Construct validation of teacher portfolio assessment

Procedures for improving teacher competence assessment illustrated by teaching students research skills

Marieke van der Schaaf
Universiteit Utrecht, Afdeling Onderwijskunde, 2005

The studies in this thesis are funded by
the Dutch Organization for Scientific Research NWO-PROO 411-21-204

Lay-out Roy Sanders
Cover design Brian Golsteijn
Printed by Febodruk bv
ISBN 90-9020054-1

Universiteit Utrecht, Afdeling Onderwijskunde

Construct validation of teacher portfolio assessment

Procedures for improving teacher competence assessment illustrated by teaching students research skills

Constructvalidering van docentportfoliobeoordeling
Procedures voor het verbeteren van de beoordeling van docentcompetenties aan de hand van het leren doen van onderzoek aan leerlingen

(with a summary in Dutch)

Proefschrift

Ter verkrijging van de graad van doctor aan de Universiteit van Utrecht op gezag van de Rector Magnificus, Prof. dr. W.H. Gispen, ingevolge het besluit van het College van Promoties in het openbaar te verdedigen op 16 december 2005 des ochtends te 10.30 uur.

door

Marieke Maria-Feikje van der Schaaf
Geboren op 5 oktober 1974 te Hoorn

Promotores

Prof. dr. K. Stokking, Universiteit Utrecht
Prof. dr. N. Verloop, Universiteit Leiden

Leden beoordelingscommissie

Prof. dr. D. Beijaard, Rijksuniversiteit Groningen
Prof. dr. B. van Hout-Wolters, Universiteit van Amsterdam
Prof. dr. J. van der Sanden, Technische Universiteit Eindhoven
Prof. dr. P. Sanders, Cito Arnhem
Prof. dr. T. Wubbels, Universiteit Utrecht

Contents

1 General introduction

1.1 Background and context

Offering students a sound education asks for a competent teacher workforce. In response to current international demands for accountability and image improvement of the teaching profession, nowadays there are two common ways to support teacher quality (Cochran-Smith & Fries, 2001; Ingvarson, 2002). The first way focuses on schools as workplaces in which teachers are supported and assessed by performance management systems. In England and Wales, for example, recently a new career structure to recruit, retain and reward good teachers is introduced in which annual assessments and appraisals of teachers are conducted by head teachers and senior school managers (Secretary of state for education and employment, 1998). The second way consists of further development of the teaching profession by certification. In the US, for example, school-independent professional standards are developed and used by The National Board for Professional Teaching Standards (NBPTS), a non-profit and non-partisan organization of whom most members are teachers. Both ways to support teacher quality are complementary in the sense that performance management systems in schools ensure teacher quality in the workplace, and professional teaching standards explicate what is expected of teachers in terms of knowledge, skills and attitudes.

In The Netherlands it is tried, so it seems, to combine both ways in one system of teacher development and assessment. Recent legislation by the 'Wet op de beroepen in het onderwijs' (Stb. 2004, 344; Stb. 2005, 460). lays down national standards, consolidating earlier moves towards developing professional profiles for experienced teachers and formulating starting competences for beginning teachers (Commissie Toekomst Leraarschap, 1993; Ministerie van Onderwijs en Wetenschappen, 1993; 1998). On the one hand, the 'Wet op de beroepen in het onderwijs' aims to guarantee teachers' minimum competences and to make employers (i.e., schoolboards) responsible for supporting teachers' development during their careers. For this purpose performance management systems (in The Netherlands also denoted as 'Integraal personeelsbeleid') are used. On the other hand, professional standards that teachers should meet are described at the national level by the teacher profession itself, coordinated by the 'Stuurgroep Beroepskwaliteit Leraarschap' (SBL) (Nedermeijer & Pilot, 2000). Teachers and other stakeholders in the field of education are involved in the development of those standards, which describe which competences are expected of teachers.

Whereas in European countries little empirical evidence has been reported on the quality of new teacher development and assessment systems, in the US rigorous evaluations of the certification system emerge. However, the results of those studies are quite puzzling. For example, frequently cited large-scale studies by Bond, Smith, Baker & Hattie (2000), Darling-Hammond, Berry & Thoreson (2001) and Darling-Hammond (2002) that show positive effects of NBPTS teacher certification of experienced teachers stand in contrast to several studies that find no effects at all or only very small effects for teacher certification (e.g. Goldhaber & Brewer, 2000; Walsh, 2001). In order to better understand the possibilities, merits and shortcomings of teacher assessment systems, there is an urgent need for research into these systems. Such research can be embedded in the context of at least four current issues.

Firstly, the way teacher competences are conceived has recently changed due to a cognitive shift in teacher research (Clark & Peterson, 1986). Quality teaching is now recognized to be a result of complex interactions between teachers' beliefs and behaviour in a certain context and teachers' thoughtful examination of these interactions. Because there are different 'good' ways of teaching and as the relation between teacher behaviours, cognitions and actions is not yet very clear, topical questions are how to describe precisely the professional standards teachers have to meet and how to decide when a teacher meets these standards (Dwyer, 1994; Sternberg & Hovarth, 1995; Uhlenbeck, 2002). We make a distinction between content standards that describe the dimensions on which teachers should be assessed and performance standards (cut-off scores) that indicate the minimum levels for a sufficient result. Furthermore, an important issue is how to decide who are the persons that should participate in standard setting activities. Research into those questions should cover the development and testing of suitable methods for setting teacher content and performance standards, including procedures to select significant stakeholders.

Secondly, the attention for teacher assessment comes up in a period in which teachers are expected to change their practices towards new (constructivist) views on learning that emphasise the development of students' skills for more active self-directed and collaborative learning in projects. Teachers are expected to expand their didactical repertoire with more process-oriented teaching strategies and new assessment approaches (performance assessment). A prominent example of these new demands is a recent large-scale educational reform in The Netherlands, which aims at the development of general skills, especially research skills in upper secondary education (ages 16-18). Research skills form an obligatory part of the final exams of Dutch secondary education since 1998. Dealing with such reforms asks for (new) teacher competences (knowledge, skills and attitudes) that mediate teacher performance. As yet little is known about the precise competences needed and the degree to which teachers possess such competences. Descriptive research of teachers' behaviour and cognitions while dealing with recent large-scale innovations can be useful in these matters.

Thirdly, recent developments in cognitive psychology have resulted in a revival of competence-based assessment. In parallel with the movement towards alternative assessment of students, teachers, too, are increasingly assessed by more authentic assessments, using instruments such as portfolios (Delandshere & Petrosky, 1994; Delandshere & Arens, 2001). Depending on its content and form, a portfolio can do justice to the fact that teaching is a complex activity and that teachers' behaviour is bound up with their cognitions and the teaching context (Andrews & Barnes, 1990; Bird, 1990; Lyons, 1998). However, the merits and pitfalls of teacher portfolios and the ways such portfolios can be validly and reliably developed and scored have not yet crystallized out. To improve the quality of scoring, insight

into raters' judgment processes is crucial. Yet, raters' cognitive representations in assessing teacher portfolios are rarely examined. Insight into raters' cognitions is needed to facilitate and enhance our understanding of portfolio assessments and to improve their quality.
Fourthly, there is much debate about the appropriateness of traditional psychometric criteria, such as validity and reliability, for alternative forms of assessment, such as performance assessment, and competence-based assessment (e.g. portfolio-assessment). Especially attaining acceptable levels of reliability is a notorious problem. The debate focuses on questions such as whether in competence-based assessment other, more specific or lower standards are needed and how validity can and should be conceptualised (Messick, 1989; Borsboom, Mellenbergh & Van Heerden, 2004). Research into the quality requirements needed and how these requirements can be attained is crucial.
These four issues are interwoven. For example one of the ways in which cognitive psychological research into teaching influences teacher assessment is through the methods it provides for examining teaching constructs. Cognitive psychological research also leads to a more detailed understanding of teaching and the ways in which assessment can promote learning and as such can contribute to psychometric assessment approaches that often suffer from a lack of cognitive psychological theoretical background.

1.2 Purposes and relevance

This thesis reports on a series of studies embedded in the context of the four issues described above. The main goal of this thesis is to develop and test suitable procedures for teacher portfolio assessment concerning teachers' competences in teaching students research skills. The major scientific relevance lays in the contribution to the third issue about improving the scoring of teacher portfolio assessment, and the first issue about developing content and performance standards, and in the connections made between the four issues. An example of the latter is that procedures for teacher portfolio assessment as developed in this thesis (related to the third issue) have been developed in connection with teaching students research skills (involving the second issue) and have been tested to a selected set of quality requirements (the fourth issue). The four issues are also interwoven in the development of assessment standards (first issue) and the quality requirements met by those standards (fourth issue). Further, we examine raters' cognitive representations (in different phases in the assessment process) and the role of teacher beliefs in teacher portfolio assessment.
The strategic relevance of these studies is twofold. Firstly, the thesis should be seen in the context of a growing demand for suitable and tested procedures needed to improve the quality and feasibility of everyday teacher competence assessment. We aim to expand a Delphi method and a Policy capturing method for developing content and performance standards. We also want to contribute to the portfolio development process by developing a portfolio using a chain model of construct validation. Secondly, recent innovations, such as teaching research skills in Dutch secondary education, have implications for teaching and for the teaching competences teachers need. Although many Dutch teachers subscribe to the relevance of the current innovations, at the same time they are uncertain about their new roles in the educational process (Stokking & Van der Schaaf, 2000). Research into teachers' competences might be formatively useful (e.g. by giving suggestions for teacher training) and might indicate what is reasonable from teachers to expect.

1.3 Research questions

This thesis focuses on three research questions, divided in several sub questions, about teacher portfolio assessment in the context of teaching students research skills. The first question focuses on procedures to improve the valid assessment of teacher competences. The second question goes into the quality of raters' scoring of teacher portfolios. The third question is about beliefs and behaviour (as components of competence) in teacher assessment, and the correspondence between beliefs and behaviour.
The first question is connected to the first issue described in section 1.1. The second question mainly contributes to the third and fourth issue. Answers on the (sub) questions 1b, 1d, and 3 may contribute to the second issue.

1. How can teacher competences concerning teaching students research skills be validly assessed?
a) How can a Delphi method be used and optimised as a procedure to develop content standards for teacher assessment in the context of educational reform? (chapter 3)
b) Which content standards can be set regarding teaching students research skills? (3)
c) How can a Policy capturing method be used and optimised as a procedure for developing performance standards for teacher assessment? (4)
d) Which performance standards can be set regarding teaching students research skills? (4)
e) How is the portfolio design related to the content standards? (5)
f) How is raters' portfolio scoring related to the content standards? (5)

2. How can rater quality in teacher portfolio assessment be improved?
a) What is the reliability of the assessments? (6)
b) How do raters assess teachers' portfolios? (6)
c) How are raters' cognitive processes related to their judgments? (6)

3. What is the correspondence between teacher beliefs and behaviour as constructed by teachers in their portfolios and as assessed by their students and by external raters? (7)

1.4 Nature of the study

The research approach or type of research used to answer the first question (chapters 3 to 5) is described by De Groot (1961) as "instrumental-nomological" and is also denoted as design-based research (Brown, 1992; Collins, 1992; Kelly, 2003; Van den Akker, 1999). The main goal of design-based research is to deal with the many uncertainties connected to complex problems in dynamic contexts and to create effective and practical interventions to solve these problems. In this thesis interventions have been focused on the systematic development of content standards and performance standards and the development of a portfolio design for teacher assessment.
Design-based research is cyclic in nature and includes one or more phases of designing educational interventions, developing, formatively evaluating and producing these interventions, and a final summative evaluation (Van den Akker, 1999). This thesis includes all three phases. The first phase is based on a literature study and on a survey study into leading edge teachers' practices in teaching students research skills (chapter 2). In the development phase, human

judgment and decision-making are crucial. To support this phase we use a Delphi method (Linstone & Turoff, 1975) and a Policy capturing method (Jaeger, 1995). After the development phase, the results are evaluated and subsequently a new development stage of further refinement starts (the Delphi method and the Policy capturing method used in this thesis included three rounds and two rounds respectively, with in between feedback and discussion). In the end, the research methods used are evaluated.
The studies reported about in chapter 6 about cognitive representations in raters' assessment of teacher portfolio and in chapter 7 about teacher beliefs and behaviour as shown in their portfolio are of a descriptive nature, using theoretical frameworks to describe the researched phenomena. In chapter 6 we analyse raters' judgment forms and retrospective verbal protocols using the Correspondent Inference Theory (Jones & Davis, 1965) and the Associated Systems Theory (Carlston, 1992; 1994). In chapter 7 the information used to search into teacher beliefs and behaviour consisted of the data in the portfolios made by the teachers, assessments of these teachers' behaviour made by their students, and assessments based on the portfolios by independent raters. To examine the role of teacher beliefs in teacher portfolio assessment, we use the Theory of Planned Behavior (Ajzen, 1985; Ajzen & Fishbein, 2005). This theory has shown to be useful for research into the relation between teachers' constructed beliefs and behaviour in several subjects (e.g. Crawly, 1990; Haney, Czerniak & Lumpe, 1996).

1.5 Outline of the thesis

Chapter two reviews contemporary insights in research into teacher competences, competence-based teacher assessment, and teaching students research skills. We especially focus on research on teacher thinking that might have implications for teacher portfolio assessment. We also discuss quality requirements needed in teacher portfolio assessment, and consider literature about improving the quality of scoring. Finally, we investigate teachers' desired and actual behaviour in teaching students research skills.
In the third chapter on the development of content standards for assessing teaching research skills, we describe the design and testing of a Delphi method based on stakeholders' judgments. In three rounds, 21 stakeholders judged and revised content standards, resulting in nine content standards.
In the fourth chapter we address the question how good teachers' assessed competences should be for a satisfying assessment and how the different competences should be weighed. Using a Policy capturing method, in two rounds nine stakeholders developed performance standards for teacher assessment on eight of the nine content standards developed in chapter three. Between these two rounds, the panellists held a structured group discussion.
The purpose of the fifth chapter was to design a teacher portfolio based on a chain model for construct validation. We focused on two links between main components in teacher portfolio assessment, namely the link between the content standards and the portfolio design and the link between the content standards and the raters' scoring.
In the sixth chapter we address the issue of rater reliability, which is a vexing problem in teacher portfolio assessment. Insight into raters' judgment processes is crucial for improving the quality of assessment. We used a mixed quantitative and qualitative approach to investigate underlying cognitive processes. In this chapter we describe how six raters systematically scored and assessed 18 portfolios.
The seventh chapter deals with teacher beliefs and behaviour in teacher portfolio assess-

ment. The purpose of this chapter is to search into the correspondence of teacher beliefs and behaviour in teacher portfolio assessment, focusing on the interactive phase of teaching. Despite the widespread idea that teacher assessment should take account of teacher beliefs in relation to their behaviour, the role of teacher beliefs in teacher assessment is far from clear. To search into this relation we distinguished teacher beliefs and behaviour at three levels: a) the theoretical conceptualisation of teacher beliefs and behaviour; b) teachers' constructed beliefs and behaviour in their portfolio; c) the teachers' beliefs and behaviour as assessed by students and external raters.

Chapter eight concludes this thesis by a general discussion. A summary is given of the main results and some limitations are discussed. The discussion also considers the roads along which cognitive psychological research into teacher thinking and raters' scoring processes might especially have an impact on actual assessment practices. Further, a research and development agenda is proposed for expanding our knowledge around the cutting face between cognition and assessment. Finally, some practical recommendations are provided.

The thesis can be read as a whole with subsequent chapters. The chapters 3 to 7, separately published or submitted for publication, can also be read separately as independent articles. For that matter, some small overlap between the chapters is inevitable.

2 Assessing teacher competences in teaching students research skills

Section 2.3 is partly published in: Stokking, K., Van der Schaaf, M., Jaspers, J., & Erkens, G. (2004). Teachers' assessment of students' research skills. *British Educational Research Journal, 30* (1), 93-116.

Section 2.4 is based on: Stokking, K., & Van der Schaaf, M. (2000). *Ontwikkeling en beoordeling van onderzoeksvaardigheden* [Development and assessment of research skills]. Utrecht, The Netherlands: ICO-ISOR, Utrecht University

2.1 Introduction

This chapter provides a general theoretical background and implications for the studies in this thesis. One of our basic assumptions is that the way in which teacher competences are conceived has implications for assessment. To conceptualise teacher competences we look at research into teacher thinking and teacher expertise and we derive some implications from this research for the assessment of teacher competences (2.2).
Secondly, we search into competence-based assessment as a recent form of more authentic assessment of teacher competences (2.3). We especially explore the potential of a teacher portfolio as an instrument in competence-based teacher assessment. We also review quality requirements which teacher assessment should meet and focus on providing high scoring quality, a vexing problem in teacher portfolio assessment. We assume that the scoring by raters during the assessment is a human cognitive activity. Psychological theories and methods to explicate and influence the role of assessor's cognitions during the scoring process might contribute to the quality of scoring.
A prerequisite for a valid assessment is a clear operationalization of the construct domain to be assessed. We search into the domain of teaching students research skills, which is the joint teaching domain in all studies in this thesis. Based on the literature we identified and specified the main components of learning environments teachers can create for their students while teaching research skills: goals and objectives, instruction and guidance, tasks or assignments, and assessment (2.4). This literature study has been followed by an empirical study into leading edge teachers' practices (Stokking & Van der Schaaf, 2000) of which we report the main results.

2.2 Teacher competences

2.2.1 The concept of teacher competences

In recent years the assessment of competences has been a central issues in the teacher profession. Whereas traditionally the term teacher competence has often been perceived in terms of behaviour or in terms of individual psychological attributes, the recent concern for teacher

competence has been expressed as a need for a more integrated concept denoting teachers' knowledge, skills, and attitudes in context while performing professional tasks (Gonczi, 1994; Hager, Gonczi & Athanasou, 1994; Eraut, 1994; Tomlinson, 1995; Velde, 1997; Westera, 2000).

Traditionally, evidence for teacher competences is based on direct observation of performances and/or products in predictable situations, reflecting a technical concept of teaching. Especially in the US, in the 1970s and 1980s teacher competences are conceived in terms of underlying stable, personal, and situation-independent characteristics providing a basis for excellent performance. This approach parallels early cognitive views of competence, which distinguish between performance and competence. This distinction is significant as it is constitutive to the idea that teacher competences are assessed indirectly, derived from teaching performances, which influences the types of evidence that should be gathered for teacher assessment. This approach is criticized for its reliance on observational checklists lacking a theoretical basis (Andrews & Barnes, 1990; Dwyer & Stufflebeam, 1996). Other criticism concern the abstractness of the competences defined, detached from teachers' everyday teaching practice, and the questionable existence of generic, situation-independent competences (Gonczi, 1994). Also it is left unclear how observed behaviours relate to each other and to the whole of teaching (Eraut, 1994; Gonczi, 1994; Tomlinson, 1995; Wolf, 1995).

Recently, scholars have been moving toward a more integrated approach, based on the notion that competence is a relational concept, bringing together the abilities of the individual and the professional tasks to be performed in particular situations. Hager and Gonczi (1994, p. 4) describe this integrated approach to competence-based training and assessment as follows:

"(…) competence is conceptualised in terms of knowledge, abilities, skills and attitudes displayed in the context of a carefully chosen set of realistic professional tasks which are of an appropriate level of generality".

Fundamental is the idea that the professional's conception of the work to be done underpins his or her knowledge and skills practised and that in the definition of the competence concept an intentional dimension should be included (Sandberg, 1994). The idea that professional performances should be seen as intentional acts brings into play that professionals interpret their tasks and actively construct meanings, whether explicit or tacit. The argument that professionals are actively involved in making sense of their tasks differs from the approach of early research into 'teacher thinking' in which the act of teaching results from an inner representation of an external task. The integrated approach instead closely parallels that of situated cognition (Brown, Collins & Duguid, 1989; Klarus, 1998), which focuses on competences as social and contextual constructs. The idea that teachers' professional tasks are subject to their individual perceptions and interpretations in specific contexts could lead to radical relativism in the sense that teaching is holistic, cannot be objectively described, and as a result cannot be assessed, but we believe that there are elements of a teacher's knowledge base that are shared by (almost) all teachers in a certain period in a certain community, especially by teachers teaching the same subject to students of about the same level (Verloop, Van Driel & Meijer, 2001).

In this thesis we adopt an integrated approach, because this approach recognises the importance of teacher cognitions and of the context in which teaching takes place. This gives rise to some implications for teacher assessment, for example that it is necessary to assess both cognition and behaviour in context. To provide a useful basis for teacher assessment more work has to be done and this is described in the remainder of this section and in section 2.4.

2.2.2 Research on teacher thinking

During the last decades, the common perspective on teacher assessment has moved from one grounded in behavioural psychology to one based on cognitive psychology in which teacher behaviour is seen as integrated with teacher cognitions and with the context of teaching (Clark & Peterson, 1986; Shulman, 1986; 1987; Reynolds, 1992; Verloop, 1995). Research into teacher cognitions and decision-making started in the 1980s and is known as the 'teacher thinking paradigm' (Clark & Peterson, 1986; Calderhead 1996). This line of research was developed starting from a reliance on concepts drawn from information-processing studies focussing on isolated pre-, inter- and post-active phases of teaching, towards an orientation on understanding teacher cognitions in context (Carter, 1990; Borko & Putnam, 1996). In this orientation, what takes place in context (i.e., the setting in which teachers work and the content of the subject they teach) is seen as a function of the history of the teacher, the setting and the subject. This implies that teachers' knowledge and beliefs about teaching, students and the subject matter are central because they function as interpretative lenses through which teachers make sense of the situation.

Though teacher knowledge and beliefs can be theoretically distinguished by viewing teacher knowledge as propositions and teacher beliefs as more subjective personal theories, in fact both terms merge into each other (e.g. Calderhead, 1996; Meijer, 1999; Den Brok, 2001). Teacher knowledge consists of declarative knowledge (knowledge of concepts, facts and principles, including knowledge about performance-relevant behaviour or knowledge what to do), and procedural knowledge or skills (knowing how to do as well as what to do). Procedural knowledge differs from declarative knowledge in that it is part of the processes underlying performance. Procedural knowledge is not limited to cognitive processes, as it can include psychomotor and affective processes as well.

Teachers' declarative and procedural knowledge appear to be relatively rich and detailed, connected to specific contexts, and may differ in level of awareness. This corresponds with findings in research into teachers' practical knowledge (cf. Carter, 1990; Verloop, 1992; Meijer, 1999). Over the years, teachers have developed practical knowledge which influences their teaching practice and the way they handle innovations and is therefore relevant to our research. This knowledge is personal and unique, can be tacit or conscious, context-specific or general theoretical (Beijaard & Verloop, 1996). This practical knowledge can be categorized in different ways. For the description of a domain of teacher competences, categorizations concerning content are most relevant. Putnam and Borko (1997), using Shulman (1986, 1987) and Grossman (1990), give the following classification: general pedagogical, subject matter, and pedagogical content knowledge. In addition, Grossman (1995) distinguishes between knowledge about the context and the self. In our study, pedagogical content knowledge about teaching students research skills is the most relevant category: what a teacher knows, what a teacher does, and why, while teaching students research skills in their subject.

The relation between teacher knowledge, beliefs and behaviour is reciprocal. On the one hand teacher knowledge and beliefs are developed by experiences derived from teacher actions. On the other hand, teacher knowledge and beliefs can be seen as concepts that filter the input of external stimuli and that mediate the output of behaviour (Kagan, 1992). These filters imply expectations and beliefs about themselves, others, the possible effects of certain behaviour, and the teaching context that may ease or complicate certain behaviour.

Teacher knowledge and beliefs may guide teacher behaviour in either a deliberate or spon-

taneous way. In the deliberate way knowledge and beliefs are retrieved or constructed in an effortful manner in a certain context and they are assumed to guide intentional behaviour and describe at various levels of detail what the teacher wants to accomplish. Fazio (1990) argues that under conditions of high motivation and sufficient cognitive ability people can effortful construct their beliefs about certain intentional behaviour. Without those two conditions beliefs about intentional behaviour will be only operative when they are routinely activated. Routine activation of beliefs only takes place when an intention raises a strongly tied attitude towards (or evaluation of) certain behaviour. Because of this, strong beliefs are generally more resistant to change than weak beliefs.
Characteristic of routine teacher behaviour in every day teaching situations is that teachers do not have time for explicit reflection and decisions are made in a split-second (Eraut, 1994; Dolk, 1997; Korthagen, 1998). Dolk (1997) calls such a state of affairs 'immediate educational situations', referring to situations in which teachers' observation, interpretation and analysis of a situation and his or her feelings and behaviour in relation to that situation are barely distinguishable. Korthagen (1998) uses the term 'Gestallt' to indicate what activates teacher behaviour in everyday teaching practices (see also Dolk, 1997).

2.2.3 Research on teacher expertise

An important theme within the teacher thinking paradigm is expert-novice research. Expert-novice studies in different subject domains shed light on critical characteristics of good teachers that could be objects for assessment. Based on theory and research in artificial intelligence and information-processing, traditional expert-novice research initially took shape in the 1970s and 1980s studying characteristics of problem-solving performance of experts in other fields than teaching (Chi, Glaser & Farr, 1988). Since the 1980s also expert-novice studies of teaching have appeared (Ericsson & Smith, 1991) and now numerous studies show consistent results concerning the differences in cognitions and behaviours of expert and novice teachers (Borko & Livingston, 1989; Leinhardt, 1988). For example, based on a review of research into teacher expertise, Berliner (1994) summarizes that expert teachers: excel mainly in their own domain and in particular contexts; develop routines for the repetitive operations that are needed to accomplish their goals; are more opportunistic and flexible in their teaching than are novices; are more sensitive to the task demands and social situations surrounding them when solving problems; represent problems in qualitatively different ways than do novices; have faster and more accurate pattern recognition capabilities; perceive more meaningful patterns in the domain in which they are experienced; may begin to solve problems slower, but they bring richer and more personal sources of information to bear on the problems that they are trying to solve.
Whereas expertise studies have resulted in important contributions to our insight into what constitutes a 'good' teacher, the nature of teacher expertise remains confusing (Berliner, 2001). Firstly, research shows that although inexperience is connected to a novice status, being an experienced teacher does not automatically imply expertise (Eraut, 1994; Bromme & Tillema, 1995). For example, Borko and Putnam (1996) argue that compared to novices experienced teachers have a better conceptual understanding of the subject they teach, but their understanding is still often limited. Secondly, there is a debate about whether expert teaching is determined by (innate) talent (Sternberg, 1996) or by deliberate practice (Ericsson & Charness, 1994). Thirdly, teachers' working conditions have proved to be of strong impact on the development of expertise and therefore the development of expertise is context sensitive.

Even the meaning of teaching expertise differs between cultures and periods (Eraut, 2000).
We conclude that expertise is a relative concept that cannot be treated from an objectivist point of view. As a result, some scholars reject the idea of expertise as denoting a fixed level of teacher competence and instead view teacher expertise as an ongoing process of acquisition and consolidation of competences needed for a high level of mastery in one or more domains of teaching (cf. Sternberg, 1999).
Traditional psychological theories of expertise development usually contain three to five stages of development, using anchor points such as: a) a novice stage, wherein numerous mistakes occur; b) a transitional stage, wherein performances improve and some routine is developed; c) an expert stage, wherein high levels of performance arise (Dreyfus & Dreyfus, 1986; Kagan, 1992). Such stage models have been criticized for the ordered and linear progress of teacher growth they assume (Bullough, 1997), under representing the unique personal dynamic processes of teachers' learning in connection with the teaching context. Nowadays, the development of expertise is seen as an irregular process with sudden improvements and unexpected environmental influences (Huberman, 1995; Day, 1999), in which the course of development can differ between individuals and between contexts (Eraut, 1994). This view parallels a social-constructivist approach in which teachers are seen as learners who actively construct and reconstruct knowledge by constantly interpreting the situations and events they are confronted with (Borko & Putnam, 1996; Putnam & Borko, 1997).

2.2.4 Implications for the research in this thesis

The conceptualisation of teacher competences as being of a cognitive, context-sensitive and developmental nature has major implications for teacher competence assessment. Uhlenbeck (2002) and Uhlenbeck, Beijaard & Verloop (2003) give a thorough overview of these implications. We now summarize the salient points. Firstly, both teacher behaviour and cognitions should be included in the assessment. Secondly, content and performance standards should be developed allowing a range of satisfying ways of teaching. Thirdly, the teacher's workplace should be involved in the assessment of teaching. Fourthly, teacher assessment should capture the full teaching process, including pre-, inter- and post-active phases.
We aim to elaborate on these implications in this thesis. In chapter 3 we develop content standards based on research in the domain of teaching students research skills. The development of performance standards is complicated by the dynamic nature of expertise development. One of the questions that arise is whether to base performance standards on a compensatory model (in which teachers can compensate for weaker competences by being strong on other aspects) or on a conjunctive model (in which certain minimum competences are required). We address this issue in chapter 4.
In addition, we focus on an integrated assessment of teacher behaviour and cognitions. In chapter 5 the development of a portfolio assessment method is described as a method to assess teacher cognitions and behaviour in context, including the different phases of teaching. Assuming that teacher competence is embedded in the teaching performances and products, evidence can be collected from teaching performances (e.g. video registrations) and teaching products (e.g. lesson plans). Assuming that teaching is knowledge-based, Eraut (1994) argues that in addition also constructed evidence is worthwhile about a teacher's cognitions underpinning his or her teaching that cannot be directly derived from the teacher's performance. In this thesis we include both kinds of evidence.
Research on teacher thinking shows the importance in competent teaching of a connection

between teacher cognitions and behaviour. Whereas the existence of a connection between teacher cognitions and behaviour has been well established, the role teacher beliefs (can) play in teacher portfolio assessment is far from clear. We try to deepen our understanding into the relation between teacher beliefs and behaviour in portfolio assessment in chapter 7.

2.3 Teacher assessment and quality requirements

2.3.1 Teacher assessment

To deal with the process of drawing inferences about teachers' competences based on their responses to especially created or naturally occurring stimuli, the terms assessment, measurement, evaluation and testing are often used interchangeably. For the sake of clarity, we prefer to distinguish these terms.
In this thesis we view assessment as a human cognitive process of capturing teaching performances and products to produce data that can be used to make inferences about teacher competences (what teachers know and do) (Mislevy, Steinberg & Almond, 1999; Brookhart, 1999). This definition implies that we refer to assessment as a broad process including measurement, scoring, analysis and interpretation (Crooks & Kane, 1996; Stokking, 2005) see *Figure 2.1*. We define measurement as the process of representing empirically ascertained causal effects of tasks on responses in a numeric system (De Groot, 1961; Borsboom et al., 2004). We view scoring as the sub process of systematically assigning values (usually numerical ones) to the responses. Finally, the scores are analysed and interpreted within a model of the teacher competences assessed.

