (2000). *Problem-based Learning in Higher Education: Untold Stories.* Buckingham: SRHE and Open University Press.

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

Scouller, K., & Prosser, M. (1994). Students' experiences in studying for multiple choice question examinations. *Studies in Higher Education*, *19*, 267-279.

Segers, M., & Dochy, F. (2001). New assessment forms in Problem-based Learning: the value-added of the students' perspective. *Studies in Higher Education*, *26* (3), 327-343.

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

Shepard, L. A. (1995). *Implications for standard setting of the National Academy of Education Evaluation of National Assessment of Educational Progress Achievement Levels, Proceedings from the Joint Conference on Standard Setting for Large-Scale Assessments*. Washington: National Assessment Governing Board and National Center for Education Statistics).

Struyven, K., Dochy, F., & Janssens, S. (2003). Students' perceptions about new modes of assessment in higher education: A review. In M. Segers, F. Dochy &

E. Cascallar (Eds.), *Optimizing new modes of assessment: In search of qualities and standards* (pp. 171-224). Boston/Dordrecht: Kluwer Academic

Sugrue, B. (1993). *Specifications for the design of problem-solving assessments in science. Project 2.1 designs for assessing individual and group problem-solving.* Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Sugrue, B. (1995). A theory-based framework for assessing domain-specific problem solving ability. *Educational Measurement: Issues and Practice*, *14* (3), 29-36.

Tenenbaum, G., Naidu, S., Jegede, O. & Austin, J. (2001). Constructivist pedagogy in conventional on-campus and distance learning practice: an exploratory investigation. *Learning and Instruction*, *11*, 87-111.

Summary

It is believed that assessment has an important impact on learning and can be a powerful means to focus on learning. In order to have a positive effect from the assessment, it is important to align the learning environment and the assessment and consequently enhance appropriate learning activities. However, a fit between the learning environment and the assessment alone is not effective enough. From several studies it is known that preferences and perceptions have its impact on how students cope with the learning environment, the assessment and the learning outcomes. This dissertation focuses on students' and teachers' perceptions of assessment and the relationship between these perceptions and learning outcomes. The studies took place in a constructivist learning environment, as designed into problem-based learning (PBL) environment at the faculty of law of the University of Maastricht.

This dissertation consists of six studies. Firstly, an introduction (overview, structure and presentation of the research questions) to the different studies in the dissertation was given.

The first study in chapter 1 was concerned about students' perceptions of their learning environments. Students in a problem-based learning environment and in a conventional lecture-based environment were asked, using a questionnaire consisting of seven key factors of constructivist learning environments, about their experiences of the educational practice. The main aim of the first study was to verify whether students in a problem-based learning environment perceive the learning environment to be more constructivist, when compared to the perceptions students have of a conventional lecture-based environment. A question of particular interest in this study was for which factors the differences between the problem-based learning environment, and the conventional lecture-based environment, are the largest. Results showed that students participating in problem based learning perceived their learning environment to be more constructivist in case of the key factors 'conceptual conflicts and dilemmas', 'sharing ideas with others', and 'arguments, discussions, debates'. The main characteristics of the first two factors, confronting students with conceptual conflicts and learning is a cooperative process, determine the strength of problem based learning in incorporating constructivist principles. The factor 'meeting students' needs' was only moderately present in the

problem based learning environment. Meeting students' needs is about issues such as students' satisfaction with their learning outcomes, the possibility for students to pursue personal goals, to benefit from their own learning difficulties, and to what extent the faculty takes students needs and concerns into consideration. The moderate presence of this factor indicates that students in this learning environment only had a relatively small say in the learning process. As a consequence one of the opportunities, which will be more in line with the constructivist principles, is to make sure that learning is more student centred, to make sure that students are in a larger extent owners of their learning process.