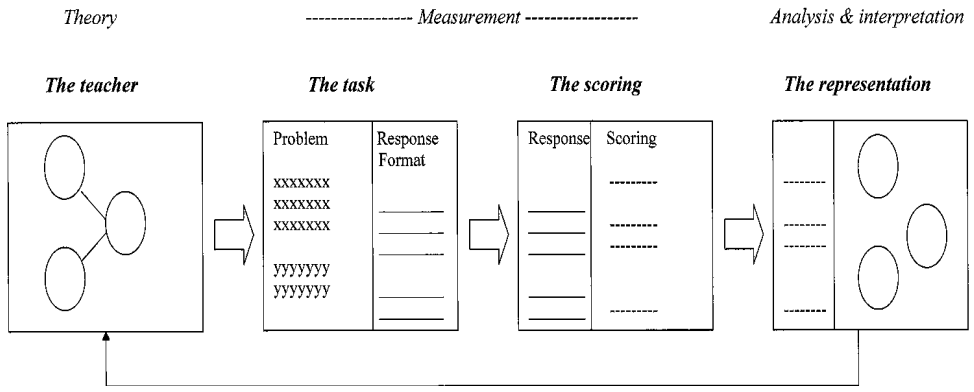


Figure 2.1 Explicitly modelling the assessment process (Stokking, 2005)

Including measurement and scoring as sub processes within the total assessment process, as we do, fits in with the meaning of the term assessment as common in the UK and departs from an accepted terminology in the US in which the term assessment is often used for measurement and the term evaluation for scoring (Vermetten, Daniëls & Ruijs, 2000). In line with the UK tradition, we use the term 'evaluation' for the appraisal of the quality of educational programmes, materials and organizations and not for the assessment of competences of people. Some authors distinguish between 'assessment' as a term for alternative, more authentic forms of assessment (Dochy & Segers, 1999) and 'test' for more traditional paper-and-pencil

tests. In this thesis we use 'competence-based assessment' to denote a more authentic approach (Gonczi, 1994).
The term *competence-based assessment* is regularly used in assessments of *professionals* (such as teachers and nurses) in Anglo-Saxon countries, which are often based on collections of evidence of work performance, and in which competence standards are used as benchmarks. Such assessments often are used together with a competence-based approach of professional education and training (Gonczi, 1994). Characteristics of such more authentic (real-life) assessments are: they have no strict time limit, they rarely have a single correct answer, and they permit the use of a variety of aids and resources. Competence-based assessments are supposed to be more valid than other more traditional methods. However, for this claim so far there is hardly empirical evidence (Linn, Baker & Dunbar, 1991; Birenbaum, 1994; Keeves, 1994).
In a similar vein, in the literature about assessments of *students* the term *performance assessment* is used where students are required to create an answer or product or to visibly demonstrate the completion of tasks to measure their achievement. Such formats might include work samples and simulations. Many authors use the terms performance assessment and authentic assessment interchangeably, which may be explained by the difficulty to imagine an authentic assessment that is not also a performance assessment. However, a distinction can be drawn concerning the degree of authenticity. Authentic assessment focuses on the context in which people perform, especially to support the fidelity of the assessment as to the way a performance in everyday work occurs, while in performance assessment this is not per se the case (Reeves & Okey, 1996).
This thesis is about teacher competence-based assessment. This kind of assessment requires fidelity as to the teachers' workplace and includes the gathering of data from complex long-term performances on a full range of tasks (e.g. from pre-active to post-active teaching phases).

2.3.2 Teacher portfolio assessment

In competence-based teacher assessment it is common to use portfolios with selected evidence of performances and products in various contexts, accompanied by teacher's comments and reflections (Wolf & Dietz, 1998). Depending on its content and form, a portfolio can do justice to the fact that teaching is a complex activity and that teacher behaviours are bound up with teacher cognitions and the teaching context (Andrews & Barnes, 1990; Bird, 1990; Lyons, 1998).
Portfolio assessment can be used as a tool to ascertain whether teachers satisfy the required competences and to formulate guidelines for professional development. In the first case, assessment is used summatively to account for the teacher's quality, with possible consequences such as merit pay and certification. In the second case assessment has a formative goal and produces information that can be used for planning activities directed at further professional development.
We endorse the idea that any teacher assessment should support teachers' professional development (Beijaard & Verloop, 1996) and that this is important on behalf of the acceptability of the assessment by the teachers involved (Andrews & Barnes, 1990). In this thesis we focus on portfolio assessment for high-stake purposes but with a formative goal as well, providing feedback about teacher's strong and weak points in teaching students research skills.
The combination of both assessment purposes, summative and formative, can and has been questioned (e.g. Barton & Collins, 1993; Paulson & Paulson, 1994). A frequently used argu-

ment is that high-stake assessments will sabotage professional development because it obstructs teachers to experiment in their teaching without being settled up. On the other hand, portfolio evidence for formative assessment purposes might be of insufficient quality to meet minimal acceptable quality requirements needed for high-stake assessment. Contrarily, there are authors who opt for portfolio assessment fulfilling both purposes at the same time, although some tension will always remain. For example Snyder, Lippincott & Bower (1998) conclude that reconciliation is possible when the assessment can be based on a broad archive of portfolio evidence gathered over a longer period of time from which teachers can select evidence for specific assessment purposes. The use of digital portfolios may turn out to be efficient in this regard (Van Tartwijk, Pilot & Wubbels, 1999). According to these authors, one of the prerequisites for combining both purposes is to clearly communicate to the teachers which portfolio evidence will be used for which assessment purpose. As a consequence, the development of portfolios in this thesis should provide enough space for teacher comments and reflections on the one hand and should be sufficiently specified to meet psychometric quality requirements on the other hand.

In a portfolio various evidence can be included and this material may be more or less structured. A portfolio typically includes evidence drawn from practice in the form of productions (documents especially prepared for the portfolio), reproductions (registrations of teacher's regular work captured on behalf of the portfolio), artefacts (regular products of teacher's work), and judgments of others about the teacher's work (e.g. students or colleagues) (Collins, 1991). In our studies all these forms are used.

As portfolios can provide a rich view of teaching in context, a portfolio is often seen as a highly valid instrument for teacher assessment. However, the data in a portfolio, being often descriptive, context-bound and personal, ask for much interpretation before they can be scored (Moss, 1994). This may decrease the reliability of these scores and reliability studies of portfolio assessment indeed often show disappointing reliability estimates. Shapley and Bush (1999) and Reckase (1995) claim that interrater reliability may increase when the portfolio content is specified and content standards and scoring rules are clearly defined. Specification of the portfolio content and the content and performance standards can also support teachers in developing a portfolio because this specification instructs them how to gather portfolio data (Wade & Yarbrough, 1996; Freidus, 1998).

Traditionally, teacher assessment is based on a psychometric approach, focussing on one-dimensional constructs, one-time measurement, description and analysis of individual differences, and norm referenced standards. This does not match well with assessment of complex constructs over a longer time span using criterion referencing with multiple cut-off scores (e.g. basic, competent, advanced), in which feedback, self-assessment and reflection are seen as important elements of the assessment process. Consequently, there is a growing interest in alternative views on assessment that can deal with complex skills and competences, repeated measurements, giving feedback suggesting actions to further growth and development, and criterion- and self-referenced standards. However, it is not necessary and even would not be wise to completely dismiss the psychometric tradition, as some authors seem to suggest, often using the term 'edumetrics'. Psychometricians are increasingly engaged in the development of models and techniques for the assessment of complex cognitive processes (such as teacher competences) and authentic forms of assessment (Messick, 1989; Embretson, 1993; Kane, 1992; Mislevy, Steinberg & Almond, 1999; Straetmans & Sanders, 2000; Borsboom et al., 2004). In our view, this shift in psychometrics reflects a reconciliation of both approaches.

Following an edumetric line methods can be developed to capture representative complex teaching tasks in context and further teachers' professional development, while a psychometric approach can support the development of techniques to analyse and preserve validity and reliability as important quality requirements.

2.3.3 Psychometric quality requirements

From the literature on psychometric quality criteria for assessments we identified 15 criteria and facets, falling into four main categories (Stokking, Van der Schaaf, Jaspers & Erkens, 2004). *Figure 2.2* provides a summary, principally derived from De Groot (1970) and Messick (1989, 1995).

The *reliability* of assessments (including the instruments, procedures and standards used) depends on reproducibility and repeatability as conditions for consistency (between moments and between raters).

The *validity* criterion covers a number of facets. The facets 2a-2c view on the *construction* of a valid instrument. Facet 2a implies that an assessment should cover the knowledge content of teaching. The assessment of competences should be in tune with the execution of the task (2b). If teaching is conceived as consisting of separate phases (preactive, interactive and postactive), this should be reflected in the scoring instructions (2c). Facet 2d represents the *core* of the validity concept: outcomes should reflect the competences to be assessed. (Messick, 1984, 1994, 1995; Cronbach, 1988; Kane, 1992; Moss, 1992; Wiggins, 1993; Crooks & Kane, 1996; Borsboom et al., 2004). The facets 2e and 2f are relevant for *validation*: to check whether the assessment results are not an artefact of the methods used and to define the domain assessed as precise as possible.

The *acceptability* criterion primarily relates to aspects such as fairness and lack of ambiguity. It also embraces the need for raters to avoid dealing with illegitimate criteria (such as their 'overall impression' of a teacher) (3a), and to effectively communicate to teachers the intentions, organization and implications of the assessment (3b) (Millman & Greene, 1989).

The *utility* criterion refers both to the feasibility for the teacher and the rater (4c and 4d) and to the clarity of the meaning assigned to the assessments and the assessment results (4a). It is partly their utility that allows assessments and assessment results to play an effective role in professional development. Therefore assessments should be designed to facilitate the raters to achieve insight into teacher quality and to give teachers meaningful feedback.

Messick (1989, 1994, 1995) has put forward a broader meaning of the term 'validity'. This subsumes all quality criteria, not only those relating to measurement, evidence and interpretation, but also those concerning the use and consequences of an assessment. For reasons of clarity we prefer to preserve the traditional, more restricted meaning of validity and to maintain the differentiation between the four main criteria described above (cf. Borsboom et al., 2004).

The literature on performance assessment is full of positive conclusions and expectations (Wiggins, 1989; Wolf, Bixby, Glenn & Gardner, 1991; Rowe & Hill, 1996). Many authors are optimistic: with this method we would be able to measure competences and even better than with traditional tests. In our view the potential of this approach is not yet clear. We believe that alternative, more authentic forms of assessment must be subjected to the same critical review as other methods or instruments claiming validity (Linn et al., 1991; Linn, 1994; Glaser & Silver, 1994; Messick, 1994, 1995; Haertel, 1999; Stokking & Voeten, 2000). Firstly, competence-based assessment emphasize less the correctness of interpretations and more

Reliable and valid measurement

1. *Reliability:* intra-and interassessor agreement (consistency): if the same assessors some weeks later assesses the products once again, or if a colleague assessor assesses the products also, the scores should not differ too much.
2. *Validity:* the assessment should measure what the assessor wants to measure (and not something else); validity has a number of facets:
    a. Content: sufficient coverage of the construct or domain one want to measure;
    b. Process: based on a model of the task as much as possible; showing the development of the relevant competences;
    c. Scoring: the scoring model and scoring rules should mirror the structure of the construct, domain or task;
    d. Specificity and unambiguous interpretation: the assessment should distinguish between more and less competent teachers (see also criterion 4b), and the scores should not be explainable in a causal sense by other factors than the competences intended;
    e. Convergence (a special case of 2d): measurements of the same construct, domain or task using different methods should correspond as much as possible;
    f. Generalizability: the degree to which the results are valid for a broader range of tasks, conditions and populations.

Adequate assessment

3. *Acceptability:*
    a. Objectivity: precise scoring instructions, minimal need for interpretation, procedures and criteria re defensible to others;
    b. Transparency: procedures (taking, scoring), criteria, norms and consequences should be clear and explainable to others and the teachers should know them beforehand;
    c. Equality: teachers should be treated equally and should get equal chances;
    d. Unbiaseness: items should not discriminate between personal characteristics which are not relevant.
4. *Practical utility:*
    a. Functionality: the results should be suitable for the functions intended, such as: give information on the teachers' progress and weak points be suitable for grading and be suitable for communication with others;
    b. Appropriate difficulty level and discriminating power;
    c. Workability: it should be possible to realize the assessment in a teaching situation;
    d. Efficiency.

Figure 2.2 Quality criteria for assessments

the acceptability of consequences but the former (the traditional validity) remains significant in view of the importance of consistency between judgments and generalizability of tasks, especially in portfolio assessment with only one rater (e.g. the teacher's own head teacher) and a small number of tasks. Secondly, the design and development of a good assessment must be attuned to the goals and content of teaching and the resulting assessment procedure must be suitable for teachers to use or participate in. Thirdly, the assessment has to be general enough to accommodate the broad variation in teachers' choices, performances and products. Satisfying such diverse requirements implies a great deal of development work in order to create good scoring rules (cf. Arter, 1993; Baker, Abedi, Linn & Niemi, 1995; Novak, Herman, & Gearhart, 1996). In practice, it is often difficult to distinguish assessees in a sufficiently reliable and valid manner. So, in many cases the acceptability and the utility of the assessment will be in danger of being compromised.

2.3.4 Construct validation of teacher assessment

In developing and validating teacher competence-based assessments it is important to derive the constructs to be assessed and to define the domain to which the assessment results can be generalized (see the validity facets 2a and 2b, and 2f, described in the previous section; see also Drenth & Sijtsma, 1990). The quality of teacher assessments for summative purposes traditionally has been evaluated mainly through studies of the predictive validity of the assessment, that is, by examining the correlation between the teacher's results on the chosen assessment and an indicator of subsequent professional performance. However, while predictive validity is important from a practical point of view it is not decisive for construct validity (Cronbach, 1988; Messick, 1989; Wolming, 1999; Borsboom et al., 2004). Essential for a valid assessment is a relevant operationalization of the construct or constructs to be assessed. Therefore, in this thesis other questions are assumed to be more important: What do the methods and instruments actually measure and are the content and performance standards used for the assessment of teacher performance relevant? (cf. Wolming, 1999). Consequently, construct validity is central.

A classical conceptualization of construct validation in teacher assessment consists of the derivation of the meaning of an assessment score from the network of relations of the assessment with other measures using patterns of correlations (a nomological network approach, Cronbach & Meehl, 1955). Criticisms of this approach are that the meaning of scores cannot be derived from empirical relationships with other measures and that this approach even confounds the meaning of the constructs measured with their practical significance (i.e., what the constructs predict) (Bechtoldt, 1959). Therefore, correlation studies do not suffice to validate the assessment of teacher cognitions and behaviour because they do not provide a theoretical understanding of the mechanisms that cause the assessment scores (Embretson, 1993; Borsboom et al., 2004). Consequently, to validate teacher portfolio assessment the causal linkage between the constructs measured (teacher competences) and the scores given by the raters should be ascertained.

To conceptualise the validation process we use a chain model of construct validation (Crooks & Kane, 1996) (see chapter 5). Such a model refers to the process of gathering evidence for links between the constructs to be assessed (e.g. teacher cognitions and behaviour) and the constructs as assessed (the assessment scores). Main components in such a chain model are (Mislevy et al., 1999) (sequence altered, see also *Figure 2.1*): (i) a (psychological) construct domain that describes teacher competences in teaching students research skills (teacher model); (ii) teacher competences as displayed in the portfolios (task model); (iii) raters' scoring of these displayed teacher competences (scoring model); (iv) the content standards and performance standards by which the competences are finally assessed (assessment model). These standards (iv) are derived from the construct domain (i). As mentioned above, content standards describe the dimensions on which teachers should be assessed and performance standards are the cut-off scores that indicate the minimum levels for a sufficient result.

2.3.5 Scoring teacher portfolio data

If the use of portfolios is to lead to more reliable and valid assessments, the development and use of a scoring model must be supported and clearly understood by the raters. In this thesis, raters make judgments about teacher portfolios against content standards. The scoring reflects judgments in which several aspects of the content standards, the portfolios and the

teachers' performances are interpreted and weighted in the mind of the raters. For example, a decision must be made regarding how content standards are to be applied to create a score, commonly choosing between an analytic and a holistic scoring approach. In analytic scoring performances or products are judged systematically on separate aspects and the scores are combined somehow (whether weighted or not) to get a judgment on the total performance. In holistic scoring the performance is directly judged as a whole. In both cases, additionally norms or performance standards should be developed for performance levels that are to be regarded as satisfactory respectively good. The accurateness with which raters employ these standards is influenced by the quality of the standards, the scoring form and the instruction and training given to the raters.

The issue of the meaning and quality of this scoring process cannot be settled without investigating and evaluating the raters' cognitive processes. In the psychometric tradition much research on scoring has focused on estimating the precision of measurement and the identification of sources of error or bias. For example, scoring reliability is often estimated by the percentage of exact agreement in scores between panellists. Exact agreement is based on similarity of classification (Popping, 1983) and presumes that the classifications given are identical. This can only be realized when raters interpret the standards in completely the same way. As people always may differ in their interpretations, these conditions cannot be guaranteed. As a result teaching standards and teacher portfolio assessments using these standards will always be based on a mix of different underlying visions and the precise meaning of the standards and judgments will not become completely clear. Therefore, we assume that the percentage of exact agreement can only be used as an indicator of consensus and that the meaning of standards and of teacher portfolio assessments can only be ascertained relative to a community of practitioners who share the assumptions behind the standards and judgments (Dylan, 1996).

As indicated before, to improve the quality of the scoring processes in teacher portfolio assessment, insights in raters' cognitive scoring processes are needed. The assumption that raters' cognitive processes influence their judgments has been fed by social cognitive psychology over the past decades, but only recently there is more interest in implicit factors that influence human scoring processes in (portfolio) assessment situations (Elander, 2004).

Rating process models in social cognitive psychology that describe scoring activities (e.g. Landy & Farr, 1980; Feldman, 1981; DeNisi, Cafferty & Meglino, 1984) have in common that raters while observing, interpreting, and evaluating teacher products and performances are using schemata. These schemata are comparable to personal constructs (Kelly, 1955): content categories used to organize and simplify information. Interpersonal filters in these schemata induce raters to look selectively at the information about the people observed and to interpret this information according to their own constructs. In this process raters' tacit knowledge (that is, implicit knowledge acquired through personal experience and often difficult to formalise) is into play. Tacit knowledge is the fundament of the 'connoisseur' model of assessment, which is based on "the ability to make fine-grained discriminations among complex and subtle qualities" (Eisner, 1991, p. 63). By making such tacit knowledge explicit, research into raters' cognitive scoring processes may contribute to the understanding and the improvement of the quality of assessments.

2.3.6 Implications for the research in this thesis

This thesis is about teacher portfolio assessment as a form of competence-based teacher assessment. Since we aim to combine both formative and summative assessment purposes and the teachers in our studies will only have scarce experience in developing a portfolio, we develop a semi-structured teacher portfolio in which we structure the portfolio elements while teachers are free in the form of the portfolio material (Shapley & Bush, 1999). Also, teachers will be provided help in developing a portfolio when needed (Freidus, 1998; Seldin, 1997; Tillema, 1998).
Further, we aim to expand a psychometric evidence-based approach by research into panellists' and raters' cognitions. We develop and test several methods based on social cognitive psychology in human judgment to understand implicit influences on panellists' and raters' judgments. In the chapters 3 and 4 we develop content and performance standards and explore panellists' underlying assumptions and perspectives to better understand the meaning of the standards developed by the use of a Delphi method (Linstone & Turoff, 1975), a Policy capturing method (Jaeger, 1995) and a nominal group technique (Delbecq, Van de Ven & Gustafson, 1975). For example, in chapter 4 we use Policy capturing involving the construction of statistical models of the relations between the judgments made by the panellists and several 'cues' (in our case content standards and exemplar portfolio evidence) that influence those judgments. In addition, the judgment 'policy' of individual panellists is captured in the form of a regression model in which the relevant cues are weighted and combined. In this way we provide a numerical insight into the weights attached by the panellists to the cues and the way in which they combine these cues in their judgments. Feeding back these policies to the raters can trigger the explication of arguments behind their judgments.
Also, in chapter 5 we research the validity of the raters' scoring. We compare the judgments of teacher portfolios (the extent to which teacher portfolios meet the content standards according to the raters) to more objective information from the portfolios themselves.
Finally, we explicitly view raters' cognitive representations in assessing teacher portfolios to facilitate and enhance our understanding of portfolio scoring and to improve its quality in chapter 6 by a content analysis of rating forms used and verbal protocols made while scoring teacher portfolios. Research of this type can answer questions about raters' ability to identify and evaluate specific aspects of teachers' work, the extent to which criteria as distinguished are in fact interdependent, and the way in which specific aspects of teachers' work influence the judgments made. Such information might lead to the development of more useful content and performance standards and result in concrete examples of how the criteria were demonstrated in teachers' work.

2.4 Teaching students research skills

2.4.1 Student research as a skill domain

In many countries, education is being directed towards new goals, new views on learning and teaching, and new assessment strategies. New goals include complex skills and competences, such as research skills and cooperation skills. Teachers who have to teach and assess such complex skills are expected to give students assignments that give them options and stimulate them to cooperate. Such less structured and more authentic (real-life) assignments rarely

have a single correct answer and often permit the use of a variety of aids and resources.
Research skills are an obligatory part of the final exams of Dutch secondary education since 1998. These skills are considered as cross-curricular, covering the natural and social science subjects, mathematics, and even aspects of language such as argumentation and coherent writing. Also a change in didactical approach is advocated, towards more independent learning by students and a more coaching role for teachers. Research skills are included in other national curricula as well, both in a more restricted form, for example limited to data gathering and analysing skills (e.g. Germany, France) and in a more comprehensive form, comparable to the Dutch skills curriculum (e.g. Great Britain) (Steenstra, 2003).
Teaching research skills asks for new teacher knowledge and skills. Developing new teaching practices takes time and can only succeed if fitting in with the teachers' habits, beliefs, goals and resources. Teachers deserve support and educational researchers, curriculum developers and teacher trainers should take an active part in the process of development and implementation.
Teachers in a number of subject areas have long been familiar with 'practical work', variously specified in the literature as practicals, fieldwork and lab experiments (Gee & Clackson 1992; Hodson, 1992; Kirschner, 1992; Duggan & Gott 1995; Meester & Kirschner, 1995; Thair & Treagust, 1997). Meester and Kirschner (1995) differentiate between 'cookbook' experiments (conducting standard tests in a laboratory environment), practical work (more freely experiencing the activities of a researcher), and experimental seminars (learning from discussions to interpret experimental data correctly). The meaning of the term 'research' as used here is broader and comparable to what is denoted by Duggan and Gott (1995) as 'investigation': research in which the approach and results are not obvious (see also Lock 1990b; Wiggins 1993). While developing an approach for teaching research skills teachers will have to decide how to conceive research.

2.4.2 Research skills as a domain for performing, learning, teaching and assessing

Analysing the potential of theoretical knowledge to describe and inform school practice De Corte (2000) made a useful distinction between four domains of theory: of expertise, of acquisition (to be called here: learning), of intervention (here: teaching), and of assessment. We link up with this.

2.4.2.1 Theory of expertise

Doing research is a broad and complex skill domain. It is a broad domain because, apart from skills in the narrow sense, also an appeal is made to knowledge, attitude, reflection and self-regulation. Such broad competences should form an integrated whole. It is also a complex domain because different sub domains are involved, in mutual coherence, organized and directed towards a specific aim. Doing research requires not only subject-specific knowledge but also conceptual, procedural and contextual knowledge and skills of a more or less general kind, and certain attitudes (towards careful execution, proper processing of data, consistent reasoning, and critical evaluation).
A distinction can be made between skill as cognitive capability and skill as task-oriented performance. If we conceive research skills as representing the cognitive capabilities necessary to carry out research and doing research as consisting of a series of steps (problem, design, execution, and so on), it would be convenient if these steps represented the skills. However, different steps can presuppose the same cognitive activities and for any given step more than

one cognitive activity may be necessary (Lock 1989, 1990a; Brown, Moore, Silkstone & Botton, 1996).
If there would be a strong coherence between the skills, there could exist one overall research skill. Studies using various conceptualisations have shown that this does not appear to be the case. Lock (1990a) employed the classification: planning, manipulation, observation, interpretation, report, self-reliance. Brown et al. (1996) distinguished: designing and planning, using and organising techniques and apparatus, observing and measuring, interpreting experimental observations and data, recording and reporting. Both studies showed that data collection and report writing are quite different skills. Only designing a research and interpreting the results were closely related. This association might be explained by the fact that both research steps are closely related to the content of the research problem and the knowledge domain concerned.

2.4.2.2 Theory of learning

An adequate theoretical model for describing and explaining the development of broad and complex skills such as research skills is not yet available. Well-known models for skill development involve psycho-motoric skills or simple cognitive skills (Fitts & Posner, 1967; Anderson, 1982) and are not well applicable in this context (Shuell, 1990; Stokking & Voeten, 2000; Stokking et al., 2004; Veenhoven, 2004).
The first question discussed above, whether research skills could be represented by research steps, is also relevant here but can be additionally addressed in another way. Research steps are sequentially ordered activities and are particularly meaningful in the context of a research project as a whole. Conceiving research steps as contextually meaningful tasks to be carried out (rather than as skills) fits well into a constructivist approach. In such an approach, the cognitive-psychometric view of skill development is abandoned, little attention is paid to theoretical modelling of the process, and the question whether the research steps also represent skills (in a cognitive-psychometric sense) becomes less important.
A constructivist approach of teaching research steps in the context of a complete, authentic research task is in line with common conceptions of research and suggestions for teaching and assessment. Hodson (1992) argued that science is holistic and can only be taught, learned and assessed as such. The idea of a step model of research is, according to Hodson, not without risks: the steps are not fully differentiated, and they are dependent on both the subject matter and the context in which the research is carried out. Kirschner (1992) also warned against limiting research to practical execution, and advocated reflecting on the interdependency of conceptual and procedural knowledge and on the overall quality of the research. Similarly, Gott and Duggan (1996) focused on comprehensive concepts of evidence and White and Frederiksen (1998) used overall criteria such as systematic work, meticulous reasoning and clear communication. In sum, it is questionable whether conceiving research as consisting of a series of steps could provide enough connection to the conceptual, procedural and contextual knowledge needed for doing and learning to do research (Smits, 2003). Obviously, all this does not alter the fact that it can be useful to (firstly) train certain sub skills, and teachers do have certain degrees of freedom here.

2.4.2.3 Theory of teaching

Main components of the learning environments teachers can create for their students are the goals and objectives, the instruction and guidance, the tasks or assignments for the students,

and the assessment. We discuss these components.

Goals. While teaching students how to do research, teachers can strive for different goals (Hodson, 1992; Kirschner, 1992; Duggan & Gott, 1995; Meester & Kirschner, 1995; Thair & Treagust, 1997): motivate students; let students work independently and in cooperation; illustrate phenomena and concepts, let students have hands-on experiences and contribute to the development of their subject knowledge; train them in practical skills (using instruments, making graphs, etc.); let them work as a researcher, systematically executing certain activities; let them think as a researcher, reasoning on theoretical concepts, research design, instruments, and data; and let them experience the 'ethos' of research and science.

The teachers who teach research to develop students' subject knowledge probably want to choose the topics themselves, fitting in with the exam requirements. It could be motivating for students and enhancing their learning profits to only present global topics, leaving the students free in their choice of specific topics and questions. However, as teachers might have little knowledge about some topics they could encounter problems in coaching students properly.

The teachers who primarily want to develop students' research skills also have to choose between structuring the process for the students or giving them more freedom of choice. The teachers can teach purposively the knowledge and skills needed for the design and execution of research, for example by presenting a standard scheme of steps and discussing it. The teachers can also choose to approach research as a complex task that the students have to learn by working on it in a realistic way. In this approach the teachers will leave more to the students, possibly leading to a larger call for guidance.

Instruction and guidance. Teachers having to decide how to teach research have to make choices on several dimensions. Firstly, they have to decide the context in which to teach the students. Nowadays it is common to assume that broad, complex skills such as research skills should preferably be taught in authentic contexts. Connected to this is a second choice: either teach the separate sub skills, followed by integration, or focus right from the start on (meaningful) wholes. There is evidence that the latter would be preferable because it fits in better with the way complex skills develop (Hodson, 1992). Thirdly, teachers can choose to install a learning sequence from easy to difficult, concerning subject content (conceptual) and/or research method (procedural).

Fourthly, teachers will have to decide how much to structure the learning process through assignments and guidance, and how much to leave to the students themselves. Obviously teachers might vary the amount of structure, fitting in with the complexity of the research and with the students' experience. Students in secondary education have little experience in doing research and thinking scientifically. A very structured learning environment, in which the students almost have no space for making choices, is not very obvious (Roth & Roychoudhure, 1993). The same goes for a very open learning environment, in which the students are thrown upon their own resources (Fradd & Lee, 1999). Working on complex assignments students learn best by 'guided reinvention' (not too open, not too structured), because both extremes lead to superficial learning (Kanselaar, Galen, Beemer, Erkens & Gravemeijer, 1999).

Additionally, the teacher can give the students opportunities to cooperate. Teachers can let students cooperate for practical reasons, such as reduction in the number of assignments to be assessed. Cooperation can also be seen as a goal on its own because students in their future life will also have to work in teams. Additionally, cooperation can be motivating the students and thereby furthering their learning process. Cooperation can also further students' learn-

ing directly if it triggers discussion and reflection on the research assignment in progress. Finally, teachers have to make decisions about the coaching they will give the students, the frequency, amount and type of feedback, and the extent to which they stimulate reflection and self-assessment (Collins, Brown & Newman, 1989).

Assignments. In a learning environment featured by independent learning, as expected in Dutch upper secondary education, a substantial part of students' learning takes place while they work on assignments. The main features of assignments which can improve the work and the learning of students are well known: the problem is recognizable for them and is seen as relevant; the problem is challenging (not too easy nor too difficult); the work should result in a final product; students are allowed to cooperate; there is a variety of tasks to be done; and there is room for students to make choices of their own (Blumenfeld, Soloway, Marx, Krajcik, Guzdial & Palincsar, 1991). Dutch teachers are used to vary the scope and difficulty of assignments and they often believe this to be easier than to vary the degree and nature of guidance (Vankan, Van Nijnatten & Ankoné, 1999).

Assignments can include various information and options: the choices that students still can or have to make; the sources and facilities to be used; the time planning; the way of cooperating; the guidance they can ask or will be given; the moments of assessment; and the assessment criteria. Especially the assessment criteria can be used to explain the learning aims and their importance and can make meaningful feedback possible. Clear goals and realistic feedback can be important motivators (Crooks 1988; Snow & Lohman, 1989; Boekaerts, 1991; Rowe & Hill, 1996; Wolfe, 1996).

Assessment. Assessing research skills by means of realistic research assignments is in tune with new assessment strategies known as performance assessment and authentic assessment. These strategies lead to more integration of teaching and assessment, in which the emphasis is less on the summative function of assessment and more on the formative function, giving feedback and supporting the learning process (Shavelson, Baxter & Gao, 1993).

In addition, teachers have to make a number of specific decisions with regard to how often and what to assess and to the assessment criteria, scoring, norms, and quality assurance. The assessment criteria will depend on the research steps or research skills emphasized by the teacher and on the teachers' goals. Teachers also must decide how to score: more analytically or more holistically.

2.4.2.4 The principle of alignment

According to the instructional design literature, to facilitate learning the different components of the learning environment have to be consistent with each other. There are indications that better results are achieved in educational settings where the curriculum and the assessment methods are aligned (Cohen, 1987). Therefore, an appropriate learning environment for research is not just a matter of adequate separate components. All components should address the same agenda and support each other: the objectives are clearly stated in terms of the research skills needed, the teaching methods support students conducting the research, and the assessment addresses those same research skills. With the term 'constructive alignment' Biggs (1996) proposes a marriage between constructivism as a framework for instructional design and the principle of alignment. The principle of constructive alignment means that teachers are expected to be clear in the goals they pursue; to choose for student-oriented instruction and guidance; to work with authentic assignments fitting with the students' level; to be authentic in their assessments; and to make these components fit together.

2.4.3 Teachers' conceptions of research: separate steps or integrated whole

In spite of the critical remarks made above about conceiving research skills in terms of separate steps, teachers' beliefs and practices often are based on this conception, and understandable so. It is hard for teachers to direct their teaching to the underlying cognitive skills and also many teachers do not have much knowledge of or experience in doing research themselves, so they gladly accept the idea of research as consisting of concrete, specified steps (Van Rens & Dekkers 2000; Van Tilburg & Verloop, 2000). This gives them something to hold on and the Dutch final secondary education exam program nicely fits in here. Analysing the exam requirements for research skills in the subjects Physics, Biology, Economics, History, Geography, Mathematics and Dutch, it was found that these requirements have been formulated in terms of research steps and that the requirements of the different subjects can be covered by a common set of ten steps (Stokking & Van der Schaaf, 2000): 1. identify and formulate a problem using subject-specific concepts; 2. formulate the research question(s); 3. make and monitor the research plan; 4. gather and select data; 5. assess the value and utility of the data; 6. analyse the data; 7. draw conclusions; 8. evaluate the research; 9. develop and substantiate a personal point of view; 10. report and present the research.
If teachers have limited knowledge and experience themselves and if they concentrate their teaching and assessment on the execution of these different steps, the effect could be that students think that doing research is taking the steps, as a fixed and not very inspiring sequence of activities. Organizing the learning environment within a constructivist framework could prevent this. From a constructivist point of view the development of complex research skills demands a learning environment in which: students learn in a gradually, independent, active and cooperating way and in a meaningful context; the teaching is focused on coaching and guiding the students' development; different subject contents can be integrated; and assessment is an integral part of the learning environment (De Corte 2000). The intended result is that students learn to monitor the consistency and scope of the research as a whole. Still, as weaknesses in certain research steps usually cannot be compensated by high quality in other steps, each sub skill has to be mastered to a certain level. So, teachers have to be able both to reach with their students a minimum level in each research step and to go beyond the separate steps in their feedback and assessment. The choices teachers make will depend on their ideas about the meaning of research within their subject and their own experience and skills in doing research, on the expectations they have about the interests of their students, on the school's policy and facilities (schedule, computers, etc.), on their experience in teaching research skills, and on the specific bottlenecks they are confronted with.

2.4.4 Teaching research skills in practice, results of a survey study

Stokking and Van der Schaaf (2000) conducted a study into upper secondary natural and social science teachers' practices in teaching their students research skills. One of the research questions was: which choices do teachers make with regard to the components of the learning environment? The goal of the study was to further our understanding of the aspects on which teachers need support.
In the fall of 1998 and the fall of 1999 two written surveys were conducted amongst teachers who were also coordinating their subject department. In 1998 the response came from 49 teachers in 40 schools of all 122 schools that had started the implementation of the new

exam programme in 1998. In 1999 the respondents were 165 teachers in 123 schools from a random sample of 300 from all (other) schools in The Netherlands that had started the implementation in august 1999. In both surveys the responding teachers had an average of 20 years of teaching experience. In the first survey the teachers were additionally asked to submit some assignments given to their students and related material and to answer a short questionnaire with questions about these assignments.

The two surveys and the short questionnaire about the submitted assignments and material consisted of both single questions and Likert-type scales. The questionnaires covered the four main components of learning environments discussed above and some additional aspects. The two questionnaires overlapped in content. Some areas were sufficiently addressed by the first survey and therefore not included in the second one; subsequent literature study led to the inclusion of some alternative and additional questions in the second one.

The main results of the study were that by giving their students research assignments, teachers want them to develop subject knowledge, to practice independent learning, and to develop research skills. The majority of the teachers first teach certain sub skills and about half of the teachers first give structured assignments and than more open ones. In their coaching, most teachers emphasize the subject content of the assignments (about 80% to a reasonable or high degree) more than the research method (about 65%), and they use a limited range of coaching approaches, mostly posing and answering questions. Most teachers give feedback in small groups or individually and only sometimes by whole class discussion. The teachers do not much stimulate reflection and self-assessment.

Students get on average two or three research assignments per year, most of them developed by the teachers themselves. Most teachers require assignments to fit close to the students' level, to contain enough subject content, to be sufficiently challenging, and to produce testable results. Most teachers give rather structured assignments and let the students work on them in cooperation.

The teachers differ quite a lot with respect to the aspects of assessment: the moments at which they assess students' research work, the information they give in advance about the assessment, the scoring approach, the weighing of aspects, and the standards they use. They attach the most weight to the subject content (60% of the teachers). They make little use of co-assessment (by colleagues or by students).

Most teachers think that research in their subject is not, slightly or only reasonably different from research in other subjects, and they inform their students about the similarities. The teachers vary in their experience with different types of research by students. According to the teachers their students have quite a lot of problems with all research steps and the teachers themselves have problems due to lack of time and shortage of adequate assignments and experience deficits in assessment competences. Most teachers cooperate little with colleagues.

The teachers seem to align the different components of the learning environment (aims, instruction and guidance, assignments, assessment) to a reasonable degree. The goals they strive for are to develop subject knowledge, to further independent learning, and to develop the research skills. The emphasis they give to the subject content is reflected in their requirements concerning the assignments that these should have enough content, should fit in with the exam requirements and cover the content of the lessons, and also in their assessment practice, in which for most teachers the aspects which have to do with the subject content (research question, conclusions, difficulty level) weigh the most. The emphasis on independ-

ent learning is reflected in the fact that about half of the teachers let gradually decrease the extent to which they structure the assignments and the fact that something more than half of the teachers regularly or often let the students cooperate.
The way teachers approach research, stepwise or holistically, differs between the components. Looking at the goals, 90% of the teachers want to teach students first certain sub skills. In the context of the instruction and guidance this percentage falls towards 70%, and during the assessment still 50% of the teachers firstly judge the separate research steps before formulating a total judgment.
Teachers welcome the conceptualisation of research in terms of steps. However, research steps are not skills and the teaching of research in terms of steps, taking into account the teachers' limited knowledge about and experience in doing research (Van Rens & Dekkers, 2000; Van Tilburg & Verloop, 2000), could be at the expense of the view that research is (and should be) an integrated whole and at the expense of the students' motivation (which might be avoided by not structuring the assignments too much, by letting students make their own choices, and by stimulating cooperation between students). The fact that the teachers in their coaching and especially in their assessment are less precisely distinguishing the steps than while formulating their goals is probably related to their experience of lack of time.

2.4.5 Implications for the research in this thesis

Our conclusion is that teachers could indeed use support, especially by furthering their knowledge about research, by supporting the development of more adequate assignments (decreasing the time they need for developing assignments and thereby making time for giving feedback and coaching students), by stimulating their skills (in giving feedback, stimulating reflection, and assessing), and by furthering the cooperation between teachers (Little, 1990). Also, teachers could be stimulated and supported to develop their knowledge about relevant theories of expertise (i.e., their conception of research), learning, teaching and assessment, which could inform their views, discussions and decisions (Eraut, 1994; Roehrig & Luft, 2004).
Additional research is needed into the optimal balance between starting with teaching sub skills or focussing on research as a whole right from the start, into the characteristics of effective assignments, into the relationships between developing research skills and developing subject knowledge, and into cross-curricular cooperation between teachers. Concerning the quality of teachers' teaching students research skills, the topic of this thesis, there is a need for explicit standards setting acceptable minimum quality levels. The development of such standards is addressed in the following chapters.

3 Developing content standards for teaching students research skills using a delphi method

This chapter is submitted as: Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (2005). Developing content standards for teaching students research skills using a delphi method. Manuscript submitted for publication

Abstract

The increased attention for teacher assessment and current educational reforms ask for procedures to develop adequate content standards. For the development of content standards on teaching research skills, a Delphi method based on stakeholders' judgments has been designed and tested. In three rounds, 21 stakeholders judged and revised content standards. The support for the standards increased over the rounds. The method resulted in nine content standards with a high degree of support and consensus. Qualitative analysis of stakeholders' comments and a Homals analysis showed that the stakeholders differ in the perspectives and preferences from which they judged the standards. The Delphi method proved to be an adequate procedure for developing teaching content standards based on stakeholders' judgments.

3.1 Introduction

The quality of education highly depends on what teachers do (the teaching tasks they perform) and on their competences to adequately fulfil these tasks. To enhance good teaching, in many countries teaching standards have been produced and discussed. Given the elusive and contested nature of what to count as good teaching, these standards can only be formulated in a relative way. It has become widely accepted that standards are shared by a community of practitioners at a certain time and place, and thus are socially situated constructs.

The need for adequate procedures for the development of standards by practitioners is urgent for at least three reasons. Firstly, in many countries, educational reforms expect teachers to change their practices towards more constructivist views on learning and teaching. Inspiring standards that can be used as a descriptive framework to rely on when dealing with the reforms can assist teachers. Since reforms can only succeed if they fit in with the teachers' beliefs, everyday practices and resources, teachers at least should have a voice in developing such standards. Secondly, as constructing teaching standards is a resource for deliberation on what it means to be a good teacher (Ingvarson, 1998), this corresponds with the tendency to give teachers as a collective more responsibility in monitoring the quality and further development of their profession. However, in reality teachers still exercise little control in these

matters. Thirdly, there is an increased attention for teacher assessment. Assessment can be formative, aiming at further professional growth, or summative, e.g. for certification or merit pay. Assessments are designed based on standards and assessors judge whether teachers' performances meet the standards.
Teaching standards are of two sorts: content standards and performance standards. Content standards are the criteria that conceptualise the desired tasks teachers have to fulfil in specific contexts and the competences needed to do this. Performance standards are the cut-off scores indicating to what extent teachers should meet the criteria.
In this chapter, we focus on the development of content standards. The study is part of a larger project on the assessment of social sciences teachers (economics, geography, history) in pre-university education in The Netherlands (ages 16-18), comparable with undergraduate college level in, for example, the American system. We focussed on a recent Dutch educational reform that is representative for the current change in many countries towards more constructivist views on learning that emphasize developing students' skills in an active, self-directed and collaborative way. In a number of national curricula this has led to the inclusion of research skills, both in a more restricted form, for example limited to data gathering and analyzing skills (e.g. Germany, France) and in a more comprehensive way regarding the whole cycle of doing research (e.g. The Netherlands) (Steenstra, 2003).
During the last decades, different methods for generating content standards have been used (Berk, 1996), all focussed on stakeholders' discussion. Main disadvantages of such methods using face-to-face communication are interdependences between panel members, such as domination, group thinking and polarization, distorting the results and threatening their reliability (Riggs, 1983). The Delphi method is a promising alternative. This method is task-oriented, reduces social influence processes, and has a cyclic character resulting in successive refinement in the answers and generating convergence (Scheele, 1975). Whereas the Delphi method has several applications, in most cases it is used as a survey technique in which panel members respond anonymously and independently to a set of questions in successive rounds. After each round the answers are analysed and feedback is given. The aim of this procedure is to realize a structured and iterative written discussion and to foster consensus between the panel members (Linstone & Turoff, 1975).
However, the feasibility of the method for generating content standards is still largely unknown. In this study we inquire into the advantages and disadvantages of the Delphi method for the development of teaching content standards. The main questions are: How can a Delphi method be used and optimised as a procedure to develop content standards for teacher assessment in the context of educational reform? Which content standards can be set regarding teaching students research skills?

3.2 Validation strategies for content standards

Today, the development of content standards is commonly based on comparisons to a group or to a domain. The scores are compared with those of a well-defined group of other candidates who took the assessment before or at the same time (norm-referenced), or with a set of specified constructs derived from a properly formulated performance domain (criterion-referenced). The first approach is criticized being not clear about the relationship between the assessment outcomes and the fact that the outcomes can be interpreted in many ways. The latter approach is criticized on the premise that performance domains can be specified

so precisely that items for assessing the domain could be sampled automatically and without doubt. Nevertheless, the second approach suits best the purposes of our study, because it enhances the interpretation of assessment outcomes while referencing to the domain that is assessed.
However, since assessments are used in social settings, assessment results are only relevant with a reference to a particular population. Consequently, any criterion-referenced assessment is attached to a set of norm-referenced assumptions (Angoff, 1974). This increases the ambiguous character of content standards and several researchers concluded that such standards are best based on standard setters' 'connoisseurship', that is "the ability to make fine-grained discriminations among complex and subtle qualities" (Eisner, 1991, p. 63) (see chapter 2). Nowadays this ability is also often associated with members of a community of practitioners (cf. Dylan, 1996), or, in a broader sense, to stakeholders.
Kane (1994) describes three types of strategies for demonstrating the validation of content and performance standards which are developed by stakeholders (see also chapter 4): i) validation checks based on external information, consisting of comparisons of the content standards with external information about the desired teacher practices; ii) validation checks based on internal information, concerning the consistency of the procedure used, which indicates the reliability. The word 'indicates' is used deliberately, because variation in panel members' judgments does not necessarily mean that the procedure as such is not reliable, since the panel members could interpret the content standards differently; iii) procedural evidence, consisting of demonstrating the soundness of the procedure followed, including the selection of the panel members involved in the process. We included all three types of evidence in our study.

3.3 Method

3.3.1 The development of a preliminary set of content standards

We use the Delphi method to modify a preliminary set of content standards. We developed this preliminary set in a four-step procedure. The first step was the specification of the domain to be measured (Berk, 1980). Given the fact that there are several ways of good teaching and that teaching is context-bound, such a domain cannot be clearly demarcated. For that reason we used 'domain' only as an indicator of the information for which the standards are expected to account (Messick, 1989). We conducted a preparatory literature study to develop a general framework for describing teaching tasks. Secondly, we conducted a literature study into teaching students research skills. This was followed by an empirical study into leading edge teachers' practices regarding developing students' research skills (see chapter 2). Thirdly, based on the results of these studies we made an initial selection of relevant teaching tasks. Fourthly, we described a preliminary set of content standards around the tasks.
1. General framework for describing teaching tasks. Currently teaching is viewed as a complex activity in which teachers' practices are inextricably bound up with teachers' unique cognitions, personality, and the context at hand (Shulman, 1986; Calderhead, 1996). Though teaching is highly personal, we believe that teaching as a profession at a general level can be characterized by elements that at least groups of teachers have in common (e.g. teachers who teach certain subjects to students of similar age level, as in our study) (cf. Verloop, Van Driel & Meijer, 2001).

A central element of teaching is that teaching tries to bring about learning on the part of the student (Fenstermacher & Richardson, 2000). Therefore, teachers conduct professional tasks. Reynolds (1992), after having analysed tasks from job descriptions, teachers' practical knowledge (Shulman, 1987), performance assessment systems, and duties of teachers (Scriven, 1988), synthesizes them into four domains. These include: tasks before teaching (e.g. planning instruction), tasks during teaching (instructing, coaching, assessing), tasks after teaching (e.g. reflecting on own teaching), and administrative tasks (e.g. management tasks).
As described in chapter 2, executing teaching tasks is mediated by teacher cognitions. Putnam & Borko (1997), using Shulman (1986, 1987) and Grossman (1990), classified the content of teachers' cognitions into: general pedagogical, subject matter, and pedagogical content knowledge. Grossman (1995) added the categories of knowledge about the (school) context and the self. Especially the category of pedagogical content knowledge is viewed as a vital domain of teaching, because it refers to essential knowledge and skills that are unique to the teaching profession. It covers the translation of subject-matter knowledge into teachers' acting (Van Driel, Verloop & De Vos, 1998). In our study, pedagogical content knowledge about teaching students research skills is central: what a teacher knows, what a teacher does, and why, when teaching research skills in their subject (Baxter & Lederman, 1999).
Content standards should leave enough room to fit diverse teaching contexts, but on the other hand they should be precise enough to discriminate between suitable and unsuitable performances (Katz & Raths, 1985; Kagan, 1990). A formulation of content standards in terms of dispositions yields a middle 'size' between too extensive and too narrow descriptions. Dispositions can be seen as characteristics of teachers that summarize the trend of their intentional actions (Katz & Raths, 1985). They address the need for the likelihood that a teacher will perform in a certain way and goes beyond the idea that a teacher could have certain knowledge and skills but not employ them. Dispositions suggest what a teacher is likely to do, rather than what a he or she can do at peak performances, and therefore should be part of the content standards. Having a disposition includes the possession of relevant skills and knowledge. Dispositions also cover components of psychological nature (e.g. enthusiast, accepting, stimulating, sensitive, flexible) and moral and ethical kind (e.g. honesty, respectful, trustful) (cf. Katz et al., 1985; Reynolds, 1992; Fenstermacher et al., 2000). In this study we use the term competences to denote teacher dispositions, because both concepts are almost analogous and the former is more common.
Elaborating on this, we describe teaching tasks in terms of task domains (e.g. tasks before, during, and after teaching) and contents or cognitive domains of teaching (e.g. general pedagogical knowledge, pedagogical content knowledge, knowledge of the school context). The combination of these two dimensions generates a matrix. The cells in the matrix contain a specification of the tasks teachers have to fulfil. We formulate the standards around these tasks, in terms of teachers' dispositions to fulfil the tasks. See *Figure 3.1*.
2. The domain of teaching students research skills. In a literature study Stokking & Van der Schaaf (2000) identified, based on the literature, main components of constructivist learning environments that teachers can create for their students when teaching research skills (see chapter 2). These are goals and objectives, instruction and guidance, tasks or assignments for the students, and assessment. According to the instructional design literature, the different components of the learning environment must be consistent with one another to facilitate learning. Better results are achievable in educational settings where the curriculum and the assessment methods are aligned (Cohen, 1987). Biggs (1996) proposes the term 'construc-

<table>
<tr><td></td><td colspan="3">Knowledge domains</td></tr>
<tr><td>Tasks</td><td>General Pedagogical Knowledge</td><td>Pedagogical Content Knowledge</td><td>Knowledge of School Context</td></tr>
<tr><td>Planning</td><td></td><td>1. GOAL</td><td rowspan="5">8b. COLLAB</td></tr>
<tr><td>Developing</td><td></td><td>2. ASSIGN</td></tr>
<tr><td>Managing</td><td>3. MANAGE</td><td></td></tr>
<tr><td>Instructing</td><td></td><td>4. THINK</td></tr>
<tr><td>Coaching</td><td>6. CLIMATE</td><td>5. TEACH</td></tr>
<tr><td>Assessing</td><td colspan="2">7. ASSESS</td></tr>
<tr><td>Reflecting</td><td colspan="2">8a. REFLECT</td></tr>
</table>

Figure 3.1 Content standards in function of tasks and knowledge domains

tive alignment' as a marriage of constructivism as a framework for instructional design with the principle of alignment. The principle of constructive alignment means that teachers are expected to be clear about the goals they pursue, to choose student-oriented instruction and guidance, to work with assignments that are suited to students' levels of learning, to be authentic in their assessment, and to make these components fit together.

This literature study was the input for an empirical study into leading edge teachers' practices. We executed oral interviews (n = 20) and two written surveys (n = 49; n = 165) with pre-university teachers in social science subjects (economics, history, geography) and natural sciences (physics, biology). Also teacher submitted research assignments and associated information (teachers' goals, ways of coaching, assessment procedures and criteria) (n = 54). The study showed that the main components as revealed by the literature are indeed important aspects of teachers' practices. Furthermore, the study revealed additional aspects that play important roles in teaching students research skills: the collaboration with colleagues (e.g. making arrangements with regard to the teachers' goals, students' assignments and the assessment) and the bottlenecks teachers experience when handling research assignments (e.g. due to lack of time for coaching and assessment, lack of modern media at school, and lack of information resources and other material needed, the support some students get from others such as their parents, and fraud) (see Stokking, Van der Schaaf, Jaspers & Erkens, 2004 and chapter 2).

3. Initial selection of teaching tasks. We selected teaching tasks that emerged to be the most relevant in the domain of teaching students research skills, and that fitted in the matrix of tasks and knowledge domains. Also, the tasks as an indication guarantee for being consistent constructs had to be measured reliably in the surveys. The selected tasks include: setting goals (Cronbach's alpha .62); selecting assignments (.88); instruction and coaching (.86); using different coaching approaches (.63); assessing students' research skills (.83); collaborating with colleagues (.93). In addition, we selected tasks regarding management of the research assignments. Finally, we added the establishment of a safe learning climate. In sum, we derived

a provisional list of these tasks: 1. Formulating goals; 2. Developing assignments; 3. Managing students working on their assignments; 4. Thinking of teaching (preparation); 5. Teaching research skills (instruction); 6. Creating a positive pedagogical climate; 7. Assessing students' research skills; 8. Reflecting (8a) and collaborating (8b) with colleagues.

4. Initial description of content standards. Relying on our data, we closely examined each category and made an initial description per task. For example, our data showed that teachers pose requirements to the research assignments they select. The teachers in general want the assignments to fit close to the students' level, to contain enough subject content, to be realizable in a certain amount of time, to be sufficiently challenging, and to produce testable results, and also offer the students options, and to be approachable in various ways. Furthermore, about 40% of the teachers almost always let the students work on a research assignment in pairs or small groups. We summarized this information in our initial description of a content standard concerning the development of assignments.

Each content standard was thoroughly described, in a half to full page per standard. We described the content standards in a document of 7 pages, consisting of a description in terms of dispositions and a more detailed description to direct the interpretation of the standards (Ryan & Kuhs, 1993). The content standards that were still rather broad were divided into two or more indicators (see *Table 3.2* and section 5.2.3 in chapter 5 for an overview). The content standards were converted into a questionnaire that formed the input to our Delphi study.

3.3.2 The sampling of panel members

In order to produce high quality standards several conditions have to be fulfilled. (a) The selected stakeholders should have (practical) knowledge of the teaching research skills in social science subjects in pre-university education and about everyday teaching practices. (b) The stakeholders have to agree with the way in which the content standards to be set will be used in the assessment of teachers (Putnam, Pence & Jaeger, 1995). (c) Although reproducible content standards may be accomplished with 5 to 10 judges (Norcini, Lipner, Langdon & Strecker, 1987), a larger group will increase the chance that a variety of expertise, experiences, and perspectives is represented. The Delphi method is suitable for 10 to 50 participants (Turoff, 1975).

Heterogeneous groups with different backgrounds and various perspectives typically develop highly acceptable and valuable results (Norcini & Shea, 1997), but if a group is not constructed with sufficient care, the quality of the representation of the population in the group and the resulting quality and acceptability of the content standards will be reduced (Linstone & Turoff, 1975; Sackman, 1975; Clayton, 1997).

Preparing the selection process of the participants, we involved groups with a considerable stake in the results of the study (Clayton, 1997): teachers in geography, history, and economics (n = 3), educational researchers (n = 3), government employees (n = 4), lobbyists (n = 2), pedagogical content experts in geography, history, and economics (n = 4), teacher educators (n = 4), and school principals (n = 2). These individuals were interviewed about the assessment of teacher competences in teaching students research skills. When asked which persons or organizations should participate in the process of developing content standards for summative teacher assessment, most interviewees mentioned practicing teachers, followed by external experts such as teacher educators and pedagogical content experts.

We used the following procedure to make up the Delphi panel. From a random sample of

115 schools we approached the heads of the geography, history, and economics departments and their school principals with information about the study (goals and planning), a description of the judgment task, a request for participation, and a detailed intake form. The intake form contained questions on general characteristics (sex, age, position, years of experience), opinions on students' research in social science disciplines, and their experience with teacher assessment (Putnam et al., 1998).
Fifteen respondents were willing to participate: 3 school heads and 12 experienced teachers: 3 in geography, 4 in history, 3 in economics, and 3 in social science subjects who were also teacher trainers. This low response can be explained by the demanding character of the study in terms of motivation, time, and expertise. In addition, 2 educational researchers, 3 pedagogical content specialists, and a lobbyist were asked to participate. Thus, 21 individuals participated in the Delphi study.

3.3.3 Data collection

The stakeholders received instructions, including a description of the aims of the study, of the way the content standards to be set will be used, and of teaching students research skills. The number of rounds that is needed depends on how stable the judgments of the panel members are and on how quickly they reach an acceptable level of (statistical) consensus (Erffmeyer, Erffmeyer & Lane, 1996; Smith & Simpson, 1995). Often, three rounds are enough. Therefore, in three monthly cycles the panel members judged the content standards and indicators. Their judgments were based on the following four aspects: content relevance, thoroughness of formulation, clarity of formulation, and correspondence of the content standards with everyday teaching practice. Firstly, they answered open questions about the proposal as a whole. Secondly, for each content standard they answered half-open questions on the four aspects. Thirdly, for each content standard and each indicator they were asked to rate the four aspects on a 4-point Likert scale. Fourthly, in order to gain insight into the way the panel members interpret the content standards, they were asked to comment on the ratings they made. Fifthly, they were asked to rank the content standards in order of importance.
To improve the content standards, in each round the panel members were asked for suggestions. After each round the standards were revised, based on their' judgments and comments, after analyses by two researchers. The panel members received written feedback concerning the means and standard deviations of the ratings of the whole group, a summary of the comments made by the other panel members, and an overview of the conclusions per round. In the third round the panel members were also asked to evaluate the Delphi method itself. They answered open and closed questions about their experiences and their level of satisfaction with the method.

3.3.4 Data analysis

After each round a key issue was the decision whether particular content standards should be accepted, revised, or deleted. This decision was based on the validation evidence (see section 3.2), and directed by the panel members' ratings of the content standards and indicators on the four aspects: content relevance, thoroughness of formulation, clarity of formulation, and correspondence with everyday teaching practice. For the analysis of the panel members' judgments per round and across the rounds we formulated the following guidelines.
Firstly, we wanted to know whether the group supports, opposes, or is ambivalent towards the content standards. We used the mean to indicate whether the panel members support

the content standards (4 = strongly supports, 3 = weakly supports, 2 = weakly opposes, 1 = strongly opposes the content standard).
Secondly, we wanted to analyse the degree of consensus between the judgments. We divide the concept consensus into statistical consensus (percentage agreement) about the content standards and substantial consensus about underlying perspectives. We allow panel members to have different valid perspectives which all can contribute to the definition on the content standards. Following the results of a previous Delphi study by De Loe (1995), beforehand we distinguished four levels of agreement: high (70% of the ratings in one category or 80% in two contiguous categories); medium (60% of ratings in one category or 70% in two contiguous categories); low (50% of ratings in one category or 60% in two contiguous categories); none (less than 60% in two contiguous categories). However, consensus can only be assumed if judgments tend to be in one direction only (Sackman, 1975). The skewnesses of the distributions of the ratings were computed to check on symmetry and to check whether the panel members' judgments tended to be in one direction.
Thirdly, the interrater reliability was checked by computing jury alphas, and the consistency between items was examined by computing Cronbach's alphas.
Fourthly, our purpose was to revise the content standards based on the judgments and comments of the panel members until a list emerged that suited the majority of panel members. To test whether the succeeding rounds resulted in increasing support, we compared the ratings between rounds by performing paired sample T-tests. To check whether the range in the answers decreased, we performed paired sample T-tests on the deviation scores.
Fifthly, to gain insight into the most important content standards, the panel members were asked to rank these standards in order of importance. The consistency in the rankings between the rounds was estimated by calculating correlations.
Finally, we analysed stakeholders' written comments for emerging themes and categories. The data of the third round was analysed by homogeneity analysis (Homals) to see whether the panel members could be grouped according to their rankings. Homals projects the differences between panel members concerning their preferences into distances between points, representing the panel members, in a two-dimensional plot.

3.4 Results

3.4.1 Results per round

First round

Results – overall judgment. The panel was asked whether the content standards satisfactorily covered the subject of our study, teaching students research skills. More than half of the panel members thought this to be the case (57%). Only 10% did not agree. Approximately one-third did not directly answer the question, but made some specific remarks. Most noticed that some content standards overlapped. Next, the panel members answered questions about the formulation of the content standards. More than half of them thought the content standards were clear (57%). Several panel members noted that the list as a whole was quite detailed, but judged the formulation of separate content standards as too global or too abstract. Finally, the panel members were asked whether they thought it possible to check whether teachers met the content standards. Almost all agreed (90%).