In chapter 2, students' approaches to learning in a problem-based learning environment, and the students' learning outcomes, were the subject of study. The students' approaches to learning were measured with the revised two factor study process questionnaire (R-SPQ-2F). In order to distinguish between students' assessment outcomes on different levels of the knowledge structure, a model of cognitive components of problem-solving was used. The purpose of the study was to explore further the relationship between students' approaches to learning and their quantitative learning outcomes, as measured by the assessment, from the perspective of the different components of problem-solving. A lot of students had high scores for deep approach to learning and low scores for surface approach. Most likely these students recognised the key factors of a constructivist learning environment and perceived the learning environment as effective to use deep approaches to learning. However, another, substantial, group of students had low scores for both deep and surface approaches. Probably these students did not perceive the learning environment as sufficient constructivist and therefore were not encouraged by the learning environment to use a deep approach to learning. Having no approach to learning at all, means that students depend heavily on guidance from the facilitators in the learning environment (tutors and/or the fellow students). Looking at the relationship between students' approaches to learning and the performance on the different components of problem-solving being measured within the assessment we did not find any significant results. The assessment was not capable of differentiating between students using a different approach to learning. It is suggested that students' perception of assessment has probably more influence on the way students study for, and respond to, the assessment questions.

The third study, in chapter 3, had two purposes. The first was to gain more insight into students' actual assessment preferences and perceptions of assessment. The second was to explore the effects of these preferences and perceptions on the students' assessment outcomes. In this study the assessment combined two assessment formats and the question types were directed towards different cognitive process levels. Parts of the Assessment Preferences Inventory (API) were used to answer four research questions: firstly, which assessment preferences do students have? Secondly, how did students perceive the assessment? Thirdly, in what ways are students' assessment preferences related to their assessment results? And fourthly, in what way are students' perceptions of the assessment related to their assessment results? The students in our study preferred traditional written assessment rather with the use of supporting material. We did not find a positive relationship between this format preference and the assessment outcomes. In only

40% of the cases there was a clear correspondence between the intended level of cognitive processes and the perceived level of processes in the assessment. It is suggested that the presence of multiple choice questions in the assessment could have caused the overestimating of the reproduction of knowledge. Despite of the fact the multiple choice and essay questions in the assessments emphasise on higher levels of cognitive processing, students seems to focus on reproduction of knowledge especially. An explanation might be that teachers do stress the reproduction of knowledge. We found, though not statistically significantly, that students with matching perceptions scored slightly better on the total assessment, compared to students with misperceptions. Probably because a lot of students have a mismatching perception of the cognitive processes measured by the assessment we did not find any direct relationships between students' perceptions of assessment and their assessment results. It is also possible that students' approaches to learning moderate the relationships between students' perceptions and their assessment results. Qualitative data about the course and the assessment revealed that a lot of students need help in building up a matching perception of what is assessed by means of the assessment formats that are used.

The research in chapter 4 was aiming at gaining more insight into the degree to which assessments are correctly aimed at in terms of level of competence in relation to the defined learning goals. The chapter contained a review of previous research into teachers' and students' perceptions of item difficulty and an empirical study, investigating the accuracy of teachers' estimations and students' perceptions of difficulty. The central question to be answered is whether teachers and students perceive item difficulty correctly. In the empirical study four research questions were formulated: firstly, to what extent are teachers accurate in estimating the difficulty level of assessment items? Secondly, to what extent do students' perceptions of the difficulty level of the assessment items correspond with their actual difficulty levels? Thirdly, to what extent do the students' perceptions of the difficulty levels of the assessment items differ from the teachers' estimations of the item difficulties? And finally, what relationship exists between students' perceptions of the difficulty levels of the assessment items and their performances on the assessment? The review revealed that the outcomes of previous research into teachers' perceptions of item difficulty are not consistent. Most research showed that estimating the difficulty of items or assessment standards is a difficult job. Teachers tend to overestimate the difficulty of easy items and underestimate the difficulty of difficult items. In contrast, students seem to be better estimators of item difficulty. Our findings from the empirical study regarding the teachers' perceptions of the difficulties of items showed that most items were more difficult for students than expected by the teachers. The formulated explanations imply that, teachers should focus on the assessment construction process, keeping this average student in mind. The findings regarding the students' perceptions of the item difficulties showed that, in contrast to the teachers, students underestimated their own performances. Strangely, students mainly underestimated their own performances on the easiest items. Examining the relationships between the students' perceptions of the difficulty levels of the assessment items and their performances on the assessments, results show that the students who performed well on the assessments

underestimated their performances the most. Additionally, the focus group interviews indicated a possible misperception of the item difficulty caused by a misperception of the cognitive processes students had to use in solving the items.