Table 3.1 Support and consensus of the content standards in rounds 1-3 (n = 21)

	First round				Second round				Third round			
	Support		Consensus		Support		Consensus		Support		Consensus	
	m	sd	a	sk	m	sd	a	sk	m	sd	a	sk
1. GOAL												
Relevance	3,7	.6	H	-1.8								
Formulation (t)	3,3	.9	H	-1.0	3,7	.5	H	-0.9	3,8	.4	H	-1.3
Formulation (c)	2,9	.9	M	-0.5	3,4	.7	H	-0.7	3,7	.5	H	-0.9
Correspondence	2,5	.9	L	0.6	3,0	.9	M	-0.5	3,2	.8	H	-1.2
2. ASSIGN												
Relevance	3,7	.6	H	-1.7								
Formulation (t)	3,3	.8	M	-0.4	3,7	.5	H	-0.8	3,6	.6	H	-1.3
Formulation (c)	3,2	.6	H	-0.9	3,3	.9	H	-1.1	3,4	.6	H	-0.3
Correspondence	2,6	.9	N	0.3	2,9	.9	M	-0.7	3,3	.7	H	-0.8
3. MANAGE												
Relevance	3,5	.5	M	0.1								
Formulation (t)	3,3	.6	M	0.0	3,5	.6	H	-0.6	3,4	.6	H	-0.5
Formulation (c)	3,3	.6	M	-0.2	3,5	.5	H	0.2	3,6	.5	H	-0.3
Correspondence	3,0	.8	N	0.3	3,0	.8	M	-0.6	3,2	.8	H	-0.6
4.THINK												
Relevance	3,1	.7	M	-0.6								
Formulation (t)	3,0	.6	H	0.0	3,8	.4	H	-1.3	3,7	.6	H	-1.6
Formulation (c)	2,7	.7	N	0.7	3,6	.5	H	-0.4	3,5	.8	H	-1.9
Correspondence	2,1	.7	N	0.0	3,1	.9	M	-0.8	3,2	.8	H	-0.7
5. TEACH												
Relevance	3,4	.6	H	-0.4								
Formulation (t)	3,1	.7	M	-0.1	3,5	.5	H	0.0	3,6	.5	H	-0.3
Formulation (c)	3,1	.7	M	-0.1	3,5	.5	H	0.0	3,6	.5	H	-0.3
Correspondence	3,0	.9	L	-0.1	3,0	.8	H	-1.4	3,2	.7	H	-0.6
6. CLIMATE												
Relevance	3,6	.6	M	-1.3								
Formulation (t)	3,1	.8	L	-0.5	3,5	.7	H	-1.1	3,5	.7	H	-1.0
Formulation (c)	3,3	.7	M	-1.0	3,5	.6	H	-0.9	3,6	.5	H	-0.6
Correspondence	2,9	.6	M	-0.7	3,0	.5	M	0.1	3,4	.6	H	-0.5
7. ASSESS												
Relevance	3,3	.8	M	-0.6								
Formulation (t)	3,1	.7	M	-0.1	3,4	.6	H	-0.8	3,4	.5	H	0.3
Formulation (c)	3,3	.8	M	-0.4	3,5	.7	H	-0.6	3,5	.7	H	-1.9
Correspondence	2,4	.8	H	-0.3	2,8	.8	H	-0.2	3,1	.8	H	-0.7
8. REFLECT												
Relevance	3,3	.7	H	-0.4								
Formulation (t)	3,1	.7	M	-0.2	3,6	.5	H	-0.4	3,6	.6	H	-1.3
Formulation (c)	3,1	.7	M	-0.1	3,6	.6	H	-1.0	3,7	.7	H	-2.8
Correspondence	2,1	.6	H	-0.1	2,8	1.0	L	-0.4	3,2	.8	H	-1.1
9. COLLAB												
Formulation (t)					3,3	.7	H	-0.4	3,7	.5	H	-1.0
Formulation (c)					3,3	.6	H	-0.2	3,7	.6	H	-1.8
Correspondence					2,7	1.0	L	-0.5	3,0	.9	H	-0.6

NOTE. m = mean (1 = weak support, 4 = strong support); a = agreement (H = high, 70% of the ratings in one category or 80% in two contiguous categories; M = medium, 60% of ratings in one category or 70% in two contiguous categories; L = low, 50% of ratings in one category or 60% in two contiguous categories; N = none, less than 60% in two contiguous categories). sk = skewness; t = thorough; c = clear.

Results per content standard. Table 3.1 gives an overview of the mean ratings per content standard. The panel supported the relevance of most content standards. It weakly supported the thoroughness and clarity of the formulation, and did not support the correspondence with everyday teaching practice of most content standards. The degree of consensus varied strongly between the content standards, but for most standards it was mediocre.
Revision. The revision mainly consisted of reformulating the content standards in a less abstract manner. This was partly realized by adding illustrations and by categorizing the standards more clearly. In addition, the standard 'Reflecting on the program and on personal actions, and collaborating with colleagues in teaching students research skills' was divided into two standards. Note that in rounds 2 and 3 the content standard 'Reflecting on the program and on personal actions, and collaborating with colleagues' was divided into two content standards. The positive answers as to the relevance of the standards indicated that it was not necessary to repeat the relevance question, and so we omitted this question in the next rounds.

Second round

Results – overall judgment. The panel was asked open-ended questions about whether they thought the reformulation of the content standards to be an improvement. Almost three-quarters of the panel members (71%) thought this to be the case. In summary, the panel thought that the content standards resulted in a better survey and were formulated more realistically. Most standards were judged as being to the point; some standards were seen as too elaborate.
Results per content standard. In the second version, 'Collaborating with colleagues' had been added as a new, separate content standard. Panel members were asked whether it was sensible to distinguish this standard separately. The majority thought this to be so to a reasonable (34%) to strong (52%) degree. *Table 3.1* shows that the panel members judged the content standards in the second round to be, on average, formulated thoroughly and clearly to a reasonable to strong degree. The correspondence with everyday practice was judged as reasonable.
Revision. The most important conclusion from the second round was that the content standards had to be formulated in an even more condensed and concrete form. The resulting changes concerned more compact formulations of the content standards and the addition of a 'shortlist' with a summary of the content standards.

Third round

Results – overall judgment. The panel was asked whether they thought the reformulations were an improvement. Three-quarters (76%) agreed and 14% partly agreed. The formulations were evaluated as more concise, clearer, a better match to the practice of teaching, more logically constructed, less redundant, and linguistically more correct. Two-thirds of the panel (62%) thought all elements were to the point.
Results per content standard. Table 3.1 shows that the content standards on average were judged to be thoroughly and clearly formulated and corresponding to everyday practice to a reasonable or strong degree. The average scores for thoroughness and clarity were all above 3,3. The average scores for corresponding to practice ranged from 3,0 to 3,4. On all aspects with regard to all standards, the degree of consensus was high.

3.4.2 Interrater consistency and scale reliability

The panel as a jury was rather consistent. *Table 3.2* shows the Jury alphas in the third round per content standard and per indicator. Overall, the agreement proved to be sufficient. The panel members only disagreed on standard 6 (CLIMATE). The Jury alphas in the three rounds for all content standards and indicators together were .53 (first round), .72 (second round) and .65 (third round).

The judgments about the thoroughness and clarity of formulation and the correspondence with teaching practice became more consistent across the rounds; see *Table 3.3*. The jury alphas were satisfactory in each round and increased over time. The relevance was only judged in the first round and the alpha was moderate, which is obvious for this aspect in a first round.

With regard to each content standard and indicator the panel members gave judgments on the relevance, thoroughness of formulation, clarity of formulation, and correspondence with everyday teaching practice. In each round regarding each standard and indicator, the judgments on these aspects were analysed for consistency and to ascertain whether, in the case of several indicators, the indicators made up a sufficiently reliable scale or had to be distinguished separately. We limit ourselves to reporting on the third round. *Table 3.3* shows the results of these item analyses.

Cronbach's alphas of all standards without separate indicators (standards 1, 4, 5, 6, 8, 9) were sufficiently high. Amongst the standards made up of different indicators, standard 3 (MANAGE) formed a sufficiently reliable scale. Standard 2 (ASSIGN) was only a weak scale. It seems sensible to distinguish its indicators separately but we nevertheless decided to keep the scale. The alpha of standard 7 (ASSESS) was very low, while the separate indicators had sufficiently high alphas.

Table 3.2 Internal consistencies (Cronbach's alphas) and interrater consistencies (jury alpha) in panellists' judgments of the content standards and indicators, third round (n = 21)

		Cronbach's alpha	Jury alpha
1.	GOAL	.67	.87
2.	ASSIGN	.56	.69
a)	selecting short-term goals	.78	.87
b)	selecting appropriate contents	.82	.91
c)	selecting an appropriate form	.58	.78
3.	MANAGE	.73	.62
a)	structuring time and room, making resources accessible	.72	.45
b)	offering insights into working and assessment procedures	.73	.74
4.	THINK	.75	.71
5.	TEACH	.74	.82
6.	CLIMATE	.63	.06
7.	ASSESS	.06	.85
a)	explicitly stating purposes of the assessment	.90	.47
b)	choosing an adequate assessment model	.83	.44
c)	conducting fair assessment practices	.71	.74
d)	communicating inferences and results to students	.62	.67
8.	REFLECT	.84	.84
9.	COLLAB	.61	.89

Table 3.3 Internal consistencies (Cronbach's alphas) of the aspects across content standards per round (n = 21)

	First round			Second round			Third round		
	m	sd	α	m	sd	α	m	sd	α
Relevance	3,33	.45	.61						
Formulation (thorough)	3,21	.41	.82	3,55	.33	.88	3,63	.29	.91
Formulation (clear)	3,10	.43	.68	3,43	.35	.91	3,61	.38	.86
Correspondence	2,46	.82	.82	2,93	.76	.90	3,24	.60	.94

3.4.3 Results across the rounds

The panel's support increased across the rounds. The panel members tended to rate the standards in the successive rounds as being more thoroughly and clearly formulated and more corresponding with practice (see *Table 3.1*). Paired sample T-tests between the ratings of the first and third rounds showed that the standards 1 (GOAL), 2 (ASSIGN), 3 (MANAGE), 4 (THINK), 5 (TEACH), and 6 (CLIMATE) were judged significantly higher on 'Thorough formulation' ($p < .05$). The standards 1, 4, 5, and 6 were rated significantly higher on 'Clear formulation' ($p < .05$). The standards 1, 2, 3, 4, 6, and 7 (ASSESS) were rated significantly higher on 'Correspondence with everyday teaching practice' ($p < .05$). The T-tests on the deviation scores were not significant, so the range in the judgments did not change.

3.4.4 The panel's underlying assumptions and perspectives

The written comments of the panel members in the three rounds were transcribed verbatim, resulting in about 40 pages of text. These were then qualitatively analysed. We now summarize the most important perspectives of the panel members in judging the content standards. The abbreviation P (1-21) stands for 'panel member', R (1-3) for 'round', S (1-4) for 'support score', and (-) for 'missing'; thus, the reference P3, R2, S4 stands for panellist 3 in round 2 giving support 4 (strong support).

1. Selecting yearlong goals for students (GOAL). The panel members thought this content standard to be relevant, for different reasons: because it is important for students to have a cumulative curriculum, and because collaboration between colleagues is important, giving the teacher focus and legitimacy in instruction, coaching, and assessing. The correspondence with everyday practice was judged less high, because teachers operate pragmatically, owing to lack of time and dependence on collaboration within the school, e.g. "Formulated very idealistically. Should be attainable, but it depends also on the department and the school agreements" (P18, R3, S-).

2. Choosing an appropriate assignment (ASSIGN). In the first round almost all panel members thought this standard to be relevant, because "Everything depends on having an adequate assignment" (P11, R1, S4; P13, R1, S4; P18, R1, S4). Otherwise, the practices differ, e.g. "The student's contribution is made to heavy, in practice the teacher will decide how the assignment will be shaped" (P2, R3, S3).

3. Preparing and managing students working on their assignments (MANAGE). Panel members who judged this content standard in the first round as highly relevant, thought organization to be a condition for working on the assignment, e.g. "Classroom management is very important in coaching independent research" (P20, R1, S4). Panel members who provided the judgment 'rather relevant', gave two considerations. The content standard was thought to be

not only the teacher's task but also the student's, e.g. "Of course these things are relevant, but part of it can be managed by the students themselves" (P6, R1, S3), or other content standards were judged as more important, e.g. "A little less relevant than other indicators, because these matters are only organizational and I guess have less influence on learning" (P19, R1, S3). The panel members who thought that students can manage things themselves judged the correspondence of the content standard to everyday practice to be less high in the first and second round, because the standard assumed to too great an extent that the teacher was in control; in the third round, when the standard was reformulated to give the students more control, they thought the standard was more recognizable, e.g. "This is roughly how things are going on in our school" (P6, R3, S3). However, some panel members judged the formulation in the third round to be less recognizable, e.g. "Much too optimistic and formulated far from reality" (P2, R3, S-).

4. Previously thinking of and selecting teaching strategies that support the development of research skills of students (THINK). Several panel members thought this content standard to be relevant because it is a condition for teaching students, e.g. "Otherwise many students will fail" (P15, R1, S4). Some panel members judged the content standard to be less relevant because it is so obvious or because it goes together with other content standards, e.g. "This standard belongs to teaching students research skills" (P18, R1, S2). In the first round the correspondence with everyday practice was judged as moderate, because it assumes quite a lot of coordination in the school or because teachers are not so far, e.g. "I am convinced that most colleagues are not able to do this adequately. For sure, it should be done this way" (P2, R3, S2).

5. Teaching students research skills (TEACH). Most panel members thought this content standard to be relevant because it is the kernel of teaching: "If this is not what it is all about, what is it?"(P12, R1, S4). Because of lack of time the correspondence with everyday practice was judged to be in general somewhat less high.

6. Creating a positive pedagogical climate (CLIMATE). The panel members who judged this content standard as highly relevant thought that creating an adequate climate is an outstanding condition for good teaching and guidance, e.g. "It is a general condition for education" (P8, R1, S4; P9, R1, S4; P21, R1, S4). Some panel members remarked that this standard is a general one, and not specific to the promotion of reform 'independent research'. For this reason a few panel members judged the content standard as somewhat less relevant. Two panel members gave the warning that the pedagogical climate can become an umbrella construct, e.g. "This could become endless" (P22, R2, S-). Finally, several panel members indicated that the climate in class depends on the school climate and the teacher's personality: "This standard depends very much on the school climate and the teacher personality; teachers will differ in this to a great extent" (P2, R3, S3); "It is important to respect differences between teachers. Not everybody will be able to create an adequate climate in the same way." (P13, R1, S3).

7. Adequately assessing research skills of students (ASSESS). Several panellists thought "Assessing is an important component in the learning process of students" (P4, R1, S3; P6, R1, S4). When judging the correspondence with everyday practice, many added that teachers have problems with assessing students.

8. Reflecting on the program and on personal actions in teaching students research skills (REFLECT). Most panel members thought this content standard to be relevant but pointed out that teachers need training because they do not reflect as a matter of course. In addition, the school climate is important here. Consequently, the correspondence with everyday practice

is judged to be rather low. "Lack of time and facilities, it is a school's task" (P1, R2, S2). "Does not happen explicitly, demands consultation and training" (P15, R2, S2). In contrast to this, a couple of panellists thought this content standard so obvious that they doubt whether it should be made explicit.

9. Collaborating with colleagues (COLLAB). Although most panel members judged this content standard to be relevant, e.g. "This has to be accentuated, the teacher is a team player" (P1, R2, S3), a few thought the standard to be less relevant because it is obvious and overlaps with other standards: "Collaboration with colleagues is obvious. This is not an independent standard!" (P10, R2, S1). In addition, opinions differed as to the correspondence with practice: "This is in our school as yet very moderately developed, people are stuck in their own discipline" (P6, R2, S2) versus "Recognizable, but it stays dependent on the employment conditions in education and on the colleagues one has to collaborate with" (P21, R3, S4).

The panel members' comments show that their judgments are based on different experiences in their own (educational) practice. Important themes that arise are the feasibility and advisability of collaboration between teachers, the extent to which students can work independently, the fact that content standards are interwoven, and the need to train teachers with regard to some standards. Although the panel members in general judge all content standards to be important, they prefer some content standards above others.

In order to map the panel members' preferences we asked them in each round to rank the content standards. Based on the rankings in the third round we constructed a data matrix with 20 cases (20 responding panel members) and 36 variables (all pairs of 9 content standards [9(9-1)/2]). The rankings were converted into pairwise comparisons, scoring 1 if the first standard was preferred to the second, and 2 if otherwise. In searching for clusters in the answers of the panel members, a homogeneity analysis (Homals) was carried out. The result was a mapping of the panel members on two dimensions with eigenvalues of .302 and .178 (total fit .48).

The eigenvalues are moderate, which means that the panel members cannot be easily distinguished. They fall into four clusters, each with its own preferences. The clusters turned out not to be related to the background and positions of the panel members. Although it is sensible not to over interpret the results, *Figure 3.2* shows that the panel members with low scores on the first dimension (n = 5) prefer the content standards 1 (GOAL) and 9 (COLLAB) above all other standards. This dimension identifies panel members who perceive teachers as team players. Panel members with low scores on the second dimension (n = 8) prefer content standards related to preparing for students' independent work, such as 2 (ASSIGN) and 3 (MANAGE), over content standards that concern the teaching process itself, such as 5 (TEACH) and 6 (CLIMATE).

3.4.5 The panel's evaluation of the Delphi method

At the end of the study the panel members evaluated the Delphi method (questions on a 4-point scale from 1 'not or to a low degree' to 4 'to a high degree', Cronbach's alpha .73). The panel members did not need more Delphi rounds to formulate the content standards even more sharply (mean 1.2; standard deviation .70), and thought additional rounds would not add much to the results (3.6; .80). They thought the Delphi procedure yielded enough (2.7; .73), and were not much in agreement with the proposition that better methods than Delphi exist to develop content standards (2.2; 1.05). The panellists judged the questions in the questionnaires in the successive rounds to be clear (3.7; .56), and thought these to be adequate to

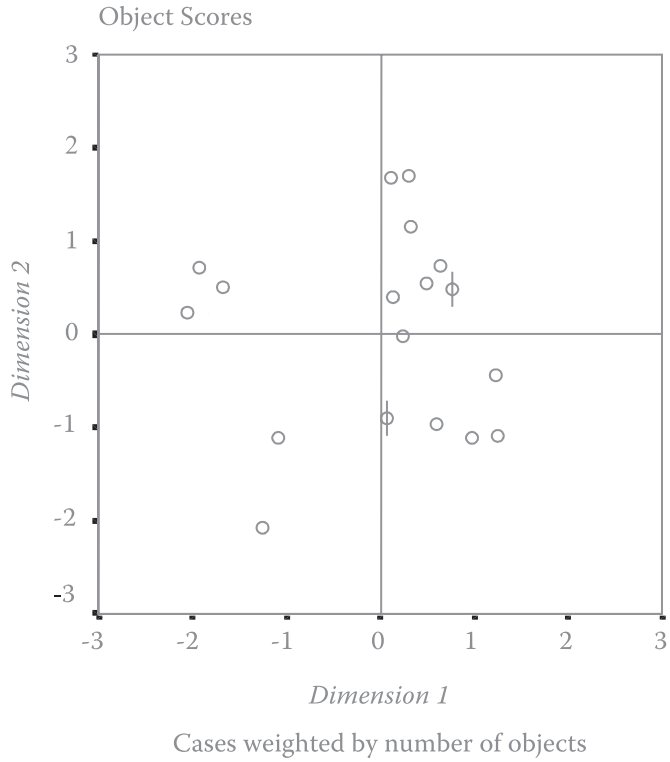


Figure 3.2 Panellists' preferences for the content standards (Homals) (n = 20)

communicate their comments on the content standards (3.1; .57). In giving their judgments in the second and third rounds, the panel members used to a reasonable degree the feedback they received from the previous rounds (2.5; .68). Finally, the Delphi method was evaluated to a reasonable degree to be a pleasant way to judge and revise content standards (2,6; .86), but at the same time judging the content standards had been experienced to be rather difficult (2.8; .87). The overall evaluation of the Delphi method (the scale mean) did not correlate significantly with the judgments of the content standards in the third round. This suggests that differences between the panel members in their evaluation of the method did not result in differences in their judgments of the proposed content standards.

3.5 Conclusions

An increased attention for teacher assessment and large-scale educational reforms ask for procedures to develop adequate content standards. The Delphi method is a promising alternative, because it has a task-orientated focus, reduces social influence processes, and has a cyclic character that results in successive refinement and generates convergence (Scheele, 1975). However, the feasibility of the method of generating content standards is still largely unknown. We tested the method in a small-scale standard setting study and summarize the

most important merits and restrictions.

Firstly, the method makes it possible to gradually improve content standards by an iterative process. The method resulted in nine content standards that are considered by the panel members to be highly relevant, thoroughly and clearly formulated and corresponding well with everyday teaching practice. The results show satisfactory interrater and interitem consistencies.

Secondly, although the ratings became more positive across the rounds, it is not fully clear what caused this change. It may be attributed to the iterative development of the content standards using the critical remarks of the panel members, to the feedback given, to their familiarization with the content standards (a learning effect), or even to the fact the panel members have simply confirmed the majority judgment (Greatorex & Dexter, 2000). This phenomenon needs further exploration, for example by longitudinal studies into possible changes in the content of panellists' cognitive representations of the content standards during the standard setting process.

Though in general statistical consensus was reached, it was difficult to reach a high level of consensus on the aspect 'Correspondence with everyday teaching practice'. The qualitative analysis of the additional remarks indicates that this could have resulted from the very different (school) practices of each panellist. Another possible explanation lies in the broad formulation of the content standards. In the questionnaires the content standards were illustrated with a few examples, but intentionally they were unrelated to a specific classroom setting. For that reason the panellists may vary in the teaching practice they connect the content standards to, and this may result in the relative less consensus. An alternative would be to illustrate the content standards using critical incidents or using more examples of practical classroom situations.

Fourthly, two important themes on which the panel members differed in their interpretations of the content standards were the feasibility and desirability of collaboration between teachers, and the degree of independence by which students can be expected to work on research assignments. It is no surprise that analysing the preference data with Homals, the panel members are divided into four clusters that differ in just these two dimensions. So, the method shows the existence of differences between panel members' preferences. However, the method does not clearly explain what causes these differences. For that reason the Delphi method needs to be supplemented. Interviews or other face-to-face methods in which the panel members' explain their underlying perspectives may be helpful as additional methods.

As expected, our study shows that the content standards can be interpreted in different ways. Judgments and preferences are likely to be explained by panel members' backgrounds such as specialities and professional skills (Plake, Melican & Mills, 1991). In our study however, the panel members' judgments and preferences are not related to their professional position nor to their discipline. A remaining explanation is that our panel members work in different (educational related) organizations. Since it is often suggested that individuals from similar backgrounds are likely to reach agreement most readily (Pula & Huot, 1993), the Delphi method may be more suitable for internal standard setting processes.

Sixthly, although we selected a broad group of stakeholders, the resulting content standards are not based on an exhausting overview of possible perceptions. So, as expected, the meanings of content standards strictly spoken can only be defined relative to the underlying perceptions of the stakeholders that participated in our study. Since the precise meaning of the content standards becomes important when using the standards (for example in assessment

situations), the persons concerned (e.g. assessors and assessees) at least should under scribe the meaning of the content standards as formulated.
Finally, the panel members judged a preliminary set of content standards already having a certain structure; they were not invited to propose a totally different set of content standards. Although the results cannot be separated from the input for the Delphi method (the questionnaire used), we tried to reduce these influences by a careful selection of content standards based on preparatory literature study and empirical research into teachers' practices, and by adjusting the questionnaire after each round only after analyses by two researchers. A pitfall is the enormous investment of time needed.
The next question is to what degree teachers have to meet the content standards for a satisfactory result. In order to be useful for teacher assessment, the content standards need to be accompanied by performance standards. Therefore, the content standards developed in this study have been used as input for a study into the development of performance standards (Van der Schaaf, Stokking & Verloop, 2003).
The need for suitable methods for developing content standards will remain for some time. In this study we wanted to make a contribution by studying the possibilities of the Delphi method as a content standard setting method. More small-scale evaluations are needed, to further improve the quality and feasibility of standard setting methods.

4 Developing performance standards for teacher assessment by policy capturing

Van der Schaaf, M.F., Stokking, K.M. & Verloop, N. (2003). Developing performance standards for teacher assessment by policy capturing. *Assessment and Evaluation in Higher Education, 28* (4), 395-410.

Abstract

There is a need for assessment of teachers' competences fostered by a growing attention given to accountability and quality improvement. Important questions are how well the demonstrated competences of teachers should be for a satisfying assessment and how the different competences should be weighted. Using a Policy capturing method, in 2 rounds, 9 stakeholders developed performance standards (or cut-off scores) for teacher assessment on eight content standards that resulted from an earlier study. Between the rounds, the panellists held a structured group discussion. Policy capturing proved to be a clear and useful method generating consistent judgments that can be described according to both a compensatory model and a conjunctive model. From the first to the second round, the consistency increased. However, whereas the panellists agreed to a substantial degree on the performance standards, they disagreed on the weights to be assigned to the content standards.

4.1 Introduction

There is a demand for accountability and quality improvement in the teaching profession. Teachers are expected to display the competences (knowledge, skills, and attitudes) needed to perform their tasks and to engage in continuous professional development. Assessment is used as a tool to ascertain whether teachers satisfy the required competences and to formulate guidelines for professional development. In the first case, assessment is used summatively in order to account for the teacher's quality, with possible consequences such as certification and merit pay. In the second case, assessment has a formative goal and produces information that can be used for planning activities directed at further professional development.

It is common to describe competences in terms of criteria (or content standards) and performance standards. The content standards describe the dimensions on which teachers should be assessed, and performance standards are the cut-off scores that indicate the minimum levels for a sufficient result. In this chapter, we focus on the development of performance standards for the purpose of summative assessment of experienced teachers.

Over the last decade, teacher competences have been increasingly assessed by competence-based assessment, accompanied by a change in instruments from behavioural observations

and knowledge tests to simulations and portfolios. The latter instruments recognize that teaching is a complex activity and that teacher behaviour is inextricably bound up with the teacher's cognition and with the situation in which the teaching takes place. In competence-based assessment, the judgment is preferably not dichotomous (in terms of right or wrong) but uses multiple cut-off scores (e.g. basic, competent, advanced). Additionally, it is usual to judge on several content standards, which are possibly differentially weighted.

In the last 10 years, different methods have been developed for generating performance standards to be used in competence-based assessment (Berk, 1996). Many methods are criterion-referenced, in which the candidate's score is directly compared with a performance standard. The performance is judged satisfactory if it surpasses this standard. Most of these criterion-referenced methods are based on stakeholders' judgments, in which stakeholders are asked to assess the minimum score necessary to satisfy the content standards.

Some methods using stakeholders are based on Policy capturing (e.g. Jaeger, 1995; Putnam, Pence & Jaeger, 1995). Policy capturing has been used in various fields to generate insight into the judges' cognitive processes (Cooksey, 1996). This method is particularly interesting as it is not only directed towards the question of how well teachers should score but also towards the weighting of the content standards and the specific judgmental model. To answer these questions, Policy capturing uses multiple regression procedures to produce a best-fitting equation or policy. Stakeholders judge a great number of teacher profiles, being configurations of scores on content standards. The judgments are analysed by multiple regression analysis, yielding a policy that indicates the performance standards and the content standards weights. This procedure can be used both for the whole panel of stakeholders (panel policy) and the individual panel members (personal policies).

Policy capturing studies are based on theoretical approaches to judgments and decision making as described in the Cognitive Continuum Theory (CCT) (Hammond, 1980). To integrate different theories concerning human judgments, during the 1980s Hammond developed the CCT building upon ideas from, amongst others, Brunswik (1952), Bruner (1957), and Newell and Simon (1972). The CCT focuses explicitly on the tension between intuitive and analytic thinking in human judgment processes. Brunswik (1952) conceptualized human judgment in a 'lens model'. The lens model parallels the idea of a distal cause in the environment that scatters its effects that humans re-combine when they see an object. In Policy capturing studies assessors also have to 'recollect' various scores (i.e., effects) on content standards (i.e., distal causes in the perception of the assessors). It is beyond the scope of this thesis to further explain the CCT.

The merits and shortcomings of using Policy capturing for developing performance standards are still largely unknown and the implementation meets some obstacles (Hambleton, 1997), of which we mention two. Firstly, the method is a difficult one for the participating stakeholders, even after training. This is problematic because the validity of the performance standards depends to a high degree on the stakeholders' competence to implement the method. Secondly, the method aims to reach consensus between the panellists. It is likely that panellists' judgments are based on different assumptions and perspectives, which probably result in different preferences for performance standards. To reach consensus all panellists should agree with the performance standards that best meet the different underlying assumptions and perspectives. Whereas the resulting performance standards will be based on a mix of different underlying visions, the precise meaning of the standards will become less clear.

In this study, we develop and test a Policy capturing method that aims to overcome these

problems. The study is part of a larger research project on the summative judgment of experienced social sciences teachers' competences (economics, geography, history) in pre-university education in the Netherlands (ages 16-18), comparable to undergraduate college level education in e.g. the American system. The teachers are expected to promote the development of their students research skills by having them work independently and in groups on assignments that result in a report or presentation.

4.2 Developing performance standards

4.2.1 Content standards

Competence is a generic term for the knowledge, skills, and attitudes required for adequate functioning in a profession. Competence can be divided into the (sub)competences needed to satisfactorily perform a set of tasks. Competences can be made visible in actions and be influenced by the context (Gonczi, 1994), and are seen as combinations of behaviour and the underlying intentions and motives. Following this definition in an earlier Delphi study, we developed content standards that became the input for this study (Van der Schaaf, Stokking & Verloop, 2005c). The content standards have been formulated both in terms of tasks and knowledge domains. The combination of both dimensions generates a matrix. One axis describes the tasks that teachers have to fulfil when teaching. The other axis describes the knowledge domains that teachers rely on when they fulfil their tasks. The cells refer to the content standards that teachers have to meet when teaching research skills to students (see Figure 3.1 in chapter 3 and section 5.2.3 in chapter 5).
The content standards are: 1. selecting goals for students (GOAL): teacher's long term goals for developing students' research skills and his or her approach to reach these goals; 2. choosing an appropriate assignment (ASSIGN): the goals, content and form of the assignment; 3. preparing and managing students to work on their assignments (MANAGE): the preparation, communication, and management of facilities (time, rooms, sources, media, etc.) that students need for fulfilling the assignment; 4. previously thinking of and selecting teaching strategies that support the development of research skills of students (THINK): the choice of and argumentation for the use of teaching strategies that meet students' knowledge, abilities and experience. 5. teaching students research skills (TEACH): the use of the teaching strategies chosen as described in content standard 4; 6. creating a positive pedagogical climate (CLIMATE): the provision of a safe, respectful, and stimulating learning environment to students; 7. adequately assessing research skills of students (ASSESS): teachers' argumentation for, and the clarity and comprehensibility of the assessment approach used (criteria, scoring, and norms), the applicability of the approach to the assignment, and the way the teacher handles the assessment and communicates the results to the students; 8. reflecting on the program and on actions in teaching students research skills (REFLECT): the extent to which the teacher is aware of strong and weak points in his or her teaching, and his or her suggestions for improvement. We decided not to include content standard 9 as formulated in chapter 3 (collaborate with colleagues), because this issue too highly depends on a teacher's workplace and is barely put into practice in many schools (Stokking & Van der Schaaf, 2000; Van der Schaaf, Veenhoven & Stokking, 2003).

4.2.2 Judgmental model

High-stake assessment, such as for certification, selection, or merit pay, must result in a single outcome. For that purpose, the scores on the content standards have to be combined into a final judgment. Two models that are frequently used for combining scores on content standards are the compensatory model and the conjunctive model. In the compensatory model, the candidate should achieve a defined total score to pass the assessment, allowing high scores on certain content standards to compensate for low scores on other content standards (Mehrens, 1990). The conjunctive model requires that the candidate attains a certain minimum score on each content standard.
Choosing between these models one should consider: (a) the perception of teaching as a profession. Is teaching a profession in which it is natural to have both strong and weak sides or should a teacher pass minimum requirements for each relevant content standard to guarantee a competent practice? (b) the supposed linkages between the content standards. If the content standards are strongly related, it is reasonable to let them compensate for one another; if not, a conjunctive model is a sensible choice. (c) the avoidance of incorrect judgments. With a compensatory model, positive and negative deviations cancel each other. With a conjunctive model, only one good or bad score, as a result of good luck or bad luck, could produce an incorrect overall judgment.
We prefer the compensatory model. It is realistic to assume that teachers have weak sides and that they still have scope to develop further. In addition, we focus on high-stake competence-based assessment, attempting to avoid incorrect judgments as much as possible.

4.2.3 Professional development

Generally, performance standards are based on an explicit model of the processes behind the competences or the way these competences develop. Often, development is placed on a continuum from less to more professional. Aiming at teaching performance standards, this approach is not wholly adequate, because development can be an irregular process with sudden improvements or unexpected environmental influences (Huberman, 1995; Day, 1999). In addition, the course of development can differ between individuals and between contexts (Eraut, 1994). Performance standards can also be based on minimally required or maximally feasible mastery levels (Wheeler & Haertel, 1993), as we do in our study.