In chapter 5 teachers observed their students during a course in a problem based learning setting and classified these students at the end of the course into four distinct groups perceived as barely competent, moderately competent, highly competent, and students with anxiety. The main aim of this study was to investigate to what extent different parts of the assessment, multiple choice questions and essay questions, discriminates between less and more competent students based on the teachers' perceptions. Also the validity of the observations in the course was questioned in this study: were the results course specific or could the same findings also be found in previous courses? From the results it can be concluded 1) that the separate parts of the assessment discriminate in an appropriate way between the different groups of students, 2) there are no real differences in performance between the two assessment parts for the different group of students, and 3) that the perception of teachers on the performance of students corresponds with the actual performance of those students. The results suggested also that, compared to the subsequent assessments the first assessment discriminates differently. It is suggested that assessment effects play a role in this phenomenon.

The purpose of the study in chapter 6 was to gain more insight into the effects of the implementation of written assessment tasks in a problem-based learning environment through quantitative and qualitative data. The implementation of assessment tasks is considered to be a necessary move, in line with constructivist learning theories, to improve learning and to align assessment more with the learning environment. The influence of the implementation of the assessment tasks on students' performances and on students' and teachers' perceptions were investigated by means of two research questions: firstly, do students who undertake the assessment tasks do better in their final exam compared to students who do not? Secondly, what are the most important concerns in students' and teachers' perceptions of the assessment tasks? The quantitative results showed that, when corrected for differences in prior academic performance, students who did make all the assessment-tasks successfully during the course, performed better on their final exam compared to students who did not make nor succeed the assessment-tasks. The fact that students did not only perform better on those parts of the final exam which were related to the assessment-tasks, but also on non-related questions indicates that the introduction of the assessment-tasks helped students to address more appropriate learning activities, going beyond the assessment-tasks and its content. Both students and teachers perceived the assessment-tasks as meaningful. The qualitative data suggest that both the students and the teachers see the benefits of the assessment-tasks. Working with the assessment-tasks was stimulating to work more intensively with the learning material. As a consequence the students put more effort in their work and were better prepared for participating in the tutorial groups. According to the teachers this resulted in more in-depth discussions. Although adapting the assessment using assessment-tasks led to positive results two issues related to feedback have to be taken into consideration: the criteria used and feedback on students' performances. The study supported the suggestion that small

steps in the change of the assessment system can result in relatively big changes in students' learning and results.

In chapter 8 the results from the studies in the dissertation were summarised and discussed. Also some suggestions for future practices and further studies were formulated. It can be concluded that a substantial part of the student population did not recognise (all) constructive key factors. A lot of students did not experience problem-based learning as student-centred and a substantial part of the students did not use any approach to learning at all. The role of the teacher is important to facilitate students to pursue their personal learning goals and to cope with their learning difficulties. In order to achieve this the teachers have to focus on students approaches to learning, study skills and strategies. Teachers are able to identify students with inappropriate study methods during interactions in the tutorial groups. Assessment-tasks can facilitate teachers to focus students on appropriate study approaches. Perceptions of assessment play an important role in the students' behaviour. Students' perceptions of assessment do differ from teachers' perceptions of item difficulty and the cognitive processes. Students do not have a clear picture of the assessment demands and the assessment composers do not have a clear picture of the average student's performance standard. Assessment composers should envision and describe the average student in terms of expectations first and then focus on the assessment construction process, keeping this average student in mind. For students it is necessary to build up a matching perception of what is assessed by means of the assessment format that is used. Assessment-tasks can also help students to get a clear picture of the assessment demands as intended by the assessment composers.

Samenvatting

Over het algemeen wordt aangenomen dat assessment het leren in hoge mate beïnvloedt. Assessment wordt zelfs gezien als een krachtig middel om studenten te richten op het leerproces. Daar zit wel een voorwaarde aan vast: om een positief effect te ervaren dient de leeromgeving en de assessment op elkaar afgestemd te worden. En dan nog zal een goede aansluiting alleen de studenten niet in voldoende mate aanzetten tot het ontplooien van de juiste leeractiviteiten. Diverse onderzoeken hebben namelijk al aangetoond dat voorkeuren en percepties op enigerlei wijze invloed hebben op de wijze waarop studenten omgaan met hun leeromgeving, hun assessment en hun leerresultaten. De onderzoeken in dit proefschrift verdiepen zich met name in de percepties die studenten en docenten hebben van hun toetsen en in de relatie tussen de percepties en de leerresultaten. Alle onderzoeken hebben plaatsgevonden in een constructivistische leeromgeving. Meer concreet in het probleemgestuurd onderwijs (PGO) zoals die vorm heeft gekregen op de Faculteit der Rechtsgeleerdheid van de Universiteit Maastricht.