4.2.4 Validation strategies for performance standards

Kane (1994) describes three types of strategies for demonstrating the appropriateness of performance standards which are developed judgmentally, based on stakeholders' assessments (cf. chapter 3). (a) Validation strategies based on external criteria consisting of comparisons of the performance standards with external information about the desired (levels of) teacher competences. As our study is directed towards a rather new element in the curriculum and external information is not yet available, we restrict ourselves to the other two types of evidence. (b) Validation strategies based on internal criteria, concerning the consistency of the judgments in the standard-setting procedure. The consistency indicates the reliability of the procedure used. We deliberately use the word "indicates", because variation in panellists' judgments does not necessarily mean that the procedure as such is not reliable, since the panellists could interpret the performance standards differently. (c) Procedural evidence, consisting of demonstrating the soundness of the chosen performance standard-setting procedure and its use, including the selection of the panellists involved in the process.

4.3 Method

4.3.1 Sample of stakeholders

A panel of nine stakeholders was selected. This number is assumed to be sufficient for setting reproducible standards (Berk, 1996). Seven stakeholders also participated in an earlier study in which content standards for teaching students research skills were developed (Van der Schaaf et al., 2005c). Hence, these panellists were familiar with the content standards. The additional two panellists were thoroughly briefed.
The panel consisted of 3 school principals (one of them formerly a history teacher and teacher trainer), 2 experienced geography teachers who were also teacher trainers, 2 experienced history teachers, one experienced teacher in both geography and history, and an experienced teacher in social science subjects who is also a teacher trainer in economics. The teacher trainers were working at teacher training institutes at or allied with several universities. All panellists were familiar with the objectives and approaches of teaching research skills to students aged 16-18. They supported the basic assumptions and goals of our study, and received a financial compensation. None of the panellists had ever participated in a Policy capturing study before.

4.3.2 Task

The panellists' task involved judging fictitious teacher profiles. Each profile consisted of a configuration of fictitious scores of one teacher on the eight content standards shown in *Figure 4.1*. That number of content standards is close to the maximum number that judges can handle in a Policy capturing procedure (Cooksey, 1996). Each content standard met one out of three anchor points, namely, the teacher satisfies the content standard: A, completely or to a strong degree; B, to a reasonable or to some degree; C, to a small degree or not at all. The profiles varied in their scores on the content standard. See *Figure 4.1* for an example.
In our study, we focused on portfolio assessment, a portfolio being taken as a collection of material about the teacher's work, in this case concerning a research assignment the teacher has given to the students. The material was made up of a self-description, the assignment, interviews, video-registrations of teaching research skills to students, judgments of students'

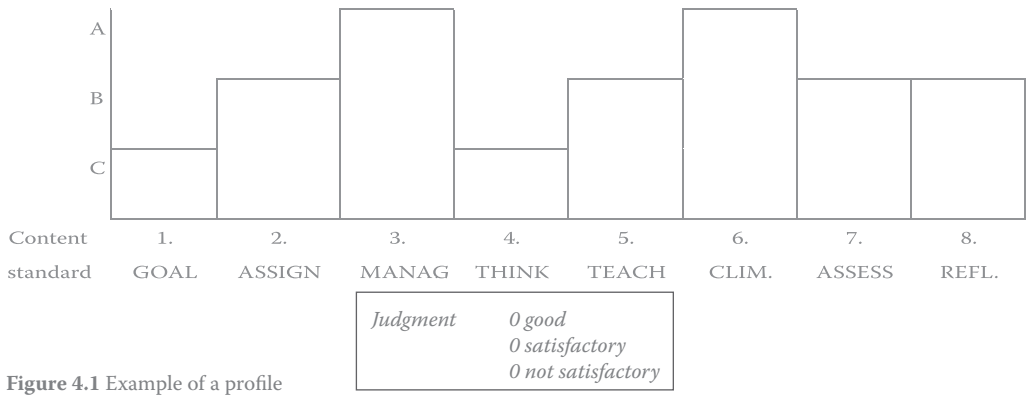


Figure 4.1 Example of a profile

NOTE. A = the teacher satisfies the content standard completely or to a strong degree; B = the teacher satisfies the content standard to a reasonable or to some degree; C = the teacher satisfies the content standard to a small degree or not at all.

work by the teacher, and evaluations by students. Before starting their task, the panellists studied an example portfolio. They also studied a manual in which all eight content standards, all anchor points for these eight content standards, and the portfolio material had been carefully and comprehensively described.
In a first round, the panellists judged 64 profiles. They judged 40 profiles in a second round. In the second round, 13 profiles from the first round were repeated to investigate whether the panellists' judgments were consistent (fitting the recommended number for replicated profiles) (Cooksey, 1996). The panellists were not informed of this. In order to minimize the effects of memory, we mixed the 13 replicated profiles with the 27 new profiles randomly.

4.3.3 Sample of profiles

Ideally, the sample of profiles represents the target domain and is drawn from empirical cases. However, in our study this was not feasible because teachers were still inexperienced in teaching students research skills. Therefore, we used a sample of fictitious profiles. With eight content standards and three anchor points 3^8 or 6561 profiles can be constructed. From these, using SPSS, we drew random samples of 64, respectively 40 profiles to be used in the first and second round (mean scores 2.02 and 1.96; standard deviations .29 and .30; minima 1.38 and 1.25; maxima 2.63 and 2.50). The ratios between the number of profiles and the number of content standards were 8:1, respectively 5:1, and match with the recommended minima (Cooksey, 1996).

4.3.4 Procedure

As Policy capturing is a difficult task and it is unclear how panellists having different perspectives could reach consensus, we used several means to support the process. Firstly, the panellists were given feedback on their judgments to increase their understanding of the judgmental process (Van de Pligt, 1997). Secondly, the panellists were given opportunities to reconsider their performance standards (Cross et al., 1984). Thirdly, the panellists discussed their judgments (Fitzpatrick, 1989). Fourthly, procedures were included to prevent negative effects of group discussion as much as possible. The procedure consisted of six steps.

1. Training (4 hours plenary)

The panellists studied a manual describing the objectives, planning, and procedures of Policy capturing. They then followed a training focused on the panel's aim, task, and procedure, including a discussion about the content standards, the anchor points, and the portfolio. The training goal was to let them practice judging the profiles, without standardizing their way of judging. The judged profiles were discussed, including the extent to which the panellists agreed and the reasons for their judgments.

2. Judgments, first round (3 hours individually)

The panellists received a booklet containing 64 profiles. It was explained that the profiles were fictitious. The panellists were asked to judge the profiles on the 3 anchor points. The panellists judged the profiles individually at home.

3. Studying feedback (1 hour individually)

The panellists received written feedback, consisting of a graph illustrating the panel policy and the personal policies, resulting from the multiple regression analysis.

4. Discussion (4 hours plenary)

The feedback (step 3) was the input for a meeting in which the panellists discussed the underlying arguments for the different policies and tried to reach consensus.
In group discussion, participants can be influenced by social comparison and by information from others. During the meeting, we tried to minimize the effects of social comparison (which would bring about unwanted effects, such as groupthink, polarization, and domination by a few panel members) and to promote the effects of information (which would increase the quality of the discussion). We used the nominal group technique (Delbecq, Van de Ven & Gustafson, 1975), consisting of a structured discussion in which the panellists alternate between working individually and in plenary sessions. During the plenary sessions they clarify their arguments and try to convince the other participants of their judgments.
Public commitment to their personal policies could make panellists resistant to changing their opinion (Fitzpatrick, 1989). To control for this effect, we divided the panel during step 4 (discussion) into two subgroups: five panellists who participated in the group discussion (panellists 1, 2, 6, 7, 8) and four panellists who had an individual discussion with a researcher (the first author) (panellists 3, 4, 5, 9). The group discussion and the discussions with the researcher, which both were video-recorded and fully transcribed, followed the phasing of the nominal group technique: (a) After recognizing their personal policies, the panellists considered how to substantiate them. (b) Each panellist presented and clarified his or her personal policy (to the other participants or to the researcher). (c) The panellists concluded their presentation in a few statements summarizing the main points. This resulted in a list of 26 statements. (d) The panellists that participated in the group discussion examined whether they could reach agreement on a panel policy. In addition, all panellists judged the 26 statements (the result of step c) on a 5-point scale (from greatly agree to greatly disagree).

5. Judgments, second round (2 hours individually)

As panellists reach better results if they have the opportunity to change their earlier policies (Jaeger, 1990), they were asked again to judge 40 fictitious profiles. They received the same instructions and were asked the same questions as in the first round.

6. Feedback

After the second round of judgments the panellists received feedback on their personal policy and the panel policy.

4.3.5 Data analysis

The panellists judged fictitious profiles of teachers, each profile consisting of scores on eight content standards. In the multiple regression models the profiles' scores on the content standards are the predictor variables, and the panellists' judgments of the profiles are the criterion variables. For multiple regression analysis to be valid, assumptions of normality of the residuals, homoscedasticity, linearity, and no collineation have to be fulfilled. We checked these by: making a histogram of the standardized residuals from the judgments of the profiles, making a plot of the standardized residuals against the predicted judgments, and calculating the tolerance of the content standards and the correlations between the content standards.
Our next questions were: (a) whether teachers' competences can be best judged according to a compensatory or a conjunctive model, (b) what weights the content standards should have

in the judgment, and (c) what the minimum overall profile score should be for teachers to be judged as "satisfactory". To answer these questions we executed multiple linear regression analyses according to a compensatory model, and multiple regression analyses applying a natural logarithm transformation according to a conjunctive model (Jaeger, 1995; Cooksey, 1996). To indicate the quality of the models, we used the proportion explained variance (R^2). The closer R^2 is to 1.0, the better the model describes the judgments.
To reveal possible differences between personal policies, we depicted them in a diagram, and computed the standardized Beta weights of the content standards in the personal policies.
In addition, we checked the consistency of the panellists' judgments by computing the correlations between the judgments of the 13 replicated profiles between the first and second round. Panellists possibly judged more consistently in the second round as a result of the group discussion. To control for this, we used the Mann-Whitney test (2-tailed) to test for differences in the consistency between the panellists who participated in the two subgroups.
Further, we analysed the panellists' scores on the 26 statements summarizing their main assumptions and perspectives. We searched for a relation between the panellists' perspectives and their personal judgmental policies by computing the partial correlations between the scores on the statements and the standardized Betas of the personal policies.
Finally, the panellists evaluated the Policy capturing method on the following aspects: the difficulty of the procedure, the recognizability of their personal policy, and the role achieved by the feedback.

4.4 Results

4.4.1 Assumption checks

The assumptions of normality, homoscedasticity, and linearity were sufficiently met. The content standards (the predictors in the model) were barely collineated: the tolerance was high (between .80 and .95 with one outlier of .75), and the Pearson correlations were low (between r = -.34 and r = .20 and mostly around r = .00).

4.4.2 Judgmental model, content standards weights, and performance standards

Firstly, the panel policy and the personal policies for both rounds were analysed. *Table 4.1* shows that the conjunctive model predicts panellists' judgments slightly better than the compensatory model. However, in our view the difference between both models is not large enough to relinquish the advantages of the compensatory model. The higher R^2 in the second round indicates that the model better represents the judgments in the second round than in the first round.
Secondly, we wanted to know what weights the content standards should be given. *Table 4.2* contains the results of the analyses, using a compensatory model. The Betas show that the panel gives content standard 5 (TEACH) and 6 (CLIMATE) more weight than the other content standards.
Thirdly, the panel policy was calculated in order to know what the minimum total score should be for teachers to be judged as "satisfactory". Every profile consisted of eight content standards with (variable) scores on anchor points A, B or C. The profiles were judged by the panellists as "good", "satisfactory" or "not satisfactory". To calculate the panel policy we attached the scores 3, 2 and 1 to the anchor points A, B and C, and also 3, 2 and 1 to the judgments "good",

Table 4.1 Adequacy of the panel policy and personal policies in terms of R^2

	Round 1		Round 2	
Panellist	Compensatory R^2	Conjunctive R^2	Compensatory R^2	Conjunctive R^2
1	0.70	0.71	0.62	0.67
2	0.43	0.45	0.60	0.71
3	0.40	0.44	0.74	0.77
4	0.51	0.53	0.74	0.79
5	0.45	0.55	0.78	0.80
6	0.62	0.73	0.55	0.68
7	0.71	0.73	0.70	0.76
8	0.80	0.74	0.88	0.89
9	0.55	0.59	0.80	0.81
Panel (total)	0.37	0.40	0.54	0.58

Table 4.2 Results of multiple regression analyses, rounds 1 & 2 compensatory model

	Round 1 (n = 575)			Round 2 (n = 360)		
	b	Standardized Beta	T (p < .001)	b	Standardized Beta	T (p < .001)
1 GOAL	0.13	0.16	4.40	0.16	0.19	4.68
2 ASSIGN	0.15	0.19	5.11	0.23	0.27	6.98
3 MANAGE	0.15	0.18	5.19	0.06	0.07	1.83[a]
4 THINK	0.12	0.17	4.45	0.15	0.17	4.10
5 TEACH	0.25	0.33	9.55	0.33	0.38	9.78
6 CLIMATE	0.21	0.27	7.66	0.33	0.35	8.66
7 ASSESS	0.15	0.19	5.34	0.19	0.21	5.10
8 REFLECT	0.10	0.13	3.72	0.14	0.15	3.85
Constant	-0.73		-4.63	-1.35		-7.64
	(se 0.16)			(se 0.18)		
Round 1 R = 0.61 F (df) = 41.49 (566) (p < .001) R^2 = 0.37 Adjusted R^2 = 0.36 Round 2 R = 0.73 F (df) = 50.98 (351) (p < .001) R^2 = 0.54 Adjusted R^2 = 0.53						

[a]p = .69

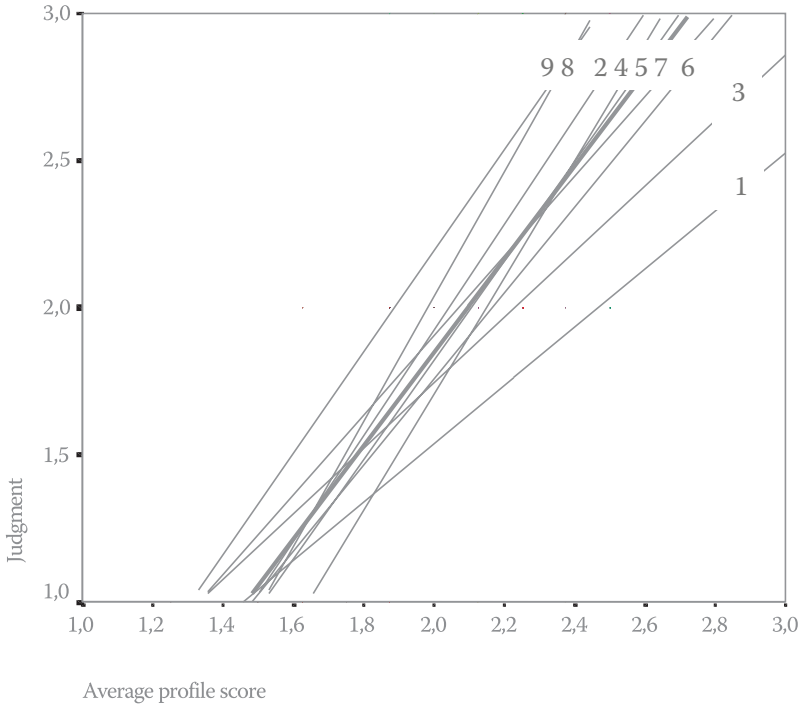
"satisfactory" and "not satisfactory". So, both the average profile score (the average of the fictive scores on the eight content standards) and the judgments can meet scores between 1 and 3. In *Figure 4.2*, the x-axis represents the average profile scores, and the y-axis refers to the judgments given. The bold line indicates the panel policy (round 2, compensatory model). The thin lines indicate the personal policies. *Figure 4.2* shows that some of the panellists judge more leniently than others. However, most of the thin lines are close together, so the panellists as a group agreed to a great extent on the performance standards. The results in the first round (not shown here) were very comparable. In both rounds, teachers needed minimally an average profile score between 2.1 and 2.2 to be judged as "satisfactory" (that is, to get a mean panel judgment of 2).

In our study "satisfactory" means that a teacher's portfolio is sufficient, and meets the panel's perception of a competent teacher to a moderate to high degree. A "satisfactory" total score can have several consequences, like certification or merit pay in the case of summative assessment, or insight in strong and weak points that need further development in the case of formative assessment.
Although most panellists were close together on the cut-off point, their opinions differed on the weights to be attached to the content standards. *Table 4.3* gives a survey per round of the weights the panellists attached to the content standards, based on the compensatory model. Most of the panellists attached comparable weights to the content standards in the second round as they did in the first round. The panellists' answers were not related to their professional position (e.g. teacher, principal, teacher educator), or to their discipline (economics,

Table 4.3 Significant (p≤ .05) standardized Betas per content standard per panellist

Content standard Round	Panel-list	1. GOAL	2. ASSIGN	3. MANAG	4. THINK	5. TEACH	6. CLIM	7. ASSESS	8. REFL
1	1	0.43	0.21	[a]	0.24	0.39	0.36	0.24	[a]
2		0.27	[a]	[a]	[a]	0.44	0.41	[a]	[a]
1	2	[a]	0.22	0.22	[a]	0.36	0.37	[a]	0.23
2		[a]	0.21	[a]	[a]	0.39	0.52	[a]	0.33
1	3	0.24	[a]	[a]	0.26	[a]	0.33	0.38	[a]
2		[a]	0.40	[a]	[a]	[a]	0.44	0.43	[a]
1	4	0.28	[a]	0.30	[a]	0.30	0.30	[a]	0.30
2		0.34	[a]	[a]	0.35	0.31	0.36	[a]	0.42
1	5	0.20	0.21	0.23	[a]	0.37	0.26	[a]	0.28
2		0.20	0.32	[a]	0.20	0.50	0.37	0.31	[a]
1	6	[a]	0.67	[a]	[a]	[a]	[a]	0.41	[a]
2		[a]	0.51	[a]	[a]	[a]	[a]	0.30	[a]
1	7	0.25	[a]	0.19	0.22	0.51	0.54	[a]	[a]
2		0.22	[a]	[a]	[a]	0.53	0.51	[a]	[a]
1	8	0.19	0.27	0.34	0.25	0.38	0.30	0.36	0.19
2		0.31	0.42	0.22	0.25	0.23	0.31	0.39	[a]
1	9	0.30	[a]	0.18	0.22	0.61	[a]	0.22	[a]
2		[a]	0.24	0.22	0.37	0.75	[a]	0.19	[a]

NOTE. [a] not significant (p > .05)



NOTE. The numbers near the policy lines refer to the panellists. The bold line refers to the panel policy. Judgment (y-axis): 1 = not satisfactory; 2 = satisfactory; 3 = good. Algorithm of the panel policy: 0.16 * GOAL + 0.23 * ASSIGN + 0.06 * MANAGE + 0.15 * THINK + 0.33 * TEACH + 0.33 * CLIMATE + 0.19 * ASSESS + 0.14 * REFLECT - 1.35.

Figure 4.2 Judgments of mean profile scores by panellists in round 2, compensatory model (R^2 = 0.54)

geography, history), or to other characteristics. Most of the correlations (Pearson) between the judgments of the 13 replicated profiles in the first and second round were $r > .75$ ($p < .05$) (with two outliers of r = .55 and r = .63), indicating consistency of the panellists' judgments. According to the Mann-Whitney test (2-tailed), there was no significant difference between the panellists who participated in the group discussion (n = 5) and those who had an individual conversation with the researcher (n = 4).

4.4.3 The panel's underlying assumptions and perspectives

Each panellist summarized the underlying assumptions and perspectives of his personal policy in a few statements. The resulting 26 statements were scored by the panellists on a 5-point scale (from 1 = I greatly agree to 5 = I greatly disagree). Exploratory factor analysis followed by item analyses resulted in 3 sufficiently reliable Likert-scales, each consisting of 5 statements. We illustrate each scale with an example of a statement: (a) The extent to which the panellists connect the content standards to each other, e.g. "All content standards are connected, without the others they could not exist" (Cronbach's alpha .85; mean 2.42; minimum 1.75; maximum 3.44; variance .54); (b) The extent to which the panellists want to reduce the judgmental model to one core, e.g. "To sharpen the judgmental model the number of content standards should be reduced" (Cronbach's alpha .84; mean 1.95; minimum 1.67; maximum

2.44; variance .13); (c) The extent to which the panellists are directed at the learning environment of students, e.g. "The choice of a research assignment is important because a good assignment is a condition for students' research skills to develop" (Cronbach's alpha .74; mean 2.22; minimum 1.44; maximum 3.33; variance .47).
The panellists' answers were not related to their professional position or to their discipline, nor to their personal judgmental policies. In some cases panellists had different reasons for the same judgments, while in other cases panellists gave different judgments for the same reasons. This observation was confirmed by the correlations between the scores on the scales with statements and the standardized Betas expressing the personal policies in the second round. Only 2 out of 24 correlations (3 scales * eight content standards) were significant.

4.4.4 The panel's evaluation of the Policy capturing procedure

The panellists thought judging the profiles to be not very difficult or easy. On a 5-point scale (1 = easy; 5 = difficult) the mean score was 2.9 in the first round and 2.6 in the second round. Profiles with unequivocal positive or negative scores on important content standards were easy to judge, whereas it was difficult to judge profiles with both positive and negative scores on important content standards. The panellists judged the latter in some cases to be not realistic or to be rarely found in practice. Secondly, almost all panellists recognized themselves in their personal policy. Thirdly, for all panellists the feedback given to them confirmed what they had previously thought to be important in judging teachers.

4.5 Conclusions

In this study, the focus was on the development of performance standards for high-stake teacher assessment purposes. During the 1990s several performance standard-setting methods were developed that can be used for performance assessment. Those methods based on Policy capturing are particularly interesting because they can be used both for the development of performance standards and to gain insight into the judgmental process (compensatory or conjunctive) as well as into the weights to be attached to the content standards (Jaeger, 1995). The disadvantages of this method are its difficulty for the judges and the fact that it is not clear how to foster consensus (which is seen as an indicator of success). In this study we attempted to counter these difficulties by: a careful selection of panellists representing different opinions and having sufficient knowledge and motivation for their task; construction of the judgmental task from eight content standards about which panellists in an earlier study had reached consensus; a training process; interim feedback; a group discussion using the Nominal group technique; and judgment of profiles in 2 rounds.
To indicate its usefulness, we now summarize the main merits and shortcomings of the Policy capturing procedure, as utilized in our study.
Firstly, the procedure produced significant and consistent regression models. The panellists' judgments in the second round were more consistent than in the first round. It is not precisely clear which intervention or combination of interventions contributed to this growth in consistency.
Secondly, we investigated whether we should follow a compensatory or conjunctive model for combining sub scores into a final judgment. Whereas it is realistic to assume teachers have weak points that they still have to develop further, and it is important to avoid incorrect judgments caused by errors, we initially preferred a compensatory model. The results show

that both models represent the panellists' judgments nearly equally well. We conclude that there is no reason to abandon our preference for a compensatory model.
Thirdly, the panellists agreed rather strongly on the performance standards accompanying the eight content standards. In both rounds, the average profile score needed for the judgment "satisfactory" turned out to lie between 2.1 and 2.2. Hence, according to the algorithm of the compensatory model, the cut-off profile score is 2.2. This outcome can be used for portfolio's that are assessed per content standard on a three-point scale (between 1 to 3).
Fourthly, the panellists did not agree on the weights to be attached to the content standards in the final judgment. This result differs from that of another study on developing performance standards (Hambleton & Plake, 1995) in which a judgment procedure in 2 rounds with interim discussion resulted in high agreement on the panel policy. This discrepancy might be explained in several ways. The Nominal group technique, in which panellists publicly present their personal policies, could discourage individuals from changing their initial response (Fitzpatrick, 1989). However, the controls conducted did not show any evidence for this. Next, it is possible that the feedback given has been perceived after all as supporting the initial personal ideas. Additionally, our study might have exerted less pressure to revise the initial policy because the panellists knew they were participating in a research project and not in an authentic summative assessment context. Also, the time span per round possibly was too short for the panellists to revise their opinions. Policy capturing could be more effective if it is undertaken by panellists working in the same context, sharing the same frame of reference, and over a longer time span.
Fifthly, our stakeholders perceived the eight content standards as being interwoven and would welcome a reduction in their number. This is not an unknown phenomenon. For example, the academic literature on assessment centres reveals that during the assessment, judges merge content standards, prefer certain content standards, or have difficulty distinguishing content standards (Arthur, Woehr & Maldegen, 2000). The interventions in our study to familiarize the panellists with the content standards, such as the training and the manual, were apparently not sufficient to prevent this.
Sixthly, in our study Policy capturing did not prove to be a difficult procedure for the participants. Profiles with consistently high or low scores on important content standards were relatively easy to judge, whereas profiles with "inconsistent" scores were perceived to be more difficult, especially if the panellists thought the profiles were unrealistic. Empirical profiles might prevent this problem.
We conclude that the Policy capturing method, as used in our study, is a useful method for stakeholders to develop performance standards. The method is feasible for the panellists and leads to consistent performance standards. The procedure followed brings about a learning effect, resulting in more consistent judgments.
This research should be seen in the context of a growing interest in teacher assessment and the need for suitable methods to develop performance standards. We wanted to contribute by developing and testing a Policy capturing method. The need for new methods and for research on this topic will remain for some time. It is important to fulfil this need for appropriate teacher assessment practices to be realized.

5 Designing teacher portfolios using a chain model of construct validation

This chapter is submitted as: Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (2005). Designing teacher portfolios using a chain model of construct validation. Manuscript submitted for publication

Abstract

A common implicit assumption in teacher portfolio assessment is that the scores given to a portfolio are based on the constructs measured. Main components in teacher portfolio assessment are the construct domain to be assessed (i.e., teacher competences), the content standards derived from this domain, the teacher portfolio design, and raters' scoring of the portfolios. In this study we focus on two links between these components, namely the link between the content standards and the portfolio design and the link between the content standards and the raters' scoring. In a panel study eight experts judged an initial portfolio design based on the Theory of Planned Behavior, using a Policy capturing method. This resulted in a high degree of support and consensus for the initial design. Subsequently, 18 teachers developed a portfolio based on this design. Next, 6 trained raters assessed the teacher portfolios according to eight content standards. The results show that both the panellists and the teachers prefer artefacts and reproductions as forms of portfolio elements above productions. The interrater reliability of 12 portfolios was satisfactory. The raters based their assessments to a substantial degree on the content standards. However, some indicators of the content standards were less referred to in their judgments. Implications of the results and suggestions for follow-up research are discussed.

5.1 Introduction

Teachers' knowledge, skills, and attitudes are often integrated into competences and are increasingly assessed by competence-based assessment instruments, such as simulations and portfolios. Depending on its content and form, portfolios may support broad and context-bound assessments. Such portfolios can do justice to the holistic nature of competences and to the personal character of teachers' work over a larger period of time (Bird, 1990) and will include authentic products and performances which are often seen as a prerequisite for a valid assessment. Also, such assessments are generally of direct interest for the teachers themselves and their teaching practice. However, due to their open character it becomes harder to anticipate the range of teachers' possible responses, which is needed to develop a relevant portfolio design and to assess the quality of teacher performance in a satisfying

way. Portfolio contents are often descriptive, context-bound and personal and ask for much interpretation before they can be assessed (Moss, 1994). Also the meaning of the content and performance standards on which portfolios are assessed can only be defined relative to the underlying perceptions of the assessors. This affects the reliability and validity of portfolio assessments. Ways in which the quality of portfolio assessment can be increased are not yet fully clear.

Although portfolio assessments vary in purposes (e.g. summative or formative) and in form (e.g. in paper or digital), a common implicit assumption is that the scores given to a portfolio are caused by the constructs assessed (i.e., teacher competences) (cf. Borsboom, Mellenbergh, Van Heerden, 2004). This assumption is critical and deserves attention, because assessment is primarily concerned with the intended competences (Linn, 1994; Linn, Baker, & Dunbar, 1991; Borsboom et al., 2004). This assumption is the foundation upon which the assessment and the resulting consequences are based (e.g. selection, merit pay, feedback and suggestions for further professional development). Therefore, a portfolio assessment should match the construct domain to be assessed and the content standards derived from this domain. Main components in the assessment process of teacher portfolios are: the construct domain to be assessed (i.e., teacher competences), the content standards derived from this domain, the teacher portfolio design, and the raters' scoring of the portfolio.

This starting point at least leads to four essential concerns in validating the scoring of teacher portfolios: (1) the domain to be assessed must be demarcated, and evidence must be provided that content standards and performance standards match this domain; (2) evidence must be given about the congruence between the portfolio design and the content and performance standards; (3) evidence must be provided about the link between raters' portfolio scoring and the content and performance standards (i.e., the standards should be applied as attended); (4) the congruence between teachers' competences shown in their portfolio and the competences assessed by the raters should be demonstrated.

Here we focus on the second and third concerns mentioned: the links between the content standards and the portfolio design and between the content standards and raters' scoring. The other two issues are addressed elsewhere (Van der Schaaf, Stokking & Verloop, 2003; 2005c; 2005e) (see chapters 3, 4, 7). The research questions are: (1) How is the initial portfolio design related to the content standards? (2) How is raters' portfolio scoring related to the content standards?

Our study is part of a larger research project on the summative and formative assessment of experienced teachers' competences in teaching research skills in upper secondary social science education. Teaching research skills is a recent Dutch secondary education exam requirement that is representative for the current change in many countries towards more constructivist views on learning and that emphasize developing students' skills in an active, self-directed and collaborative way.

5.2 Construct validation of teacher portfolio assessment

5.2.1 Quality requirements for teacher portfolio assessment

Teacher portfolio assessment has to satisfy quality requirements concerning reliability and validity. Reliability refers to stability over time and consistency between assessments. Validity refers to the question whether the assessment is measuring the intended constructs, i.c.

teacher beliefs and behaviour. As mentioned in chapter 2, based in part on Messick (1989; 1995) we conceive validation criteria covering a number of facets (see also Stokking, Van der Schaaf, Jaspers & Erkens, 2004):

a) content: sufficient coverage of the construct or domain one wants to measure;
b) process: the assessment should be based on a model of the task as much as possible and show the development of the relevant competences;
c) scoring: the scoring model and scoring rules should mirror the structure of the construct, domain or task;
d) specificity and unambiguous interpretation: the assessment should distinguish between more and less competent teachers, and the scores should not be explainable in a causal sense by other factors than the competences intended;
e) convergence (a special case of d): measurements of the same construct, domain or task using different methods should correspond as much as possible;
f) generalizability: the degree to which the results are valid for a broader range of tasks, conditions and populations.