Dit proefschrift bestaat uit 6 onderzoeken en begint met een introductie. De introductie bestaat uit een overzicht van de in het proefschrift opgenomen onderzoeken, de structuur van het proefschrift en de presentatie van de onderzoeksvragen.

Hoofdstuk 1 beschrijft het eerste onderzoek en heeft betrekking op de wijze waarop studenten hun leeromgeving percipiëren. Studenten in een probleemgestuurde leeromgeving en studenten in een meer conventionele leeromgeving, waarin het volgen van hoorcolleges centraal staat, werden met behulp van een vragenlijst bevraagd naar hun ervaringen in het gevolgde onderwijs. De vragenlijst bevatte vragen gericht op 7 kernfactoren van een constructivistische leeromgeving. Het voornaamste doel van dit onderzoek was nagaan in welke mate de studenten in een probleemgestuurde leeromgeving, in vergelijking met de studenten uit de meer conventionele leeromgeving, de leeromgeving als meer constructivistisch percipiëren. Bijzonder interessant was daarbij de vraag voor welke factoren het verschil het grootst is. Volgens de resultaten percipieerden de studenten in de probleemgestuurde leeromgeving de leeromgeving op een aantal factoren als meer constructivistisch. Dit was in geval van de kernfactoren

'conceptuele conflicten en dilemma's', 'ideeën met anderen delen' en 'betogen, discussies, debatten'. De belangrijkste kenmerken van de eerste twee factoren, het confronteren van studenten met conceptuele tegenstrijdigheden en leren is een coöperatief proces, bepalen in feite in welke mate de constructivistische principes zijn geïntegreerd in het probleemgestuurd leren. De factor 'ingaan op de wensen van studenten' werd door studenten in de probleemgestuurde leeromgeving als matig herkend. Deze factor heeft betrekking op aspecten zoals tevredenheid van studenten met hun leerresultaten, de mogelijkheid om je eigen leerdoelen na te streven, te leren van je eigen leermoeilijkheden en de mate waarin de opleiding rekening houdt met de wensen en belangen van studenten. Dat deze factor slechts matig herkend werd door de studenten geeft aan dat studenten in deze leeromgeving relatief weinig invloed konden uitoefenen op het leerproces. Gevolg is dat, wil men de leeromgeving meer in lijn te brengen met de constructivistische principes, men dient te zorgen dat het leren meer studentgecentreerd wordt, dat studenten zelf in grotere mate het leerproces kunnen beheren.

In hoofdstuk 2 werden de studieaanpak en de (leer)resultaten van de studenten onderzocht. De studieaanpak werd met behulp van een vragenlijst (de zogenaamde 'revised two factor study process questionnaire') gemeten. Een model van cognitieve componenten van probleemoplossen werd gebruikt om de assessmentresultaten op de verschillende niveaus van de kennisstructuur te kunnen onderscheiden. Het doel van het onderzoek was de relatie tussen de studieaanpak en de kwantitatieve leerresultaten, zoals die door middel van de assessment getoetst werd, vanuit het perspectief van de onderscheiden componenten van probleemoplossen in kaart te brengen. Uit de resultaten blijkt dat een groot deel van de studenten een grondige studieaanpak hanteren. Waarschijnlijk herkennen deze studenten de kernfactoren van een constructivistische leeromgeving. Voor deze studenten is een grondige studieaanpak een effectieve methode om te gebruiken in de gepercipieerde leeromgeving. Echter, een ander substantieel deel van de studenten heeft geen studieaanpak (scoort laag op zowel een grondige studieaanpak als oppervlakkige studieaanpak). Deze studenten lijken de leeromgeving als niet voldoende constructivistisch te percipiëren en werden niet door de leeromgeving gestimuleerd om een grondige studieaanpak te hanteren. Zonder studieaanpak zijn deze studenten erg afhankelijk van anderen in de leeromgeving (tutoren en/of medestudenten). De assessment zelf was niet in staat te differentiëren tussen studenten met een verschillende studieaanpak. We konden geen verband aantonen tussen de studieaanpak en de prestatie op de verschillende componenten van probleemoplossen, zoals gemeten door middel van de assessment. Gesuggereerd wordt dat de wijze waarop studenten de assessment percipiëren waarschijnlijk meer invloed uitoefent op de wijze van studeren en de wijze waarop de studenten omgaan met de assessmentvragen.

Het derde onderzoek in hoofdstuk 3 had twee doelstellingen. Ten eerste wilden we meer inzicht krijgen in welke voorkeur studenten hebben voor asssessmentvormen en de wijze waarop studenten hun assessments percipiëren. Ten tweede wilden we de effecten van deze voorkeuren en percepties op hun

assessmentresultaten onderzoeken. De assessment bestond in dit onderzoek uit twee verschillende assessmentvormen en waren de vraagtypen gericht op verschillende niveaus van het cognitieve proces. Delen van de zogenaamde 'assessment preferences inventory' werden gebruikt om vier onderzoeksvragen te beantwoorden. De studenten in dit onderzoek gaven aan een voorkeur te hebben voor traditioneel schriftelijke assessments. Het liefst met de mogelijkheid gebruik te maken van ondersteunend materiaal. We hebben geen positieve relatie gevonden tussen een voorkeur voor een assessmentvorm en het resultaat op de assessment. Bij slechts 40 % van de assessmentdeelnemers was er een duidelijke overeenkomst tussen het door de toetsconstructeurs veronderstelde niveau van de cognitieve processen en de gepercipieerde cognitieve niveaus in de assessment. Gesuggereerd wordt dat de aanwezigheid van meerkeuzevragen in de assessment bij studenten tot overschatting van het reproduceren van kennis heeft geleid. Ondanks het feit dat de meerkeuzevragen en de essayvragen in de assessment de nadruk op hogere niveaus van het cognitieve proces legden, lijken de studenten zich vooral te richten het reproduceren van kennis. Een verklaring kan zijn dat docenten het reproduceren van kennis te veel benadrukken. We vonden dat, hoewel niet statistisch significant, studenten met een overeenstemmende perceptie iets beter op de totale assessment scoorden in vergelijk met studenten met verkeerde percepties. Waarschijnlijk omdat veel studenten verkeerde percepties van de te hanteren cognitieve processen bij het oplossen van de vraagstellingen hebben, vonden we geen directe relatie tussen hun percepties en hun studieresultaten. Het is ook mogelijk dat de studieaanpak van studenten een moderatorvariabele is en de relatie tussen percepties en assessmentresultaten beïnvloedt. Uit de kwalitatieve data (de evaluatie van de cursus en de assessment) werd duidelijk dat veel studenten hulp nodig hebben bij het vormen van een juiste perceptie over wat getoetst wordt door de gebruikte assessmentvormen.

De doelstelling van hoofdstuk 4 was meer inzicht te krijgen in welke mate de assessments zich op een juiste wijze richten op het, in de leerdoelen gedefinieerde, competentieniveau. Het hoofdstuk bevat een overzicht van eerder onderzoek en een empirisch onderzoek. Het overzicht heeft betrekking op eerder onderzoek naar de wijze waarop docenten en studenten de moeilijkheidsgraad van vragen percipiëren. Het empirisch onderzoek onderzocht de nauwkeurigheid van de inschatting van docenten en de perceptie van studenten ten aanzien van de moeilijkheid van assessmentvragen. De kernvraag is of docenten en studenten de moeilijkheidsgraad van een vraag op een juiste wijze percipiëren. In het empirisch gedeelte werden vier onderzoeksvragen geformuleerd. Ten eerste: in welke mate kunnen docenten een nauwkeurige inschatting maken van de moeilijkheidsgraad van een assessmentvraag? Ten tweede: in welke mate komen de door de studenten gepercipieerde moeilijkheidsgraad overeen met de werkelijke (statistische) moeilijkheidsgraad? Ten derde: in welke mate verschilt de door de studenten gepercipieerde moeilijkheidsgraad met de inschatting van de docenten? Ten slotte: is er een verband tussen de door de studenten gepercipieerde moeilijkheidsgraad en hun prestaties op de assessment?