The term validity has been given broader meaning (Messick, 1989, 1994, 1995; Crooks & Kane, 1996) subsuming all quality criteria, not only those relating to measurement, evidence and interpretation, but also those concerning the assessments' use and consequences. For reasons of clarity we prefer however to preserve the traditional, more restricted meaning of validity (cf. Borsboom et al., 2004).

5.2.2 A chain model of construct validation

The facets mentioned above which are most important in the construction of a valid assessment instrument are the first three facets and these are therefore central in our study. An assessment process includes four salient components that are into play in covering these facets (Mislevy, Steinberg & Almond, 1999) (sequence altered, see also Figure 2.1): (i) a psychological construct domain that describes teacher competences in teaching students research skills (teacher model); (ii) teacher competences to be displayed in the portfolio (task model); (iii) raters' scoring of these displayed teacher competences (scoring model); (iv) the content standards and performance standards by which the competences are finally assessed (assessment model). Content and performance standards (iv) are constructed from the psychological construct domain (i). Content standards describe the dimensions on which teachers should be assessed and performance standards are the cut-off scores that indicate the minimum levels for a sufficient result.
For a portfolio assessment to be valid these different components have to be consistent with each other (Kane, 1992; Crooks & Kane, 1996). This means that portfolio assessment is not just a matter of adequate separate components but that all components should be linked. Subsequently, theoretical rationales and empirical evidence should support the adequacy of these links (Cronbach, 1988, Crooks & Kane, 1996; Embretson, 1993; Kane, 1992; Messick, 1989, 1994). We view the construct validation of portfolio assessment as a process of reasoning and of weighing evidence for the presence of the links between the components.
Certain characteristics of constructs and the context and aim of the assessment may impact the kind of evidence needed to support the adequacy of the links. For example, the construct to be assessed can be characterized in terms of the combination of several dimensions,

such as concerning the level of detail and the type of construct. The level of detail in which the construct is defined (conceptual) and measured or observed (operational) ranges from extremely atomistic to highly holistic. The type of construct can range from verifiable performances that retain a constant meaning across diverse contexts to highly context-bound personal constructs that underlie these performances (Harvey & Wilson, 2000). Generally, teaching can be described as a holistic construct but is also characterized by contextuality and heterogeneity, making statistical verification and comparability difficult to achieve.
Absolute evidence demonstrating the validity of teacher portfolio assessment cannot be found. The evidence to validate an assessment is never definitive but rather more or less probable and amenable to correction by new evidence (Cronbach, 1988). Therefore, in our view, the question whether the scores given to a portfolio are caused by the constructs assessed cannot be answered in any absolute sense. The answer will always depend on the question: when, under what conditions, and for what kind of people etc., are the constructs and the portfolio assessment scores related?
There is not a one-to-one relationship between inferences and types of evidence, but some types of evidence may be especially useful to support particular links in the teacher portfolio assessment process. For example the link between the task model and the assessment model (concerning the content facet) implies that an assessment should cover the knowledge content of teaching. Evidence for this link often relies on expert judgments about the degree to which the construct domain and the content standards derived from this domain (i.e., what teachers should know and be able to teach) are reflected in the task model (the teacher portfolio) (Crocker, 1997). Evidence for the link between the assessment of competences and the execution of a task (the process facet) may be provided by theories and models about the performance concerned (in this case: constructing a teacher portfolio). Also empirical evidence that includes an analysis of the process of executing the task, for example gathered by using think aloud methods, may be useful. Evidence for the congruence between raters' scoring and the content standards (the scoring facet) is about the structure of the scoring task and the adequate use of the content and performance standards. Useful evidence concerning this issue may for example derive from think aloud studies and interviews.
In two sub studies we focus on the links between the teacher portfolio (ii, task model) and the content and performance standards (iv, assessment model), and between raters' scoring (iii, scoring model) and the content and performance standards (iv, assessment model). The first link (ii-iv) addresses the question whether the teacher portfolio matches the content standards. The second link (iii-iv) is about the question whether raters' judgments are based on the content standards as taught in the rater training (Borman, 1977), decreasing the likelihood that judgments are based on raters' subjective personal representations.
Now we first discuss the characteristics of the content standards that are the input for the two sub studies.

5.2.3 Content standards and performance standards

In earlier studies (Van der Schaaf et al., 2003, 2005c, see chapters 3 and 4) we successfully developed content standards and performance standards and these became the input for this study. In one previous study we researched into teacher competences needed to develop students' research skills. We developed content standards, most of them made up of indicators, that describe what teachers should know and be able to do in teaching students research skills according to three teaching phases (pre-active, interactive, post-active). We

used a Delphi technique (Linstone & Turoff, 1975) (see chapter 3) and a Policy capturing method (Jaeger, 1995) (see chapter 4) to help achieve conceptual consensus in a group of 21 stakeholders (including the panellists and raters in this study). These content standards and indicators were:

Pre-active

1. Selecting goals for students (GOAL): a) teacher's long-term goals for developing students' research skills; b) the approach used by the teacher to reach these goals.
2. Choosing an appropriate assignment (ASSIGN): a) the goals are related to subtasks of doing research; b) the content is related to doing research in the subject at hand; c) the form of the assignment is authentic, offers students options, and is approachable in various ways; d) the assignment closely fits the students' level.

Interactive

3. Preparing and managing students to work on their assignments (MANAGE): a) the preparation and the communication; b) management of facilities (e.g. time, rooms, sources, media) that students need for working on the assignment.
4. Previously thinking of and selecting teaching strategies that support the development of students' research skills (THINK): a) the choice of and argumentation for teaching strategies that meet students' knowledge, abilities and experience. The teaching strategies fit: b) teachers' goals; c) the assignment; d) the assessment.
5. Teaching students research skills (TEACH): a) correct use of the teaching strategies chosen, The instruction and guidance fit: b) the concerning subtasks of doing research; c) the operational stage in the research assignment; d) students' level.
6. Creating a positive pedagogical climate (CLIMATE): a) the provision of a safe, respectful, and stimulating learning environment for students.

Post-active

7. Adequately assessing the research skills of students (ASSESS): a) teacher's argumentation for, and the clarity and comprehensibility of the assessment approach used (content standards, scoring, and norms); b) the applicability of the approach to the assignment; c) the way the teacher handles the assessment and d) communicates the results to the students.
8. Reflecting on the program and on actions in teaching students research skills (REFLECT): the extent to which the teacher is aware of strong and weak points in a) his or her teaching and b) the educational program. His or her suggestions for improvement of c) his or her teaching and d) the educational program.

In another study (Van der Schaaf et al., 2003), we used Policy capturing to develop performance standards for the above-mentioned content standards (i.e., the cut-off scores that indicate to what extent teachers should meet the content standards for a sufficient result). We also developed, using multiple regression analysis, an algorithm for estimating teachers' total portfolio scores (the algorithm calculates a weighted mean analytic score). The study resulted in three anchor points on the 5-point scales used in rating teachers' competences on the basis of their portfolios, namely that the teacher satisfies the content standards: completely (score

5), to a reasonable degree (score 3), or not at all (score 1) (see Appendix 1).

5.3 Developing a portfolio design in congruence with the content standards

5.3.1 Overview

In this section we study the link between a portfolio design and the content standards during the initial portfolio design development. Developing a portfolio design is a cyclic process including one or more phases of designing, developing, evaluating, and revising. We started with constructing an initial design based on the content standards and on the Theory of Planned Behavior (Ajzen, 1985; Ajzen & Fishbein, 2005). We used this theory to describe the relation between teacher behaviour and cognitions as described in their portfolios. Then, eight panellists judged and weighted the elements in the initial portfolio design in a panel study and they offered suggestions for improvement of the design. Then, we revised the initial design based on the judgments and suggestions of the panellists. Next, three teachers constructed a portfolio using the revised design in a pilot, intended as a try-out of the portfolio design and also used as a try-out for the rating procedure in the second sub study. After studying the three pilot portfolios, the panellists judged the content relevance and representativeness of the initial portfolio design in a questionnaire and they indicated again the weight per portfolio element.
Finally, 18 teachers constructed a portfolio based on the developed portfolio design. A prerequisite for a reliable and valid portfolio assessment is that teachers do not seriously experience sources of bias during the preparation of a portfolio because otherwise the portfolio may give a false picture of their cognitions and behaviour and the portfolio probably cannot be linked to the content standards. Consequently after the portfolio construction, the teachers evaluated the content relevance and representativeness of the portfolio design in a questionnaire.

5.3.2 Developing an initial portfolio design

5.3.2.1 Theory of Planned Behavior

The content standards, described above, were the input for the portfolio design in this study. Since the content standards include both teacher behaviour and cognitions in different phases of teaching and these should be aligned with the portfolio design, it is important to include evidence of underlying teacher cognitions in the portfolio design. These underpinning teacher cognitions are context-related and exist partly as tacit knowledge that cannot be easily articulated. As a result, such evidence cannot be directly derived from the teacher's performance (Eraut, 1994). As implicit clusters of cognitions can only be studied once they are first made explicit, we assume that teacher cognitions reported in their portfolios neither present pure reports nor representative samples of their cognitions per se. Rather, teachers construct representations of their cognitions.
To develop a portfolio design that includes indicators for teachers' underlying cognitions, as explicated in their portfolios, we used the Theory of Planned Behavior (TPB) (Ajzen, 1985; Ajzen & Fishbein, 2005). TPB has proved its practicability in research into the relation between explicated teacher cognitions and behaviour in (social) sciences and in educational change

(e.g. Crawley, 1990; Haney, Czerniak & Lumpe, 1996). The theory is based on three clusters of cognitions that cause variance in people's goals and intentions to engage in certain behaviour (Bandura, 1998; Fishbein, Triandis, Kanfer, Becker, Middlestadt & Eichler, 2001) (see *Figure 5.1*). These clusters are: 1) the expected effects of particular behaviour (behavioural beliefs); 2) the expected support by significant others, e.g. colleagues (subjective norms); 3) the expected factors that may ease or complicate the performance of behaviour, e.g. perceived facilitation for teaching students research skills (control beliefs). On the whole, the evaluation of the clusters is assumed to produce overall positive or negative goals and intentions to perform particular behaviour. For example, teachers who think that they have the skills and support needed to teach students research skills may develop a strong sense of self-efficacy, whereas teachers who assume that they lack some of the resources and support needed are likely to have much weaker control beliefs and as a result may have weaker intentions in teaching

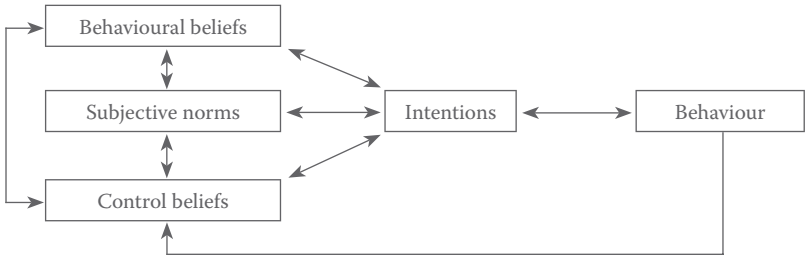


Figure 5.1 Relations between beliefs, intentions and behaviour (Ajzen & Fishbein, 2005)

students research skills.

5.3.2.2 Initial portfolio design

Portfolios normally contain selected evidence of performance and products in various contexts accompanied by a teacher's comments and reflections (Wolf & Dietz, 1998). The evidence included in a portfolio may be more or less structured. A portfolio can comprise several types of evidence, such as artefacts (regular products of a teacher's work); reproductions (registrations of the teacher's regular work, captured on behalf of the portfolio); attestations (records of the teacher's work made by others); and productions (documents especially prepared for the portfolio) (Collins, 1991).

Since the teachers who participated in our study did not have any experience in constructing a portfolio, we structured the content of the elements while teachers were free in the type of the material (verbal or written) and also free to add other material (Shapley & Bush, 1999). The initial portfolio design was about teaching students research skills in several phases of the teaching process (pre-active, interactive, post-active). The design contained seven elements (see *Figure 5.2*), being a number of productions, one reproduction, and several artefacts. The content of the elements focused on aspects of the Theory of Planned Behavior (see *Figure 5.1*): background factors (e.g. years of teaching experience, subject, class, and courses followed in teaching students research skills), teachers' beliefs in teaching students research skills (e.g. their goals for engaging students in research), subjective norms (e.g. arrangements with colleagues), control beliefs (e.g. perceived facilities for student research, perceived difficulties

Phase and Content standard	Element	Form	Content
Pre-active	1. Self description	production	Teacher's background, context, behavioural beliefs
GOAL	2. Series of assignments and clarification	artefact and production	Teaching behaviour, control beliefs, subjective norms
ASSIGN	3. Research assignment and clarification	artefact and production	Teaching behaviour, teaching intentions, control beliefs
Interactive	4a. Interview before lesson	production	Control beliefs, subjective norms
MANAGE			
THINK	5. Videos	reproduction	Teaching behaviour
TEACH			
CLIMATE	4b. Interview after lesson	production	Control beliefs
Post-active	6. Assessment	artefact	Teaching behaviour
ASSESS			
REFLECT	7. Reflections	production	Control beliefs, subjective norms

Figure 5.2 Elements, forms, and contents of the initial teacher portfolio design

of students in doing research), and behaviour (e.g. the way teachers instruct and coach their students).
This initial description of the portfolio was the input for a panel study to assess the content relevance and representativeness of the design and to decide the weighting to be attached to the portfolio elements in a score model.

5.3.2.3 Panel study and portfolio preparation

5.3.2.3.1 Participants

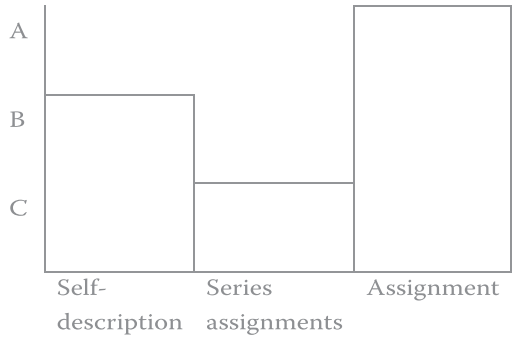
Panellists. We selected eight panellists that participated in earlier studies in which the content standards and performance standards were developed (Van der Schaaf et al., 2003; 2005c). In these earlier studies, the panellists judged the content standards as formulated thoroughly and clearly (mean scores above 3.3 on a 4-point scale) and as sufficiently recognizable in practice (means from 3.0 up to 3.4). On all content and performance standards the degree of consensus was high. The panellists were experienced teachers, school principals and trainers. None of the panellists had ever developed teacher portfolios before.
Teachers. From a random sample of 115 upper secondary schools in the Netherlands (20% of the population) we approached the heads of the economics, geography, and history departments with information about the study and a request for participation. Twenty-one teachers from twenty-one schools were willing to participate: 10 history, 7 economics, 4 geography. This low response can be explained by the demanding character of the study in terms of the required motivation, time, and experience in teaching students research skills. Three of the teachers (one randomly selected per subject) constructed a portfolio in a pilot study. Eighteen developed a portfolio in the main study.

5.3.2.3.2 Procedure

The procedure of the panel study contained the following steps (the steps 1 to 4 parallel the

steps in the Policy capturing study in chapter 4).

1. *Training.* In a plenary training the panellists were informed about the aim, task and procedure of the panel study. They were instructed about the content standards, the anchor points and the initial proposal of the portfolio design (Appendix 1). After that they trained in judging profiles (see chapter 4; Van der Schaaf et al., 2003).
   The panellists' task to be trained involved judging fictitious teacher portfolio profiles per content standard. A configuration of fictitious scores on one content standard is shown in *Figure 5.3*. Each portfolio element met one out of three anchor points, namely the element satisfies the content standard: A, completely or to a strong degree; B, to a reasonable or to some degree; C, to a small degree or not at all. The profiles varied in their scores on the content standards.
2. *Judgments.* The panellists received a booklet containing a series of fictitious profiles of teachers that they had to judge individually on the three anchor points. With 2 or 3 portfolio elements per content standard and three anchorpoints 32 or 9, respectively 33 or 27 profiles can be constructed per content standard. From these, per content standard (the columns in Appendix 2) we drew samples of 6 profiles per content standard with 2 portfolio elements in the initial design, respectively 9 profiles per content standard with



Content standard 1 (GOAL), anchor points: A, the teacher satisfies the content standard completely or to a strong degree; B, the teacher satisfies the content standard to a reasonable or to some degree; C, the teacher satisfies the material to a small degree or not at all

Figure 5.3 Example of a profile

   3 portfolio elements in the initial design (content standard 2 (ASSIGN) with only 1 portfolio element was not judged in the Policy capturing study). The ratios between the number of profiles per content standard and the number of portfolio elements were 3:1, which does not meet the recommended minimum of 5:1 (Cooksey, 1996). In total the panellists judged 54 profiles on three anchor points. They judged the profiles individually at home.
3. *Feedback.* The panellists received feedback, containing the group results of the multiple regression analysis (see *Table 5.1*) and a summary of the personal judgments.
4. *Discussion.* The feedback was the input for a meeting in which the panellists discussed their underlying arguments for their judgments and tried to reach consensus. We used

the nominal group technique (Delbecque, Van de Ven & Gustafson, 1975), consisting of a structured discussion in which the panellists alternate between working individually and in plenary sessions. During the plenary sessions they clarified their arguments and tried to convince the other panellists of their judgments. We improved the initial portfolio de sign based on the results of this meeting.

5. *Portfolio pilot study.* A prerequisite for sound judgments of teacher portfolios is that the portfolios show interpretable evidence. To check for this, in a pilot study three independent teachers (economics, history, geography) voluntarily developed a teacher portfolio according to the revised portfolio design.
6. *Judgments.* As panellists reach better results if they have the opportunity to change their earlier judgments (Jaeger, 1990), the panellists were asked again to judge the content relevance and representativeness of the portfolio and the weighting of the portfolio components. They filled out a questionnaire concerning the content relevance and representativeness of the portfolio per content standard (see Appendix 3). Further, they again indicated the weights of the portfolio elements, by individually dividing 100 points over

Table 5.1 Results of the multiple regression analysis (method enter)

	b[1]	Beta	T (p<.05)	Partial correlation	R	R^2	F (df) (p <.001)
1.GOAL					.72	.52	27,51 (77)
Self-description	0.11	0.13	n.s.	.11			
Series assignments	0.37	0.38	3.01	.32			
Assignment	0.44	0.54	5.15	.51			
3. MANAGE					.77	.59	36,43 (77)
Interview	0.27	0.32	3.63	.38			
Video	0.48	0.58	6.72	.61			
Assignment	0.44	0.33	3.80	.40			
4. THINK					.81	.65	47, 78 (51)
Interview	0.41	0.52	6.27	.66			
Video	0.82	0.62	7.50	.72			
5. TEACH					.80	.64	45,29 (77)
Interview	0.29	0.33	4.52	.46			
Video	0.55	0.70	9.72	.74			
Assignment	0.54	0.65	8.82	.71			
6. CLIMATE					.90	.81	107,18 (51)
Interview	0.33	0.38	6.24	.66			
Video	0.71	0.83	13.45	.88			
7. ASSESS					.89	.79	94,59 (51)
Assignment	0.82	0.86	13.22	.88			
Assessment	0.25	0.17	2.67	.35			
8. REFLECT					.79	.63	42,85 (77)
Self-description	0.60	0.58	7.41	.65			
Interview	0.16	0.18	n.s.	.20			
Reflections	0.49	0.56	6.05	.57			

NOTE. Content standard 2 (ASSIGN) is not judged in the Policy capturing study, because this content standard was based on only one portfolio element in the initial design (the research assignment). [1]) se 0,1

the eight required portfolio elements (Swanborn, 1999) and they individually indicated the weights per portfolio element per content standard in percentages.

7. *Portfolio preparation.* In the main study 18 teachers prepared a portfolio based on the portfolio design developed in the previous steps 1 to 6. The teachers received instructions including a description of the aims of the study, the content standards, and the assessment procedure. All portfolios were constructed within a few months time span, focusing on teaching students research skills.
8. *Teacher evaluation.* We investigated whether the teachers experienced possible sources of bias during the preparation of their portfolios. In total 21 teachers (3 in the pilot study and 18 in the main study) developed portfolios in accordance with the revised portfolio design in our study. After the portfolio construction, they filled out a questionnaire containing questions about sources of bias on a 5-point scale (1 = I strongly disagree; 5 = I strongly agree) (see Appendix 4). Also they answered open questions regarding their portfolio preparation.

5.3.3 Data analysis

Panel study. We used multiple regression analysis to analyse the judgments on the fictitious profiles in step 2 (see chapter 4 for a detailed description of the analysis process). Further, we analysed the weights given in step 6.

We analysed the transcriptions of the plenary meetings (steps 1 and 4) for emerging themes. The key issue in the data analysis was the decision whether particular portfolio elements should be accepted, revised, or deleted. We based this decision on the panellists' ratings on the aspects content relevance and representativeness in the questionnaire in step 6. Firstly, we wanted to know whether the panellists supported, opposed, or were ambivalent towards the initial portfolio design (4 = strongly supports, 3 = weakly supports, 2 = weakly opposes, 1 = strongly opposes the initial portfolio design). We used the panel's mean rating to indicate whether the panellists supported the design. Secondly, we analysed panellists' written comments for emerging themes and categories.

Teacher evaluation. We investigated the consistency between the items in the questionnaire by computing Cronbach's alphas and we computed frequencies of teachers' evaluations. Further, we qualitatively analysed teachers' comments on open questions concerning their portfolio preparation for emerging themes and categories.

5.3.4 Results

Regarding the multiple regression analysis, the usual assumptions of normality, homoscedasticity and linearity were sufficiently met, except for the content standards 1 (GOAL) and 8 (REFLECT). The portfolio elements of these content standards (the predictors in the model) were collineated; the Pearson correlations were between .35 and .64.

Table 5.1 shows that the explained variance per content standard lies between R^2 = .52 and R^2 = .81 (R between .72 and .90). The results indicate that the panellists prefer artefacts (e.g. assignments) and reproductions (e.g. video) as portfolio elements. We should be careful with interpreting these results due to the fact that the ratios between the number of profiles per content standard and the number of portfolio elements were low and the portfolio elements of content standards 1 (GOAL) and 8 (REFLECT) were collineated.

After being confronted with these results, in a nominal group discussion (step 4) the panellists gave reasons for their choices. Important themes that arose were that they thought

artefacts and reproductions are trustworthier than productions. Further, they mentioned that productions would require extra time investment of teachers and would not fit teachers' daily work in the way artefacts and reproductions do. Also, they found it easier to judge profiles with high scores on portfolio elements that they saw as more reliable (i.e., artefacts and reproductions) than profiles with high scores on productions and low scores on artefacts and reproductions. In addition, though in general the experts judged the video as an important portfolio element, they thought the content standard 4 (THINK) should be based on the content of interviews only and not also on the video. In their eyes the video could only be used to check whether teachers teach like they say they think they teach and not to check teachers' previously thinking and selecting of teaching strategies. Finally, the panellists suggested adding students' evaluations of teacher performance in a short questionnaire, focussing on the content standards 2 (ASSIGN), 3 (MANAGE), 5 (TEACH), and 6 (CLIMATE).

Next, we revised the initial portfolio design by removing the portfolio element 'video' from content standard 4 and by adding a short student questionnaire (see Appendix 5 and chapter 7).

In the pilot study (step 5), three teachers successfully constructed a portfolio based on the revised portfolio design.

Exploratory factor analysis followed by item analyses of the questionnaire data (step 6, see Appendix 3) resulted in three sufficiently reliable Likert scales: (a) the relevance of the portfolio elements (8 items), e.g. 'the portfolio elements are relevant for teaching students research skills' (Cronbach's alpha .85; mean 4.2; sd .49). (b) the usability of the portfolio (7 items), e.g. 'it is clear to teachers how to develop a portfolio' (Cronbach's alpha .89; mean 4.1; sd .59). (c) the fairness of the portfolio (4 items), e.g. 'it is fair to use the portfolio model for teacher assessment' (Cronbach's alpha .91; mean 4.0; sd .75). So, in general the panellists supported the portfolio design.

Finally, we asked the panellists about the relative weight they give each portfolio element. Six of the eight panellists responded to this question and individually divided 100 points over the eight required portfolio elements (Swanborn, 1999). In total, the results showed that most panellists favour the video (range 15 points to 35 points), the assignment (range 10 to 20) and student evaluations (range 10 to 20). As in the Policy capturing study, they mostly questioned the self-description (range 0 to 5), and the reflections (range 5 to 10), because these elements are not part of teacher's authentic practice.

Based on the results of this phase, we constructed a judgment manual for the raters including the content standards, the anchor points, the portfolio elements per anchor point, and an indication of the weighting of the portfolio elements (see Appendix 1). The elements self-description and reflections are given less weight in the scoring model of the assessment.

Teachers' evaluation of the portfolio design. The questionnaire data included three scales (Appendix 4). The first scale contained questions about the usability (5 items), e.g. 'constructing the portfolio following the manual is usable' (Cronbach's alpha .88; mean 3.5; sd .95). The second scale was about the representativeness of the portfolio (3 items), e.g. 'the portfolio reflects what I think and do when teaching students research skills' (Cronbach's alpha .68; mean 4.2; sd .54). The third scale regarded the process of preparing a portfolio (7 items), e.g. 'If the content of the portfolio was more structured, I would be more able to construct a portfolio that reflects my practice' (this item is recoded in the analysis) (Cronbach's alpha .83; mean 3.9; sd .71).

Furthermore, directly after the portfolio construction the teachers answered open questions

regarding the representativeness of their portfolio and they were given the possibility to make slight adjustments in their portfolios. Although all teachers confirmed that their portfolios represented their practices, four teachers made refinements. Three teachers made small distinctions in the assessment of student research (e.g. "In practice the scoring model is a bit more focused on the product of the research and less on the preparation of the research"). One remark concerned the lack of time that caused less individual coaching of students than planned. Finally, we asked teachers about the difficulty of preparing a portfolio. In general, teachers conceived the gathering of artefacts and reproductions to be easy parts of the preparation process, while they saw the construction of productions as difficult (especially those productions in which they had to give account of their own thoughts and motivations). In total, the results show that, despite of the difficulty in constructing productions, according to the teachers few systematic biases occurred.

5.4. Scoring teacher portfolios in congruence with the content standards

5.4.1 Overview

The second sub study was about a condition for accurate assessment, namely that raters' judgments are based on the assessment standards as taught in the rater training (Borman, 1977). The accurateness with which raters employ these standards will be influenced by the quality of the standards, the quality of the scoring form, and the instruction and training given to the raters. Therefore, we verified these conditions by analysing raters' comments made in the judgment forms used while rating the teacher portfolios. The scoring of the portfolios reflects judgments in which several aspects of the standards, the portfolios, and the teachers' performances are interpreted and weighted in the mind of the raters. For example, a decision must be made in regard to how the content standards are applied to create a score.

5.4.2 Method

5.4.2.1 Participants

From the sample of panellists we selected six raters who successfully participated in the previous studies. The raters, who, as it happened, were all male, consisted of a school principal who was also an experienced teacher, two experienced geography teachers who were also teacher trainers, two experienced history teachers, an experienced teacher of both geography and history, and an experienced teacher of social science subjects who was also a teacher trainer in economics. The teacher trainers were working at several university teacher training institutes. All raters received financial compensation for participating. None of the raters had assessed teacher portfolios before participating in our project.

5.4.2.2 Procedure

The raters were thoroughly trained in assessing teacher portfolios. They studied a manual describing in detail the objectives, planning, and procedures of the assessment, the content standards, the anchor points, and the portfolio elements (see Appendix 1). The raters first individually studied an example portfolio. They then participated in a training session (four hours plenary) (see chapter 6 for details). After the training, the raters individually rated three portfolios in a pilot (see before, section 5.3). After the pilot, the raters received feedback about their scoring and suggestions for improving their judgments. Suggestions included

the accurate use of the standards, the assessment procedure, and points of special interest concerning the interpretation.
Next, in a 9-month period, we assigned the 18 portfolios in the order they became available. Rater 1 judged six portfolios (mainly economics), rater 2 four portfolios (history), rater 3 nine portfolios (mainly history), rater 4 five portfolios (geography and history), rater 5 seven portfolios (geography, history and economics), and rater 6 five portfolios (mainly economics). We aimed at using fixed couples per subject, but due to some availability problems we had to allow for a few departures from this scheme.
The raters used judgment forms to rate the portfolios. The judgment forms were designed according to the Correspondent Inference Theory (Jones & Davis, 1965). Firstly, raters illustrated each content standard with significant portfolio material. Secondly, they described their interpretations of the material. Finally, raters assigned a score to the portfolio for each content standard, using a 5-point scale with anchor points (analytic score). They added an overall (holistic) judgment on a five-point scale, considering the total of the evidence from the separate scores (holistic score). The rating process produced two types of scores: (unweighted) mean analytic scores focused on the mean of the scores for the eight content standards and holistic scores being overall judgments of the portfolios.

5.4.3 Data analysis

We used the percentage exact agreement in holistic scoring and the Cronbach's alpha coefficient on the content standards per portfolio (the jury alpha) between the rater pairs to estimate interrater reliability.
We checked whether raters' judgments were based on the content standards by coding the content of raters' judgments in their written judgment forms and comparing this with the full descriptions of the content standards and indicators in our study (see before, section 5.2.3). A researcher (the author) and an assistant researcher qualitatively analysed the content of 36 raters' judgment forms (two judgments per portfolio). We verified per content standard per indicator whether it was referred to by the raters in their written judgments. We selected the indicators referred to in more than half of the judgment forms (see *Table 5.2*). For each reference, we checked whether the rater's comment was positive (score 1), neutral (or balanced) (score 2) or negative (score 3). In the pilot study on three at randomly chosen portfolios comparing the scores of the researcher and the assistant researcher it turned out that the Cohen's kappa on this dimension positive-negative was .79.

5.4.4 Results

The eight content standards formed a reliable scale (Cronbach's alpha .76). An Anova did not show any rater variance. The jury alphas for the separate analytic scores were satisfactory for 12 rater pairs, ranging from .39 to .76. The jury-alphas were low or even negative for six pairs, ranging from -.80 (for one pair) and -.11 to .22. The raters appointed exactly the same holistic scores on the five-point scale to 35% of the judgments, a difference of half a point showed up in 12% of the judgments, a difference of one point in 47% of the judgments, and a difference of 1.5 points in 6% of the judgments (this concerns one rater pair).
We checked whether raters' judgments were based on the content standards (the link between (iii) scoring model and (iv) assessment model) by coding the content of raters' written judgments in their judgment forms. We selected the content descriptions and indicators (see section 5.2.3) referred to in more than half of the judgment forms (n = 18, 2 judgment

forms per portfolio). We correlated the comments given on the indicators with the mean analytic score computed per content standard. *Table 5.2* shows the mean of raters' written judgments on the content standards in their written judgment forms on a three point scale (1 = positive comment; 2 = neutral or balanced comment, 3 = negative comment) and Pearson's correlations between the mean of raters' written judgments and raters' mean analytic scores (bold) on the indicators referred to in more than half (n >18) of the 36 judgment forms. Twenty correlations were significantly positive ($p < .05$). All correlations were between .37 and .83.

Table 5.2 Significant Pearson's correlations ($p < .05$) between the mean of raters' written judgments and raters' mean analytic scores (bold)

	n	r	mean	sd
1.GOAL	36		**2.03**	.43
a) teacher's long term goals	27	.68	2.27	.76
b) approach used to reach these goals	33	.82	2.06	.84
2. ASSIGN	36		**2.24**	.44
a) goals related to subtasks research	35	.61	2.48	.73
b) content is related to research in the subject	23	.69	2.59	.59
c) form authentic, offers options, variously approachable	22	.37[a]	2.42	.70
d) fits close to students' level	26	.63	2.42	.78
3. MANAGE	36		**2.18**	.50
a) preparation and the communication	28	.38	2.54	.74
b) management of facilities	29	.50	2.41	.82
4. THINK	36		**1.99**	.58
a) choice and argumentation of teaching strategies	26	.72	2.46	.81
d) teaching strategy fits assessment	22	.57	1.95	.90
5. TEACH	36		**1.89**	.49
a) use of teaching strategies	31	.74	2.06	.85
b) fits subtasks of doing research	21	.74	2.00	.84
d) fits close to students' level	27	.55	1.81	.83
6. CLIMATE	36		**2.20**	.39
a) climate	28	.74	2.60	.74
7. ASSESS	36		**1.81**	.52
b) clarity and comprehensibility of assessment approach	32	.65	1.81	.86
c) the way the teacher handles the assessment	26	.61	1.85	.83
d) communication of the results to the students	19	.71	1.57	.84
8. REFLECT	36		**1.78**	.44
Aware of strong and weak points in:				
a) his or her teaching	22	.83	1.81	.91
b) the educational program	17	.68	1.77	.86
His or her suggestions for improvement in:				
c) his or her teaching	30	.78	1.71	.77
d) the educational program	19	.59	1.63	.83

NOTE. [a] = not significant

5.5 Conclusions

The purpose of this study was to design a teacher portfolio based on a chain model for construct validation. We researched into the questions (1) How is an initial portfolio design related to the content standards? (2) How is raters' portfolio scoring related to the content standards?

We developed an initial portfolio design based on the content standards created in an earlier study (Van der Schaaf et al., 2003, 2005c) and on the Theory of Planned Behavior (Ajzen, 1985; Ajzen & Fishbein, 2005). This initial design was the input for a panel study in which eight panellists assessed the content relevance and representativeness of the design and weighted the portfolio elements. We used a Policy capturing method with discussion to support the judgments of the panellists. The panel study resulted in a slightly revised portfolio design (e.g. adding a short student questionnaire). Twenty-one teachers (3 in a pilot and 18 in the main study) prepared a portfolio based on this final design.

After the revisions had been made, the panellists agreed rather strongly with the portfolio elements accompanying the content standards. According to the teachers, the revised design was quite feasible. Further, both the panellists (in the judgment study) and the teachers (that prepared portfolios) preferred artefacts and reproductions as portfolio elements to productions. In the Policy capturing study, the panellists said to find it easy to judge those profiles consisting of portfolio elements that they find valuable. Accordingly, they found it more difficult to assess profiles containing portfolio elements concerning teacher cognitions than those mainly concerning teacher behaviour. This result is significant, because research shows that the reliability of rater assessments is generally higher for those variables that people regard as being appropriate descriptors (Amelang & Borkenau, 1986). This implies that for a reliable portfolio assessment, the portfolio should involve portfolio elements that are relevant in the eyes of the raters.

As raters and teachers experienced more difficulty with productions (intended as a basis for assessment of underlying cognitions), one can consider designing a portfolio model based on artefacts and reproductions only; this parallels the idea of a teacher dossier (Peterson, 1995). This restriction could have the effect of improving the reliability of portfolio scoring. Artefacts and reproductions are less subject to multiple interpretations, because these forms are commonly defined in terms of specific behaviours or products that are more readily observable than teachers' underlying cognitions. However, an insurmountable obstacle of such a restriction would be that the portfolio is less congruent with the construct domain (i.e., teacher competences), resulting in a decrease of the validity of the design. Moreover, teacher cognitions (including beliefs and intentions) partly provide the interpretative context within which raters make assumptions about the different elements in a portfolio. Information of teacher's underlying knowledge and motivations can even be seen as the 'glue' that holds the different (interpretative) parts of the portfolio together. These drawbacks feed our conclusion that in addition to evidence of teacher performances also information about teachers' cognitions underpinning his or her teaching is essential (Eraut, 1994).

In our study six raters systematically assessed 18 portfolios. The interrater reliability of the analytic scores of 12 portfolios was satisfactory. An Anova showed only slight rater effects. In all but one case raters' holistic scores per portfolio per rater pair differed by not more than one point (on a 5-point scale). We gathered evidence of the correspondence between the content standards and the scoring of teacher portfolios by analysing the raters' comments

in written judgment forms while rating the portfolios. The results indicated that raters quite strongly based their assessments on the content standards as trained.

A weak point of teacher portfolios is the limited number of portfolio elements. In our study teachers only used four artefacts, three productions, and one reproduction to show their competences in teaching students research skills. Such a limited sample may impact the generalizability of portfolio assessment (the degree to which the assessment is valid for a broader range of tasks) because the error variance tends to increase relatively as the sample size decreases. Therefore, we prefer not to use portfolios as stand-alone assessments but to include them in bigger assessment batteries in which portfolios supplement other assessment instruments. Beside statistical indicators concerning the generalizability of portfolios, additional indicators are desirable. These may concern the teachers' ability to interpret the meaning of the portfolio instructions and the variables, and the raters' ability to interpret the meaning of the content standards and to link these standards to their judgments based on the portfolios.

Our study gives cause to further research. In the first place, experimental studies in which variables in portfolio designs are varied will be constructive for this purpose (e.g. Smith & Tillema, 2001; Stokking & Van der Schaaf, 2004). Recently we started such a quasi-experimental study comparing different portfolio designs and their impact on students' competence development in pre-university education.

Secondly, as a collection of the teacher's work material, a portfolio supports broad and context-bound assessments, giving an important role to the teachers themselves. Depending on its content and form, portfolios also fit into the holistic character of competences and the personal character of a teacher's work over a larger period of time (Bird, 1990). Little is known about underlying cognitive processes of teachers in preparing a portfolio, but such information would be very useful to improve portfolio designs. Especially think aloud studies of teachers' thoughts during the preparation of a portfolio might be valuable in this respect.

Thirdly, one way to contribute to the theoretical understanding of the assessment scores given is by showing that particular rater cognitions generate certain assessments scores. Therefore, research into raters' cognitions should play an important role in portfolio validation research. We initiate such research in the following chapter by exploring raters' cognitive representations in think-aloud protocols.

6 Cognitive representations in raters' assessment of teacher portfolios

Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (2005). Cognitive Representations in Raters' Assessment of Teacher Portfolios. *Studies in Educational Evaluation, 31*, 27-55.

Based on: Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (2005). De invloed van cognitieve representaties van beoordelaars op hun beoordeling van docentportfolio's. *Pedagogische Studiën, 82* (1), 3-26

Abstract

Portfolios are frequently used to assess teachers' competences. In portfolio assessment, the issue of rater reliability is a notorious problem. To improve the quality of assessments insight into raters' judgment processes is crucial. Using a mixed quantitative and qualitative approach we studied cognitive processes underlying raters' judgments and the reliability of these judgments. Six raters systematically assessed 18 portfolios. The interrater reliability of 12 portfolios was satisfactory. An Anova showed slight rater effects. We used the Correspondent Inference Theory (Jones & Davis, 1965) and the Associated Systems Theory (Carlston, 1992; 1994) to analyse judgment forms and retrospective verbal protocols. Raters' cognitive representations on the dimensions abstract–concrete and positive–negative were significantly related to the judgments given and to the reliability of these judgments.

6.1 Introduction

In many countries interest is growing in the assessment of teachers' competences (knowledge, skills, and attitudes). This interest is fostered by current needs for accountability and quality improvement in the teaching profession (Cochran-Smith & Fries, 2001). Paralleling the movement towards alternative assessment of students (Boud, 1990; DeCorte, 1996; Dierick & Dochy, 2001), teachers' competences, too, are increasingly assessed by competence-based assessments, using instruments such as portfolios (Delandshere & Petrosky, 1994; Delandshere & Arens, 2001). Depending on its content and form, a portfolio can do justice to the fact that teaching is a complex activity and that teachers' behaviour is bound up with their cognitions and the teaching context (Andrews & Barnes, 1990; Bird, 1990; Lyons, 1998).

The increased use of competence-based assessment prompts discussions about the appropriateness of traditional psychometric criteria, such as reliability and validity, for such new forms of assessment. Reliability is often seen as a prerequisite for validity, but attaining acceptable levels of reliability in portfolio assessments is a notorious problem. This is due to the fact that assessing portfolios involves complex interactions between teachers' competences, the portfolio, the standards used, raters' characteristics, and raters' interpretations. Several studies have shown that in assessment, raters' cognitive activities are a major source of variance (DeNisi, Cafferty & Meglino, 1984; Feldman, 1981; Landy & Farr, 1980). Those cognitive

activities can be described from the point of view of the content of a rater's cognitions (e.g. focussing on a teacher's behaviour or on the rater's personal beliefs), their type (e.g. interpretation or judgment), and their character (e.g. more or less explicated, situated and personal). In this study we focus on the content and the type of raters' cognitive activities in assessing teaching portfolios, especially raters' cognitive representations.
To describe raters' cognitive activities while assessing other people several general models can be used (Gilbert, 1989; Jones & Davis, 1965). However, raters' cognitive representations in assessing teacher portfolios are rarely examined. Insight into these representations is needed to facilitate and enhance our understanding of portfolio assessments and to improve their quality. This study is focused on three questions:

1. What is the reliability of teacher portfolio assessments?
2. How can raters' cognitive representations during assessment be described?
3. What is the relationship between raters' cognitive representations, their judgments and the reliability of these judgments?

The study is part of a larger research project on the assessment of experienced teachers' competences. In two previous studies we developed both content standards (criteria) and performance standards (cut-off scores) for teaching research skills in upper-secondary (ages 16–18) social science education (Van der Schaaf, Stokking, & Verloop, 2003, 2005c). Teaching research skills is a recent Dutch curricular innovation which fits well with the change in many countries towards more constructivist views on learning and developing students' skills in an active, self-directed and collaborative way.

6.2 The reliability and validity of raters' judgments

6.2.1 Psychometric quality criteria

Portfolio assessment is a tool for ascertaining whether teachers satisfy required competences and for generating guidelines for professional development. In the first case, summative assessment is used to account for the teacher's expertise, with possible consequences such as certification and merit pay. In the second case, assessment has a formative goal and produces information that can be used for planning further professional development. In this article, we focus on the assessment of experienced teachers for the purposes of both summative and formative assessment.
Portfolio assessments should meet certain quality criteria (see chapter 2). Main quality criteria are reliability and validity. Reliability refers to the extent of stability between measurements and between raters. If the same rater repeats the assessment after an interval of several weeks, or if another rater also assesses the portfolios, the judgments should not differ too much. Validity refers to the question whether the assessment is measuring the intended construct, competence or task performance. Today the traditional psychometric criteria of reliability and validity are further differentiated and extended with criteria of acceptability and utility (Messick, 1989; Moss, 1992; Stokking, Van der Schaaf, Jaspers & Erkens, 2004). Acceptability is about the assessment's objectivity, transparency, equality and unbiasedness, and utility is concerned with functionality, feasibility and efficiency.
In our view, for both summative and formative aims validity should have priority (Linn, 1994;

Linn, Baker, & Dunbar, 1991; Borsboom, Mellenbergh, & Van Heerden, 2004), because assessment is primarily concerned with the intended competences. Transparency, unbiasedness, and functionality should also be satisfactory. In formative assessment, reliability, objectivity, and equality may receive less priority, because the results only indicate possibilities for improvement. Practical utility is also important, in order to be able to provide frequent and useful feedback. In summative assessment, reliability, objectivity and equality are also important (Stokking et al., 2004).

How portfolio assessments precisely can meet these criteria remains as yet unclear (Linn, 1994). In this study we focussed primarily on the reliability criterion, because reliability is a basic requirement and in portfolio assessment attaining acceptable levels of reliability has proved to be particularly difficult (Burns, 1999; Johnson, McDaniel & Willeke, 2000; Linn, 1994; Reckase, 1995; Shapley & Bush, 1999).

To estimate the interrater reliability of portfolio assessments different approaches are available, including different ways of summarizing ratings across raters, which may also impact the validity of the assessment results. Ideally the method of estimating interrater reliability is consistent with the underlying assumptions of the assessment study at hand. We shall now describe some available approaches and arguments for choosing among them.

Firstly, reliability is often estimated by the percentage of exact agreement in scoring between raters. Exact agreement is based on similarity of classification (Popping, 1983) and presumes that the classifications given are identical. This can only be realized when raters interpret the portfolio artefacts and the standards in completely the same way. As people always may differ in their interpretations (Huot, 1993; Kane, 1992; Van der Schaaf et al., 2003), these conditions cannot be guaranteed. Therefore, interrater reliability in terms of percentage agreement can only be used as an indicator of reliability.

Secondly, interrater reliability can be estimated by computing Cronbach's alpha coefficient for the raters' judgments on the content standards per portfolio (the so-called jury alpha). An underlying assumption is that two raters do not necessarily have to share a common interpretation of the standards and the portfolios, as long as each rater is consistent in classifying the portfolios according to his or her own interpretation of the standards. It is important to recognize that although the alpha may be high, the raters' judgment means may be different. Using correlations as an indicator for interrater reliability, we depart from the idea that reliable judgments should have similar mean scores *per se* (Murphy & De Shon, 2000). As subgroups of raters can have comparable interpretations, different interpretations within a group of raters are not necessarily idiosyncratic and rater variance effects do not necessarily indicate rater errors. That is, assuming the ratings are accurate. A condition for accuracy is that raters' judgments correspond with the performance theory (i.e., the standards) as taught in the training given to the raters (Borman, 1977).

Thirdly, reliability is also discussed using analysis of variance. Variability in scoring is generally based on, amongst other things, individual rater bias and systematic differences between raters, due to factors such as rater background and rater expectations. In generalizability studies, analysis of variance estimates the contribution of multiple sources of error to the score variance. Although some performance assessment studies report that the variance related to the judges makes a negligible contribution to the overall variance (e.g. Shavelson, Baxter, & Gao, 1993), this result does not occur when the assessment task is more complicated (Dunbar, Koretz, & Hoover, 1991).

As to the validity criterion of portfolio assessments, verbal protocols are often used to cap-

ture a rater's thoughts while assessing. This is also relevant in our study. Two main forms of retrospective judgment protocol invalidity are reactivity and nonveridicality (Russo, Johnson, & Stephens, 1989). In the first case, a change of the primary cognitive processes occurs due to either the verbalization process or a prolonged response time. Nonveridicalities (inaccurate reflections on the underlying primary cognitive processes) contain errors of omission (not reporting some thoughts owing to overlooking or forgetting) and fabrication (reporting mental events that did not occur). Fabrication must be taken especially seriously because fabricated data are indistinguishable from other protocol data. Several researchers have warned against retrospective verbal protocols, and in particular stimulus-cued methods may lead to substantial fabrication (Ericsson & Simon, 1980, 1984; Russo et al., 1989). We assume that raters do not present a pure report of their underlying mental processes. Rather, they construct a representation of their judgments containing crucial information about the processes involved, including the communication within the context shared by the rater and the researcher (Long & Bourgh, 1996). In sum, verbal protocols provide useful and unique information about mental activities when fulfilling cognitive tasks, but they do not transparently reflect raters' mental processes when they are involved in silent assessment tasks. For that reason, verbal protocols are most valuable when used in combination with other data gathering instruments.
More details about the methods we used to further and control for the reliability and validity of the judgments in our study are described in the method section.

6.2.2 Raters' cognitive representations

Over the past decades, the assumption that raters' cognitive representations influence their judgments has been fed by social cognitive psychology. Several rating process models have been built that describe assessment activities (e.g. Landy et al., 1980; Feldman, 1981; DeNisi et al., 1984). Those models have in common that raters are using schemata while assessing, predicting and understanding another person's behaviour. These schemata are comparable to personal constructs (Kelly, 1955): content categories used to organize and simplify information. Interpersonal filters in these schemata induce raters to look selectively at the information concerning the people observed and to interpret this information according to their own constructs.
Rating process models as described in the literature are often related to Jones and Davis' (1965) Correspondent Inference Theory. This theory assumes that in the process of drawing inferences about others' behaviours, several activities occur (Gilbert, 1989). Raters categorize others' behaviour (e.g. "Irene speaks in a strict way") and relate this to feasible corresponding characteristics (e.g. "Irene must be a strict sort of person"). Meanwhile, they correct their characterizations by considering situational information (e.g. "Irene's students are very noisy, so Irene's strictness is reasonable. Maybe Irene is not such a strict person after all"). During the process, raters may express positive and negative utterances (Huot, 1993). Gilbert (1989) suggests that cognitive activities differ in their degree of consciousness. Our general impressions of others often remain implicit (Carlston, 1994). Further, the need to process information about the context of another's behaviour is not always obvious; when people are inattentive or unmotivated, they may fail to consider such situational information.
Consequently, when designing rating formats and training procedures it is important to be attuned to natural cognitive rating activities. Research into raters' judgment activities shows that it is possible to influence those activities. Various rater training procedures have been

developed for this purpose (Lievens, 2001; Woehr & Huffcutt, 1994). Sceptics assume that rater training may only have minor effects on rater judgments because raters undergo socialization processes in their (professional) lives. Training periods and work experiences cannot easily replace mental representations formed during many years (cf. Huot, 1993).
Nevertheless, raters' cognitions form an essential domain that needs further exploration in order to understand better, and to improve, raters' judgmental processes (Day & Sulsky, 1995).

6.2.3 Associated Systems Theory

One way to study these judgmental processes is by using the Associated Systems Theory (AST) (Carlston, 1992, 1994). This theory focuses on several forms of human cognitive representations that operate simultaneously when forming impressions of other people. Furthermore, AST has proved to be usable for researching raters' judgments. For example, Schleicher and Day (1998) successfully used this theory to reveal impressions formed by assessors in an assessment centre setting.
The Associated Systems Theory is about person perceptions, impression formation, and social cognition. The theory focuses on the origin, organization and use of different kinds of mental representation of (other) people and events. Expanding on previous research in social psychology and neurology (Fiske, 1992; Martindale, 1991), two main beliefs within AST are doing causes thinking and thinking is for doing. The first principle says that people's cognitive representations develop by experiences derived from their actions. Translated to our study, raters' mental representations are influenced by their teaching experience, rater training, and experience in scoring portfolios. Secondly, mental representations can be seen as concepts that mediate the input of external stimuli and the output of actions (behavioural responses) (Norman, 1985). Thus, cognitive representations are fundamental to activities such as assessing portfolios.
AST offers a starting point for classifying assessors' representations. Carlston (1992, 1994) models AST on two dimensions (see *Figure 6.1*).

	Concrete		**Abstract**
Target	(1a) Visual appearance	(1b) Categorizations	(1c) Attributed personality traits
	(2a) Behavioural observations	(2b) Mix	(2c) Evaluations
Self	(3a) Behavioural responses	(3b) Orientations	(3c) Affective responses

Figure 6.1 Structural representation of the AST taxonomy

1. Concrete versus abstract. Concrete representations (left column) may include details of time and place of their recording (e.g. observing someone's physical appearance). The forms in the centre column recapitulate multiple observations and are therefore somewhat more abstract (e.g. attributing personal characteristics (e.g. lazy) based on someone's appearance). Abstract representational forms (right column) reflect general characteristics of the person and are not restricted to particular situations. Describing other persons in terms of personal characteristics in general is more cognitively demanding than providing concrete descriptions of someone's physical appearance. So this dimension also represents an expected increase of the cognitive activities needed.
2. Target-referenced versus self-referenced. Although mental representations are always more or less subjective and rater-bound, the strength of this varies. Target-referenced representations focus primarily on the target or person being assessed. Self-referenced representations concern rater's personal reactions to the assessee. Generally, personal reactions are based on a relatively stable cognitive structure and, compared to attitudes, may therefore be more difficult to change (Fazio, 1994). As raters always mentally interact with the assessment situation and with the assessee, assessing generally involves a mix of target-referenced and self-referenced representations (see middle row of *Figure 6.1*). Hence, their personal schemata filter their observations, their interpretions of their observations, and their final judgments (Tulving, 1983). So, even if an assessor does not physically interact with the situation and the person assessed, the assessor is always mentally involved.

Carlston (1992; 1994) specifies raters' cognitive representations as the cells in a matrix (*Figure 6.1*). Each cell represents a different type of cognitive representation.

1a Visual appearance: visual images of physical appearances are primary in our impressions of others. In our study, this is reflected in paraphrasing or referring to teachers' activities in the portfolio.
1b Categorizations (typing): sorting into sets and labelling of representations of others. In our study, this demands an interpretation or explanation of teachers' activities, as derived from the portfolio.
1c Attributed personality traits: describing others in terms of traits or character types. In our study, this refers to describing teachers' competences in accordance with the content standards.
2a Behavioural observations: representations of others' behaviours are intermeshed with perceptions of their appearance and with features of one's own role in the recorded events. This category supposes that the rater mentally or physically interacts with the person assessed in a certain context, which is the case in assessments (Conway, 1990; Tulving, 1972, 1983).
2c Evaluations: verbal concepts used to embody affective feelings (e.g. "I think she has empathy with the students, which is good").
3a Behavioural responses: acts of the rater directed at the person assessed. This category is not relevant in our study because the raters do not interact with the teachers assessed.
3b Orientations: tendencies or predispositions to respond to a person in a particular manner; e.g. approach and avoidance tendencies of raters. This category is not researched in our study.
3c Affective response: "an abstract representation of affect that is linked to physiological structures involved in the primary experience of emotion" (Carlston, 1994, p. 6), e.g. laughing or crying. This category is not relevant in our study.

For validity reasons, it is important that raters alternately use concrete and abstract representations. One prerequisite for content validity is that concrete representations, such as observations of research assignments and video recordings, enhance the assessment's fit to the portfolio content. For the sake of acceptability of the judgment, in terms of its fairness and lack of ambiguity, the raters should also make clear on which concrete data the judgments were based. Furthermore, the assessment results should be suitable for the intended functions, such as providing comments on the teacher's strengths and weaknesses. Therefore, in the interest of practical utility, concrete examples that illustrate judgments are needed to give feedback to the teachers assessed. On the other hand, abstract representations are necessary for ratings to be accurate, according to the performance theory (the standards) as taught in the rater training. Abstract representations are also needed to predict with sufficient accuracy a teacher's performance in situations other than those described in the portfolio and in future employment. Furthermore, accurate interpretations of teachers' portfolios (interpretations that combine concrete and abstract representations) are needed to distinguish correctly between more and less competent teachers (the specificity aspect of validity) as well as to compare judgments on the same content standard based on different portfolio material (the convergence aspect of validity) (see Stokking et al., 2004).
We assume that an increase in the raters' use of target-referenced representations will increase the quality of their judgments of teachers' competences. The use of target-referenced representations increases the likelihood that the judgments are based primarily on teachers' competences and less on raters' subjective representations.

6.2.4 Standards for teaching students research skills

In one previous study we studied teachers' tasks needed to develop students' research skills and the competences (knowledge, skills and attitudes) needed to perform those tasks. We developed content standards that describe what teachers should know and be able to do in teaching students research skills (Van der Schaaf et al., 2005c). Using a Delphi method, in three rounds 21 experts, including most of the raters in the present study, judged and revised a preliminary set of content standards. They judged the standards on a 4-point scale from weak support (score 1) to strong support (score 4). The method resulted in eight standards with a high degree of support. In sum these content standards were (see chapter 3):

1. Selecting goals for students (GOAL): teacher's long-term goals for developing students' research skills and the approach s/he uses to reach these goals.
2. Choosing an appropriate assignment (ASSIGN): the goals, content and form of the as signment.
3. Preparing and managing students to work on their assignments (MANAGE): the preparation, communication, and management of facilities (time, rooms, sources, media, etc.) that students need for working on the assignment.
4. Planning and selection of teaching strategies that support the development of students' research skills (THINK): the choice of and argumentation for teaching strategies that meet students' knowledge, abilities and experience.
5. Teaching students research skills (TEACH): the use of the teaching strategies chosen.
6. Creating a positive pedagogical climate (CLIMATE): the provision of a safe, respectful, and stimulating learning environment for students.
7. Adequately assessing the research skills of students (ASSESS): teacher's argumentation

for, and the clarity and comprehensibility of the assessment approach used (criteria, scoring, and norms), the applicability of the approach to the assignment, and the way the teacher handles the assessment and communicates the results to the students.
8. Reflecting on the program and on actions in teaching students research skills (REFLECT): the extent to which the teacher is aware of strong and weak points in his or her teaching, and his or her suggestions for improvement.

In a second previous study (Van der Schaaf et al., 2003), using Policy capturing we developed the performance standards (the cut-off scores that indicate how good teachers should meet the content standards for a sufficient result). Nine experts, again including the raters in the present study, judged a great number of teacher profiles, which were simulated configurations of scores on the eight content standards. The judgments were analysed by linear multiple regression analysis, yielding a policy including the performance standards and the content standards' weights (the content standards TEACH and CLIMATE being given most weight). This study resulted in three anchor points on the 5-point scales rating teachers' competences on the basis of their portfolios, namely that the teacher satisfies the content standards: completely (score 5), to a reasonable degree (score 3), or not at all (score 1). See Appendix 1 for an example.

6.3 Method

6.3.1 Sample of raters

Having teaching experience themselves helps raters assess teachers' competences (Pula & Huot, 1993). We therefore selected raters with teaching experience. Preparing the selection process in the first-mentioned previous study (described above) we asked 22 stakeholders with different educational positions which kinds of persons should participate in the study. Most stakeholders mentioned practicing teachers and external experts such as teacher educators and pedagogical content experts. We chose to use external raters, that is raters not working in the same schools as the teachers to be assessed, in order to avoid raters biasing their perceptions or judgments on episodic memories and local orientations. We selected raters who successfully participated in the two previous studies (described above) in which content standards, performance standards, and a portfolio assessment procedure were developed. So, the raters were familiar with the project and they subscribed to the standards and the assessment procedure (prerequisites for any assessment process to be successful). The raters, who, as it happened, were all male, consisted of a school principal who was also an experienced teacher, two experienced geography teachers who were also teacher trainers, two experienced history teachers, an experienced teacher of both geography and history, and an experienced teacher of social science subjects who was also a teacher trainer in economics. The teacher trainers were working at several university teacher training institutes. All raters received financial compensation for participating. None of the raters had assessed teacher portfolios before participating in our project.

6.3.2 Sample of teachers

From a random sample of 115 upper secondary schools in the Netherlands (20% of the population) we approached the heads of the economics, geography, and history departments with

information about the study and a request for participation. Twenty-one teachers from twenty-one schools were willing to participate: 10 history, 7 economics, 4 geography. This low response can be explained by the demanding character of the study in terms of the required motivation, time, and experience in teaching students research skills.

6.3.3 Teachers' portfolio preparation

Teachers assembled the following artefacts during a period of a few months (cf. Appendix 2).

1. Self-description, including a description of the teacher's experience and vision concerning teaching research skills.
2. A series of research assignments given to the students during the upper-secondary education years.
3. The results of two interviews concerning teachers' practical knowledge and intentions regarding instructing and coaching students research skills.
4. Two video recordings of lessons in which the teacher instructs and coaches students doing research.
5. A research assignment that is central to the portfolio, including the learning goals of the assignment and the teacher's reasons for the assignment content and form.
6. The assessments of students' work by the teacher (including the goals, content standards, and scoring used).
7. Teachers' reflections on their weaknesses and strengths and on how to improve their teaching.
8. Student evaluations of the teacher (see Appendix 5).

For illustration purposes, the teachers also added examples of students' work.
We strengthened the generally weaker points of the portfolio instrument (low reliability and laboriousness) by specification of the portfolio content and context. All teachers who participated in this study received an extensive report containing the judgments given (the scores to be described below) and also concrete feedback summarizing the contents of the completed judgment forms (see also below).

6.3.4 Rater training procedure

The raters studied a manual that thoroughly described the objectives, planning and procedures of the assessment, the content standards, the anchor points, and the portfolio materials. The raters first individually studied an example portfolio. They then participated in a training session (four hours plenary) during which each rater independently practised one analytic and one holistic assessment. The raters expressed their decisions on a judgment form on which they were encouraged to write comments. After individually scoring the portfolios, the group discussed their analytic and holistic assessments, with a focus on the arguments underlying the judgments given and their evidential base in the portfolio artefacts. The raters were taught how to use the rating format and to explain their arguments.
After the training, the raters individually rated three portfolios in a pilot (one randomly selected portfolio per subject), intended as a try-out of the rating procedure and to give the raters feedback in order to improve their judgments. In addition, we gathered judgment forms and retrospective verbal protocol data. We used these data to develop a coding scheme to analyse verbal protocols and judgment forms (see below). After the pilot, the raters received

feedback about the reliability of their scoring and suggestions for improving their judgments. Suggestions included the accurate use of the standards and the assessment procedure, points of special interest concerning the interpretation of and distinguishing between the standards, and argumentations of the judgments given relating the judgments to concrete examples of portfolio material.

6.3.5 Instrumentation

We used three instruments to capture raters' cognitive representations: judgment forms, retrospective verbal protocols, and retrospective open-ended interviews.

Judgment forms. Judgment forms were designed according to the Correspondent Inference Theory (Jones et al., 1965). Firstly, raters illustrated every content standard with significant portfolio material. Secondly, they described their interpretations of the material. Finally, raters assigned a score to each content standard, using a 5-point scale with anchor points (see Appendix 1) (hereafter: analytic scores). Raters added an overall judgment on a 5-point scale, considering the total of evidence from the separate scores on the content standards (hereafter: holistic score). Finally, for each portfolio, the raters reported whether they followed the prescribed procedure.

Retrospective verbal protocols. Raters explicated their thoughts in verbal protocols (Ericsson & Simon, 1984). Verbal protocols can be used concurrently (during the performance of a judgment task) or retrospectively (after the completion of the task). The first demonstrates the verbalization of raters' cognitive processes in short-term memory (thinking aloud) and the latter reveals mental representations that are stored in long-term memory (Ericsson & Simon, 1984). We used retrospective protocols by which raters recalled their thinking within two weeks after rating the portfolios, using their completed judgment forms as memory cues. We used retrospective protocols because portfolio assessment is time consuming (on average, four hours per portfolio) and mentally demanding, so concurrent verbalization would be too challenging.

Retrospective open-ended interviews. Following the verbal protocol sessions, we conducted open interview sessions about raters' judgment approaches. The interviews focused on the procedure followed while judging the portfolios, the (cognitive) steps taken to build an overall impression of a teacher resulting in a holistic judgment score, pitfalls encountered during the assessment process, and the extent to which the judgments were target-referenced.