Uit het overzicht blijkt dat de resultaten uit eerder onderzoek met betrekking tot de door docenten gepercipieerde moeilijkheidsgraad niet eenduidig zijn. De meeste onderzoeken concludeerden dat het inschatten van de moeilijkheidsgraad van vragen of de assessmentstandaarden een moeilijke klus is. Docenten hebben de neiging de moeilijkheid van makkelijke vragen te onderschatten en de moeilijkheid van moeilijke vragen te overschatten. Studenten blijken beter in staat te zijn de moeilijkheidsgraad van een vraag in te schatten. Uit onze bevindingen uit het empirisch gedeelte van het onderzoek, met betrekking tot de door docenten gepercipieerde moeilijkheid van de vragen, bleek dat de meeste vragen voor studenten moeilijker waren dan door de docenten werd ingeschat. Uit één van de in het hoofdstuk beschreven verklaringen volgt dat docenten tijdens het constructieproces zich meer zouden moeten richten op de gemiddelde student. In tegenstelling tot de perceptie van docenten onderschatten studenten hun eigen prestaties. Vreemd genoeg onderschatten de studenten vooral hun prestaties op de makkelijke vragen. De resultaten op de laatste onderzoeksvraag laten zien dat de 'betere' studenten (studenten met hogere scores) hun prestatie het meest onderschatten. De aanvullende interviews gaven als mogelijke verklaring aan dat een verkeerde perceptie van de moeilijkheidsgraad veroorzaakt werd door een verkeerde perceptie van de te gebruiken cognitieve processen bij het oplossen van de vraagstukken in de assessment.

In hoofdstuk 5 werd aan docenten gevraagd hun studenten (in een probleemgestuurde leersetting) te observeren en vervolgens aan het eind van het blok te classificeren. Deze classificatie had betrekking op het competentieniveau van de studenten: onvoldoende competent, voldoende competent, zeer competent en faalangstige studenten. Het doel van het onderzoek was na te gaan in welke mate de verschillende onderdelen van de assessment in staat was te discrimineren tussen meer en minder competente studenten. Ook werd de validiteit van de observaties onderzocht door middel van de volgende onderzoeksvraag: zijn de resultaten blokspecifiek of worden in de voorgaande blokken dezelfde resultaten gevonden? Uit de resultaten volgt dat 1) de afzonderlijke onderdelen van de assessment op een juiste wijze discrimineren tussen de onderscheiden groepen studenten, 2) er zijn geen grote verschillen in prestatie tussen de twee assessmentonderdelen voor de onderscheiden groepen studenten, en 3) de door docenten gepercipieerde prestatie van studenten komt overeen met de prestatie van de studenten op hun assessment. De resultaten geven aan dat de eerste assessment op een andere wijze tussen groepen studenten discrimineert dan de daarop volgende assessments. Assessmenteffecten lijken hierbij een belangrijke rol te spelen.

Het doel van het onderzoek in hoofdstuk 6 was meer inzicht krijgen in de effecten van de implementatie van schriftelijke assessmenttaken in een probleemgestuurde leeromgeving. Er werd kwantitatieve en kwalitatieve data gebruikt om het effect aan te tonen. De implementatie van assessmenttaken wordt gezien als een noodzakelijke stap, meer in lijn met de constructivistische leertheorieën, om het leren te verbeteren en om de assessment beter aan te laten sluiten op de leeromgeving. De invloed van de implementatie van de

assessmenttaken op de prestatie van studenten en de percepties van docenten en studenten werden onderzocht door middel van twee onderzoeksvragen. Ten eerste: presteren de studenten die de assessmenttaken maken in vergelijking met de studenten die deze taken niet maken, beter op de assessment aan het einde van het blok. Ten tweede: wat zijn de meest belangrijke aspecten ten aanzien de assessmenttaken zoals die door de studenten en docenten gepercipieerd werden. De kwantitatieve resultaten tonen aan dat, de verschillen in resultaten op de voorgaande assessments onder controle gehoudend, studenten die alle assessmenttaken succesvol hadden afgerond tijdens het blok hoger scoorden op de assessment aan het eind van het blok dan studenten die niet alle assessmenttaken succesvol hadden afgerond. Deze studenten presteerden niet alleen beter op de vragen die gerelateerd waren aan onderwerpen waarop de assessmenttaken betrekking hadden, maar ook beter op overige vragen in de assessment. Dit geeft aan dat de assessmenttaken studenten helpen gepaste leeractiviteiten te ontplooien. De invloed ervan is groter dan op de aan de assessmenttaken gerelateerde leerstof alleen. Zowel studenten als docenten percipieerden de assessmenttaken als zinvol. De kwalitatieve data geven aan dat studenten en docenten profijt ondervinden van de assessmenttaken. Het werken met de taken stimuleerde meer intensief te werken met de leerstof. Gevolg van dit was dat studenten meer tijd in hun werk staken en beter voorbereid waren op de onderwijsbijeenkomsten. Volgens de docenten leidde dit tot betere discussies. Ondanks de positieve geluiden blijkt feedback een kritische factor te zijn. Met betrekking tot feedback dienen twee aspecten in ogenschouw genomen te worden: de te gebruiken criteria bij de beoordeling van de assessmenttaken en de wijze waarop feedback gegeven wordt op de uitwerking van de assessmenttaken. Al bij al heeft het onderzoek aangetoond dat kleine veranderingen in het assessmentsysteem tot relatieve grote veranderingen kan leiden in de wijze van studeren en in leerresultaten.

In hoofdstuk 7 worden de resultaten uit de onderzoeken samengevat en bediscussieerd. Ook worden suggesties voor verder onderzoek en voor de onderwijspraktijk gegeven. Geconcludeerd wordt dat een substantieel deel van de studenten niet (alle) kernfactoren in het probleemgestuurd onderwijs herkent. Veel studenten ervaren het probleemgestuurd leren niet als studentgecentreerd en een deel van de studenten lijkt geen studieaanpak te hebben (geen oppervlakkige en geen grondige). De rol van de docent is belangrijk in het faciliteren van studenten in het nastreven van de eigen leerdoelen en om te gaan met leermoeilijkheden. Het is de taak van de docent om zich te richten op studieaanpak, studievaardigheden en strategieën. Docenten zijn weldegelijk in staat studenten met ongepaste studiemethodes te identificeren tijdens de interacties in de onderwijsbijeenkomsten. Assessmenttaken kunnen als een middel gezien worden om docenten te faciliteren in het richten van studenten op een gepaste studieaanpak. Percepties van assessment spelen een belangrijke rol in het studiegedrag van studenten. De wijze waarop studenten de assessment percipiëren verschilt met de wijze waarop docenten de moeilijkheidsgraad van de assessment en de benodigde cognitieve processen voor het oplossen van de vraagstukken in de assessment percipiëren. Studenten hebben over het algemeen geen duidelijk beeld van de eisen die de assessment aan hen stelt.

De assessmentsamenstellers hebben over het algemeen geen goed beeld van de gemiddelde student. Samenstellers dienen de gemiddelde student eerst in gedachte te nemen en te beschrijven in termen van verwachtingen voordat begonnen wordt met het construeren van de vraagstukken. Tijdens de constructie van de assessmentvragen dient de omschreven gemiddelde student in gedachte gehouden te worden. Voor studenten is het nodig een overeenstemmende perceptie op te bouwen van de gebruikte assessments. Assessmenttaken kunnen studenten helpen een beter beeld op te bouwen van de eisen zoals bedoeld door de assessmentsamenstellers.

Curriculum Vitae

Gerard van de Watering was born on September 24th 1966 in Bergen op Zoom (the Netherlands). After he obtained his teaching certificate in electro techniques at the Pedagogische Technische Hogeschool (PTH), he studied Educational Sciences at the Katholieke Universiteit Nijmegen (KUN). His thesis focussed on promoting independent learning by means of cooperative settings. After his graduation he worked in secondary and higher vocational education as a teacher and educational developer. In 2000 he started as an assistant professor at the department of Educational Innovation and Information Technology (EDIT) at the faculty of law at the University of Maastricht. His research and development interest focus on assessment and evaluation, student-centred learning environments, independent learning and study skills.

Publications

Dierick, S., Dochy, F., & Watering, G. van de (2001). Assessment in het hoger onderwijs: over de implicaties van nieuwe toetsvormen voor de edumetrie. *Tijdschrift voor Hoger Onderwijs, 19* (1), 2-18.

Dierick, S., van de Watering, G., & Muijtjens, A. (2002). De actuele kwaliteit van assessment: ontwikkelingen in de edumetrie, in F. Dochy, L. Heylen & H. v. d. Mosselaer (red.), *Assessment in onderwijs*. Utrecht: Lemma, 91-122.

Dochy, F., Gijbels, D., & van de Watering, G. (2004, June). *Assessment engineering: Aligning assessment, learning and instruction.* Paper presented at the EARLI-Northumbria Assessment Conference, Bergen, Norway.

Gijbels, D., van der Rijt, J., & van de Watering, G. (2004). Het bindend studieadvies in het hoger wetenschappelijk onderwijs: worden de juiste studenten geselecteerd? *Tijdschrift voor Hoger Onderwijs, 22* (2), pp 62-72.

Gijbels, D., & van de Watering, G. (2004, november). *Het integreren van assessmenttaken in de leeromgeving.* Paper gepresenteerd op het VFO-congres (Vlaams Forum voor Onderwijsonderzoek), Brussel, België.

Gijbels, D., van de Watering, G., & Dochy, F. (2005). Integrating assessment-tasks in a problem-based learning environment. *Assessment & Evaluation in Higher Education*, *30* (1), pp. 73-86.

Gijbels, D., van de Watering, G., & Dochy, F. (2005). Op weg naar een integratie van leren, instructie en toetsing. *Onderzoek van Onderwijs*, *34* (5), pp. 56-59.

Gijbels, D., van de Watering, G., & Dochy, F. (2005, August). *The relationship between students' approaches to learning and the assessment of learning outcomes.* Paper presented at the 11th EARLI conference, Nicosia, Cyprus.

Gijbels, D., van de Watering, G., & Dochy, F. (2006). New learning environments and constructivism: The students' perspective. *Accepted for publication in Instructional Science.*

Gijbels, D., van de Watering, G., Dochy, F., & van den Bossche, P. (2005). The relationship between students' approaches to learning and the assessment of learning outcomes. *European Journal of Psychology of Education*, *XX* (4), pp. 327-341.

Van der Rijt, J., Prop, A., van de Watering, G., de Rijdt, C., Dochy, F., & Gijbels, D. (2006, mei). Studentpercepties van Blended Leraning in Hoger Onderwijs. Paper gepresenteerd op de Onderwijs Research Dagen, Amsterdam, Nederland.

Van der Rijt, J., van de Watering, G., Gijbels, D., & Dochy, F. (2005, mei). *De moeilijkheidsgraad van toetsen in het Hoger Onderwijs. Het vermogen van docenten en studenten om de moeilijkheidsgraad van toetsvragen juist in te schatten.* Paper gepresenteerd op de Onderwijs Research Dagen, Gent, België.

Van de Watering, G. (2001, August). Assessment & edumetrics: new lines in evaluating quality of assessments. Paper presented in the expert panel on Assessment and Edumetrics at the 9th EARLI (European Association for Research on Learning and Instruction) conference, Fribourg, Switzerland.

Van de Watering, G., Gijbels, D., van der Rijt, J., & Dochy, F. (2005). *Students' assessment preferences in relation to their study-results in new learning environments*. Paper presented at the 11th European Association for Research on Learning and Instruction conference, Nicosia, Cyprus, August.

Van de Watering, G., & Claessens, S. (2003). Verschillen tussen de perceptie van tutoren op de leerprestatie en de toetsprestatie van hun studenten. *Tijdschrift voor Hoger Onderwijs, 21* (3), pp. 199-214.

Van de Watering, G., & Claessens, S. (2003, August). *The discrepancy between tutors' perceptions of student performance and achievement results.* Paper presented at the 10th EARLI (European Association for Research on Learning and Instruction) conference, Padova, Italy.

Van de Watering, G., & Dierick., S. (2002). Kwaliteit van assessment: de bruikbaarheid van klassieke toetsen binnen studentgericht- en competentiegericht onderwijs, in: F. Dochy, L. Heylen & H. v. d. Mosselaer (red.), *Assessment in onderwijs*. Utrecht: Lemma, 61-90.

Van de Watering, G., & van der Rijt, J. (2006). Teachers' and students' perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. *Accepted for publication in Educational Research Review*.

Gelezen op het vliegveld van Dar-es-Salaam, wachtend op het vliegtuig terug naar huis:

Where there is a purpose there is no failure.
(Swahili gezegde)