6.3.6 Coding scheme

We used the pilot that was part of the training also to check for the quality of the scoring procedures and to develop a coding scheme to describe raters' cognitive representations. In the pilot, the six raters individually judged three portfolios (one per subject). In the judgment forms, all raters indicated that they followed the prescribed judgment procedure. The jury alphas of the analytic scores were .75, .44 and .63. In 50% of all holistic scores, the assessors showed exact total agreement, 33% of the scores showed a difference of 0.5 point, and 17% showed a difference of one or more points. These results are comparable to those obtained by LeMahieu, Gitomer, and Eresh (1995), who reported exact agreement among teachers' scores rating student portfolios ranging from 46% to 57%.

The eight content standards formed a consistent scale (Cronbach's alpha .79), which was a prerequisite for calculating unweighted and weighted mean analytic scores. An Anova did show slight rater variance. See *Table 6.1*.

Table 6.1 One-way analyses of variance between teachers and between raters on holistic scores (h) and weighted mean analytic scores (wa) (pilot, 3 portfolios)

Effects		df	Sum of squares	Mean square	F (p)
Teachers	h	2	2.028	1.014	4.244 (.03)
	wa	2	4.819	2.409	19.146 (.00)
Raters	h	5	.944	.189	.486 (.78)
	wa	5	.432	.086	.165 (.97)

During the pilot, all six assessors retrospectively verbalized their judgments of one portfolio (from a history teacher) during a two-hour session with the researcher. Based on Ericsson et al. (1984), the raters were given general instructions to continue verbalizing while reconstructing their rating processes. The six protocols were taken at the raters' schools or private homes and were tape-recorded and fully transcribed. The raters were allowed breaks as needed. Despite the design of the judgment forms in accordance with the Correspondent Inference Theory, raters' judgment processes as retrospectively verbalized appeared to be associative, in that they skipped from one topic to another. We decided to separate segments per portfolio material per content standard (i.e., per cell in Appendix 2). The intercoder agreement (Cohen's kappa) of the segmentation between two coders (the first author and an independent researcher) of three randomly chosen verbal protocols (three out of the six available) was .89. This result corresponds with the range of the reliabilities of segmentation between .80 and .90 as discussed in Ericsson et al. (1984).

The coding of the segments was primarily based on Associated Systems Theory. We used the codes visual appearance, categorization and personal trait to describe utterances on the dimension concrete versus abstract. We also accounted for the type of response given on the dimension positive versus negative (Huot, 1993). In accordance with the Correspondent Inference Theory, as a third category we encoded verbalizations about the teacher's situation. To check this initial coding scheme, the two coders coded the three verbal protocols from the pilot (the same two coders and three randomly chosen protocols as mentioned above). Discussions between the two coders resulted in some adjustments to the initial scheme. To accurately capture raters' individual representations on the dimensions concrete-abstract and positive-negative we decided to encode the protocols per line. In addition, we added two more categories: rater's own judgment process and judgment procedure. Finally, to facilitate reliable coding, we developed discourse markers for every category. In sum, we coded on the dimensions concrete-abstract and positive-negative per line, and on three other categories, situation in which teachers operate, raters' own judgment process, and judgment procedure per segment (see Appendix 6).

The five categories included in the final coding system could satisfactorily be distinguished in the verbal protocols and the judgment forms (see *Table 6.2*). With one exception, the kappas were above .65, which is acceptable for intercoder agreement (Popping, 1983).

To obtain scores on the dimensions abstract–concrete (visual appearance, categorization, attributed personality traits, coded as 1, 2, 3) and positive–negative (positive, neutral, negative, also coded as 1, 2, 3) we averaged the rater's positions on both dimensions and then averaged the scores across the protocol lines per portfolio per rater. The frequencies on the other three categories (teachers' situations, raters' judgment processes, and judgment procedures) were counted per portfolio per rater per subcategory.

Table 6.2 Cohen's Kappas of categorizations based on the retrospective verbal protocols and judgment forms

Category of comments	Verbal protocols	Judgment forms
Dimension abstract–concrete (encoding per line)	.71	.65
Dimension positive–negative (encoding per line)	.66	.79
Situation in which teachers operate [a]	.75	1.0
Raters' own judgment process (metacognitive) [b]	.69	.48
Judgment procedure [c]	.65	.80

NOTES

a Subcategories: own background, method used, students' characteristics, time available, teacher's experience, school policy, expectations.

b Subcategories: references to the manual while rating; notifications about lack of information for giving an adequate assessment; explaining own judgment process; pitfalls and uncertainties while rating; questioning own assessment process; notifications and corrections of mistakes; references to earlier assessed portfolio material; references to earlier assessed portfolios.

c Subcategories: judgment criteria and standards, portfolio material, assessment procedure,weighting of the content standards.

6.3.7 Data collection

After having used three portfolios in the pilot as described above, the remaining 18 portfolios (9 history, 6 economics, 3 geography) were rated, each one independently by two raters. We organized the rater pairs according to their major expertise. As most of the raters had expertise in several subjects, the composition of the rater pairs could vary according to availability of the raters. In a 9-month period, we assigned the 18 portfolios in the order they became available. Rater 1 judged six portfolios (mainly economics), Rater 2 four portfolios (history), Rater 3 nine portfolios (mainly history), Rater 4 five portfolios (geography and history), Rater 5 seven portfolios (geography, history and economics), and Rater 6 five portfolios (mainly economics). We aimed at using fixed couples per subject, but due to some availability problems we had to allow for a few departures from this scheme.

All raters retrospectively verbalized their rating process of two randomly chosen portfolios (after scoring) from the portfolios they had judged. The protocols were fully transcribed, resulting in 7216 lines containing 310 segments (split per content standard per portfolio, according to the cells in Appendix 2). The independent coder (the one who earlier showed intercoder agreement with the researcher) coded the verbal protocols of 12 portfolios (two portfolios per rater) and the 36 judgment forms of all 18 portfolios, using the coding scheme developed earlier.

6.3.8 Data analysis

We describe the analyses concerning the reliability of raters' judgments, raters' cognitive representations, and the relationships among raters' cognitive representations, their judgments, and the reliability of these judgments according to the three research questions.

6.3.8.1 The reliability of the raters' judgments

The rating produced three types of score: unweighted mean analytic, weighted mean analytic, and holistic. The unweighted mean analytic scores focused on the mean of the scores on the

eight content standards. The weighted analytic scores used the weights per content standard to calculate total scores (see chapter 8). The holistic scores were overall judgments considering the total of evidence from the separate scores on the content standards.
Firstly, we analysed the reliability of the eight content standards as a set by computing Cronbach's alpha. Secondly, we analysed the extent of agreement in holistic judgments within rater pairs. As an additional indication of the consistency of raters' judgments, we verified whether there were statistically significant differences between the means of the three score types: unweighted mean analytic, weighted mean analytic, and holistic.
Thirdly, we examined the interrater reliability for the separate analytic scores by calculating the Cronbach's alphas (jury alphas). The accuracy of the ratings, that is the correspondence between raters' judgments and the performance theory as taught in the training, was checked by coding the content of raters' judgments and comparing the results with the full descriptions of the standards in our study.
Finally, we conducted variance analyses on the unweighted and the weighted mean analytic scores and the holistic scores to estimate the variance related to the raters.

6.3.8.2 Raters' cognitive representations

Data from the retrospective interviews were qualitatively analysed. The coded data from the verbal protocols and judgment forms were analysed in a descriptive way (frequencies and means) and the differences between raters were analysed using one-way analysis of variance.

6.3.8.3 Relationships between raters' cognitive representations, their judgments and the reliability of these judgments

The assumption that judgments are a function of cognitive representations entails the idea that raters who differ in their comments should also differ in their judgments. We used linear multiple regression analysis (method enter) to explore how well the scores on the categories of the comments given in the judgment forms and the retrospective verbal protocols predicted the unweighted mean analytic, weighted mean analytic, and holistic scores. In the analyses of the judgment forms we eliminated the category rater's own judgment process, because this category was not satisfactorily distinguished during the coding process (see *Table 6.2*). To indicate the quality of the prediction, we used the proportion of explained variance (R^2). For linear multiple regression analyses to be valid, assumptions of normality of residuals, homoscedasticity, linearity, and absence of collineation have to be fulfilled. Therefore we checked for these assumptions.
Finally, to check whether the raters' cognitive representations possibly influenced the reliability of their portfolio judgments, we calculated the Pearson correlations between the differences in the category scores within rater pairs and the jury alphas.

6.4 Results

6.4.1 The reliability of the raters' judgments

The eight content standards formed a reliable scale: the Cronbach's alpha was .76. To 35% of the judgments the raters gave exactly the same holistic score on the 5-point scale, a difference of half a point showed up in 12% of the judgments, a difference of one point in 47% of the judgments, and a difference of 1.5 points in 6% of the judgments (one rater pair).

The jury-alphas for the separate analytic scores were satisfactory for 12 rater pairs, ranging from .39 to .76. The jury-alphas were low or even negative for six pairs, ranging from -.80 for one pair and ranging from -11 to .22 for five pairs. We compared the content of raters' judgment forms to the full descriptions of the content standards. In all ratings, the raters referred sufficiently to the content standards according to the training.
To check the agreement between the three types of scores, we calculated correlations and conducted paired-samples T-tests (see *Table 6.3*). The results showed that the score types were highly correlated. However, the mean of the holistic scores (3.12; sd .78) is significant lower than the mean of the weighted mean analytic scores (3.60; sd .96), while the former corresponds closely with the mean of the unweighted mean analytic scores (3.11; sd .59).
To estimate whether the scoring variability was due to rater effects, we conducted one-way analyses of variance (see *Table 6.4*). The results showed slight rater effects. Due to the structure of the available data we could not check for possible interaction effects between raters and teachers.

6.4.2 Raters' cognitive representations

6.4.2.1 Retrospective interviews

In the retrospective interviews, the raters said they used the rating procedure according to their training. They also said they found the procedure useful. They spent an average of four hours per portfolio to produce all judgments and ratings.
Five raters (1, 3, 4, 5, 6) primarily strove to build a coherent image of the teacher based on

Table 6.3 Pearson correlations and paired samples T-tests for holistic (h), weighted mean analytic (wa), and unweighted mean analytic (ma) scores (18 portfolios)

	First mean	Second mean	r	Difference (sd)	df	T (p)
h – wa	3.12 (h)	3.60 (wa)	.85 *	–.44 (.50)	35	–5.28 (.00)
h – ma	3.12 (h)	3.11 (ma)	.85 *	.04 (.42)	35	0.59 (.56)
wa – ma	3.60 (wa)	3.11 (ma)	.99 *	.49 (.38)	35	7.59 (.00)

* $p < .00$

Table 6.4 One-way analyses of variance between teachers and between raters on holistic (h), weighted mean analytic scores (wa), and unweighted mean analytic (ma) scores (18 portfolios)

Effects		df	Sum of squares	Mean square	F(p)
Teachers	h	17	12.90	.76	1.54 (.17)
	wa	17	19.64	1.16	1.67 (.14)
	ma	17	7.40	.44	1.62 (.16)
Raters	h	5	.97	.19	.29 (.92)
	wa	5	2.04	.41	.41 (.84)
	ma	5	1.07	.21	.58 (.72)

the portfolio material. They subsequently explained this image using concrete examples. In building an image, they perceived the videos as very helpful. Four raters firstly watched the videos and made notes (filtering out relevant data). Then they went through the portfolio. Finally, they filled out the judgment forms and went through the portfolio per content standard for a second time. In an alternative process, two raters directly started filling out the judgment form. They went through the videos and portfolios only once. T-tests between these two raters and the other raters showed that these variations in procedure did not significantly affect the rating outcomes.
Some raters discovered that previously rated portfolio material and content standards influenced the rating process of later portfolio material and content standards. Rater 1 was aware of this process: "I should not use the self-description to assess content standard 3 (MANAGE), but sometimes it stays in my head and it influences my rating, although it should not". Most of the raters said that to control for this, they corrected this process afterwards. Rater 4: "In rating teachers, I have a tendency to follow my intuition. Since I know that is not correct, I force myself to follow the procedure as trained conscientiously". Rater 5 explained the need to be aware of the danger of looking for episodes in the portfolio that confirmed his preconceptions. "Every time I correct myself for this mechanism by saying to myself, "Watch out, you are doing it again". Rater 2: "A few days after assessing a portfolio, I re-read my assessment to control if it really justifies the portfolio".
The images built were also influenced by the raters' own experiences. Four raters (1, 3, 5, 6) confirmed that they based their judgments partly on their own knowledge about teachers. Rater 3: "You know what is realistic to expect from teachers. Carrying your experiences with you colours the judgments given".

6.4.2.2 Judgment forms and retrospective verbal protocols

Table 6.5 shows that all raters used a mix of concrete and abstract representations in their comments. On average, the verbal protocols elicited more comments from the raters than the judgment forms, especially on the dimension concrete–abstract. In general, comments made in the verbal protocols were less concrete than comments in the judgment forms. Furthermore, the raters differed in the number of comments they made per portfolio.
One-way analyses of variance of raters' mean scores on the dimensions concrete–abstract and positive–negative in judgment forms and verbal protocols showed that the raters differed in their mean scores. The results on the dimension concrete–abstract were $F = 4.19$, $p < .001$ (judgment forms), $F = 0.06$, $p < .001$ (verbal protocols). On the dimension positive–negative, the results were $F = 6.24$, $p < .001$ (judgment forms), $F = 0.04$, $p < .001$ (verbal protocols). Other differences in their comments were not significant.

6.4.3 Relationships between raters' cognitive representations, their judgments and the reliability of these judgments

To explore whether the categories of comments could explain the judgments given, we conducted linear multiple regression analysis, using raters' unweighted mean analytic, weighted mean analytic, and holistic scores as the content standard variables and the scores on the categories of cognitive representations as predictors. For all three score types the assump-

Table 6.5 Means and frequencies of comments per rater per dimension in verbal protocols and judgment forms

Sources	Concrete Abstract			Positive Negative			Situation teacher	Own judgment	Judgment procedure
Rater	n	m	sd	n	m	sd	n	n	n	n
1										
Verbal protocol	2	1.74	.70	530	2.69	.67	62	42	14	108
Judgment form	6	1.54	.57	357	1.98	.98	56	10	49	7
2										
Verbal protocol	2	1.96	.60	138	2.31	.75	20	1	0	13
Judgment form	4	1.51	.54	219	2.13	.95	24	6	30	4
3										
Verbal protocol	2	1.89	.66	1079	2.34	.84	128	50	29	104
Judgment form	9	1.44	.53	1456	1.77	.94	176	17	72	19
4										
Verbal protocol	2	1.91	.68	554	1.86	.89	94	26	33	80
Judgment form	5	1.55	.56	302	2.32	.88	78	6	11	7
5										
Verbal protocol	2	1.95	.60	594	1.97	.90	68	37	27	87
Judgment form	7	1.45	.51	504	1.74	.97	76	14	26	2
6										
Verbal protocol	2	1.94	.64	613	2.47	.80	83	31	23	79
Judgment form	5	1.47	.53	852	2.16	.91	156	9	22	1

tions of normality, homoscedasticity and linearity were sufficiently met. The categories of cognitive representations based on the retrospective verbal protocols (n = 12, concerning two randomly chosen portfolios per rater) did not show significant relations with the judgment scores, but those based on the judgment forms (n = 36) did. The explained variances of the unweighted mean analytic scores (R^2 = .65, F = 2.59, df = 13, p = .03), the weighted mean analytic scores (R^2 = .63, F = 2.68, df = 13, p = .03), and the holistic scores (R^2 = .59, F = 2.36, df = 13, p = .04) were quite similar.

Linear multiple regression analysis of the unweighted mean analytic scores on the categories of comments in the judgment forms (n = 36) showed that 65% of the variance in the ratings could be explained by the category scores. The category scores were barely collineated (the correlations were between – .22 and +.31 and mostly near r = .00). The standardized partial beta coefficients showed that the comments made on the dimensions concrete–abstract (beta = .35; T = 2.13; p = .05) and positive–negative (beta = .71; T = 4.4; p = .00) contributed significantly to the prediction of the portfolio judgments given. The other categories (concerning the teacher's situation and the judgment procedure) did not make a statistically significant contribution to the prediction.

Since the scores on the cognitive representation dimensions concrete–abstract and positive–negative significantly contributed to the prediction of the judgments given, we also checked their relationship with the interrater reliability of these judgments. Per portfolio (n = 18), we calculated the difference scores between the two raters on both dimensions and then we correlated these difference scores with the jury alphas. The results showed that the closer the correspondence within the rater pairs' comments in the judgment forms as categorized on the two dimensions, the higher their interrater reliability (r = .44, p = .09 for the dimension concrete–abstract; r = .53, p = .04 for the dimension positive–negative).

6.5 Conclusions

We researched into the questions: What is the reliability of teacher portfolio assessments? How can raters' cognitive representations be described? What is the relationship between raters' cognitive representations, their judgments and the reliability of these judgments? We used a mixed quantitative and qualitative approach to study the reliability of judgments, raters' cognitive representations (using the Correspondent Inference Theory (Jones & Davis, 1965) and the Associated Systems Theory (Carlston, 1992, 1994)), and the relationships between cognitive representations, judgments and reliability.

Six raters systematically assessed 18 portfolios. The interrater reliability of the analytic scores of 12 portfolios was satisfactory. Variance analyses showed only slight rater effects. Raters' holistic scores in all but one case differed by no more than one point (on a 5-point scale).

The raters used cognitive representations on the dimensions concrete-abstract (e.g. visual manifestations) and positive–negative (positive and negative comments), the situation in which teachers operate, their own judgment process, and the judgment procedure.

Although we used the judgment forms completed during the judgment sessions as cues for producing the retrospective verbal protocols, the comments in the verbal protocols differed markedly from those in the judgment forms. The comments made in the verbal protocols were generally less concrete than those in the judgment forms. The cognitive representations in the judgment forms were significantly related to the judgments given. The differences in cognitive representations within pairs were significantly related to the interrater reliabilities.

What is the meaning of these results? Firstly, several researchers have argued that the complex and context-bound character of portfolio assessments asks for the use of a different reliability standard than used for standardized forms of assessment in which interrater reliability coefficients often exceed .90 (Nunnally, 1978). After all, many variables that are fixed in standardized tests are variable in portfolio assessments. Koretz, McCaffrey & Stecher (1993) assert that interrater reliability coefficients of .80 or higher are often regarded as reasonably strong for performance assessments. Gentile (1992) reports that coefficients above .80 strong, and above .65 are considered good for portfolio studies. We could only partly meet these content standards. Further research should expose the coefficients attainable for portfolio assessments.

Secondly, the raters based their holistic judgments of the portfolios on the unweighted means of the analytic scores given. This is obvious from the correspondence between these two score types. This tendency could be considered as undesirable, however, since it is possible that a content standard is more important than another (and thus should get more weight). In one of our previous studies (Van der Schaaf et al., 2003) we let the raters formulate a panel policy in which they decided that some content standards should be weighted more heavily than others. The tendency to use unweighted means is strong, however, as the holistic scores and the unweighted mean analytic scores corresponded highly although the raters were instructed and trained in weighting the scores according to the policy.

Thirdly, the retrospective interviews revealed that the raters strove to build a coherent image of the teacher they were assessing. The raters' experience was that the images were influenced by previously rated portfolio material and content standards, as well as raters' own experiences as a teacher. These results are comparable with those of other studies in diverse domains. Zwaan and Brown (1996), for example, found that skilled readers generate explanations to combine data across sentences, building a coherent mental representation.

Fourthly, a prerequisite for valid assessments is that raters base their judgments on teachers' portfolio content. Rating quality is improved by raters using target-referenced representations, relating judgments primarily to teachers' competences rather than to raters' subjective, self-referenced perceptions. Forming concrete representations enhances this process. On the other hand, abstract representations are needed to predict a teacher's performance in situations other than those described in the portfolio. Ideally then, raters should alternately use concrete and abstract representations.

Fifthly, the judgments given could significantly be predicted by the raters' representations as categorized on the basis of their comments in the judgment forms on the dimensions concrete–abstract and positive–negative. These representations explained a high 65% of the variance in the portfolio ratings. Furthermore, the results showed that the higher the correspondence within rater pairs in these representations, the higher the interrater reliability. These dimensions of raters' cognitive representations as revealed in our study could therefore also explain the reliability of the assessment.

Since portfolios are personal and context-bound instruments, portfolio assessment is cognitively demanding and complex, even after training. This affects the reliability of portfolio assessments (expressed in the interrater reliability and rater effects). Although our study showed that certain cognitive representations may contribute to reliable portfolio assessment, it also shows that raters differ in their representations, which seemingly reflects different interpretations of the portfolios, regardless of the agreement in the ratings given.

These results give way to the question whether consensus approaches would be preferable as an assessment strategy and whether such approaches could replace independent ratings. The value of consensus approaches depends on the importance of interrater consistency in portfolio assessment. For summative assessment purposes—e.g. certification or merit pay—portfolios contain a sample of "best work" and high interrater reliability will be necessary to make adequate decisions. However, for formative purposes and using portfolios that contain broad and various samples of work, it can be questioned whether it is likely that two independent raters will arrive at identical judgments. In this case, it might be more important that raters agree on the consequences of the judgment given, rather than on the actual scores achieved. Delandshere and Petrosky (1994) proposed a procedure using confirmation of prior judgments (similar to getting a second opinion from a physician) rather than replications. That such a confirmation approach might be useful for portfolio assessment was also concluded by Linn (1994), referring to the work of Moss (1994) on validity versus reliability and illustrating how committees examining the qualifications of candidates can arrive at an integrated decision. Further research should demonstrate the consequences and practical utility of confirmation approaches.

Portfolio assessments are contextually embedded. Our study revealed that external raters involve situational information (e.g. student characteristics and school policy) in their rating processes. They may partly color this information with their own teaching experiences. Internal raters, working in the same educational organization as the teachers being assessed, could ensure more validity and consistency in their ratings. It is often suggested that individuals from similar backgrounds are likely to reach agreement more readily (Pula & Huot, 1993). For that reason, portfolio assessment may be more suitable for internal rating processes (Linn, 1994). However, internal ratings of teacher portfolios will probably echo subjective impressions and personal relationships between the rater and the teacher assessed. As a result, the ratings will be more idiosyncratic and this will have its own negative effects on the rating

reliability. It can be expected that internal ratings of teacher portfolios will be more self-referenced than those of external raters (Carlston, 1994). To increase the accuracy of internal ratings, rater training should pay attention to the importance of target-referenced rating. Future research should examine such possibilities.

7 Teacher beliefs and teacher behaviour in portfolio assessment

This chapter is submitted as: Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (2005). Teacher beliefs and teacher behaviour in portfolio assessment. Manuscript submitted for publication

Abstract

What is the correspondence between teacher beliefs and behaviour as constructed by teachers in their portfolios and as assessed by their students and by external raters? We qualitatively analysed the beliefs and behaviour of 18 teachers as described in their portfolios. In addition, each portfolio was independently assessed by two trained raters on eight content standards and the teachers' classroom behaviour was assessed by their own students in a questionnaire (n = 317). Linear multilevel analysis showed that the students' assessments of their teachers' classroom behaviour could be significantly predicted by the raters' assessments of the teachers' beliefs and behaviour as described in their portfolios. However, the raters' assessments themselves could not be significantly predicted from the beliefs and behaviour as described in the portfolios. We searched for differences between the teachers by using exploratory Q-principal component analysis on those raters' assessments that had a reasonable to good degree of inter rater reliability and we tested the two groups of teachers found for differences in the students' assessments of their classroom behaviour. Teachers with high raters' assessments on the content standard THINK ("the choice of and argumentation for teaching strategies that meet students' knowledge, abilities and experience") had significantly higher student assessments than teachers who were judged low on this content standard. Implications of the results and suggestions for further research are discussed.

7.1 Introduction

Many countries face a growing interest in the assessment of teachers. This interest is fostered by current needs for accountability and quality improvement in the teaching profession (Cochran-Smith & Fries, 2001). Quality teaching is now recognized to be a result of complex interactions between teachers' beliefs and behaviour in a certain context and teachers' thoughtful examination of these interactions. It is generally assumed that only by becoming aware of their beliefs, teachers can further develop their repertoire (Borko & Putnam, 1996; Shulman, 1996; Cochran-Smith & Lytle, 1999). Saroyan & Amundsen (2001) even argue that the most competent teachers conscientiously try to align their beliefs with their behaviour in an attempt to attain specific instructional goals.

Since teacher beliefs appear to be crucial for the improvement of teaching behaviour, it is nowadays assumed that teacher assessment should take account of teacher beliefs in relation to their behaviour (Beijaard & Verloop, 1996). Accordingly, teacher assessment is now often based on instruments that focus on an inclusive view of teaching, for example portfolios (Andrews & Barnes, 1990; Delandshere, 1994). Depending on the content and form of the portfolio and on the integration of the portfolio with a teacher's working context, a portfolio may do justice to the fact that teaching is a complex activity and that teacher behaviour is inextricably bound up with teacher cognition and the teaching context (Bird, 1990; Lyons, 1998). Portfolio assessments can be used summatively, e.g. as a tool to ascertain whether teachers satisfy required standards and formatively, to formulate guidelines for professional development.
Although the existence of a connection between teacher beliefs and behaviour has been well established, the role of teacher beliefs in teacher portfolio assessment is far from clear. This is due to the fact that teacher beliefs are mostly held unconsciously and are context dependent, so they hardly allow for accurate assessment. Further, teachers' beliefs as constructed in their portfolios do not necessarily correspond to teachers' beliefs as assessed by the raters.
Research into the correspondence between teachers' beliefs and behaviour in teacher portfolio assessment is important to deepen our understanding of this correspondence, to increase our knowledge about what can and cannot be assessed by teacher portfolios, and to improve the validity of portfolio assessments. Therefore, in this paper we try to give some answers to the question: What is the correspondence between teacher beliefs and behaviour as constructed by teachers in their portfolios and as assessed by their students and by external raters?
We research into this question by comparing plain portfolio data (including self descriptions and video registrations of teachers' lessons) with assessments by students and raters. The study is part of a larger research project on the assessment of experienced Dutch teachers' beliefs and behaviour and builds upon earlier studies in which we developed standards (both content standards and performance standards) for teaching research skills in pre-university social science education (Van der Schaaf, Stokking & Verloop, 2003, 2005c). Our research project is focused on a topic that is representative of the change in many countries towards more constructivist views on learning and that emphasizes developing students' skills in an active, self-regulated and collaborative way, namely teaching students research skills.

7.2 Teaching students research skills

Learning to do research implies the development of cognitive abilities to investigate phenomena in a subject domain (Wiggins, 1993; Duggan & Gott, 1995). Carrying out research requires subject-specific knowledge, more general conceptual, procedural and contextual knowledge and skills, and certain attitudes (including careful execution, proper processing of data, consistent reasoning, and critical evaluation) (Lee & Fradd, 1998; Stokking & Van der Schaaf, 2000).
In the Dutch exam requirements research is seen as consisting of a series of activities or steps and the following steps can be distinguished: 1. identify and formulate a problem using subject-specific concepts; 2. formulate the research question(s); 3. make and monitor the research plan; 4. gather and select data; 5. assess the value and utility of the data; 6. analyse the data; 7. draw conclusions; 8. evaluate the research; 9. develop and substantiate a personal point of view; 10. report and present the research (Stokking & Van der Schaaf, 2000).

A distinction can be made between skill as cognitive capability and skill as task-oriented performance. If research skills are conceived as representing the cognitive capabilities necessary to carry out research and research is seen as consisting of a series of steps, it would be convenient if these steps would represent the skills. However, different steps can assume the same cognitive activities and for any given step more than one cognitive activity may be necessary (Lock, 1989, 1990a; Brown, Moore, Silkstone & Botton, 1996). Therefore, it is questionable whether conceiving research as consisting of a series of steps does enough justice to the conceptual, procedural and contextual knowledge needed for doing and learning to do research (Kirschner, 1992; Gott & Duggan, 1996; White & Frederiksen, 1998; Smits, 2003). Obviously, this does not alter the fact that it can be useful to train certain sub skills.
In teaching students research skills, teachers will have to decide how much they structure the students' learning environment through assignments and guidance and how much they leave to the students themselves. Students in pre-university education have little experience in doing research and in thinking scientifically. A very structured learning environment in which the students almost have no space for making choices is not very obvious (Roth & Roychodhure, 1993). The same goes for a very open learning environment. Students learn best by 'guided' reinvention (not too structured, not too open), because both extremes lead to superficial learning (Kanselaar, Galen, Beemer, Erkens & Gravemeijer, 1999). Additionally, the teacher can give the students opportunities to cooperate, which motivate the students and further their learning process. Cooperation can also further students' learning directly if it triggers discussion about and reflection on the research assignment and the research process.
To teach students research skills teachers have to fulfil professional tasks. These tasks include: pre-active tasks (tasks before teaching, e.g. goal setting), interactive tasks (tasks during teaching, e.g. instructing, coaching), and post-active tasks (tasks after teaching, e.g. assessing and reflecting on own teaching) (Reynolds, 1992; Van der Schaaf, Stokking & Verloop, 2005c). In this study we focus on the interactive tasks.

7.3 Correspondence between beliefs and behaviour in teacher assessment

7.3.1 Teacher beliefs and teacher behaviour

Teacher beliefs are important mediators of teacher behaviour (Calderhead, 1996; Cohen, 1987). However, the relation between teacher beliefs and teacher behaviour is far from clear. This is due to the fact that teacher beliefs are messy constructs with different interpretations and meanings (Kagan, 1992; Pajares, 1992). We nevertheless found some convergence of ideas about the nature of teacher beliefs in literature.
Firstly, current definitions of teacher beliefs focus on teachers' assumptions which affect what they notice in any set of circumstances and what they regard as possible, the goals they will set, and the knowledge they will bring into those circumstances (Clark & Peterson, 1986; Kagan, 1992; Pajares, 1992). Since teacher beliefs shape the way teachers perceive and interpret classroom interaction and influence their construction of intentions in response to those interactions, we define teacher beliefs as: "an integrated system of personalized assumptions about the nature of a subject, its teaching and learning" (Artzt & Armour-Thomas, 1998, p. 8).
Secondly, it is generally assumed that teacher beliefs differ in specificity and in strength depending on the context (as perceived), that they tend to be activated in clusters, and that in-

compatible beliefs may contend for priority (Schoenfeld, 1998; Aguirre & Speer, 2000; Ajzen, 2002). This implies that not all teachers' beliefs will play a role in their behaviour. Only the most salient beliefs will influence the execution of teaching tasks.
Thirdly, beliefs may guide teacher behaviour either deliberately or spontaneously. In the deliberate way beliefs are retrieved or constructed in an effortful manner in a certain context and they are assumed to guide goal formation and behaviour. Teacher goals are "cognitive constructs that describe at various levels of detail what the teachers wants to accomplish" (Aguirre & Speer, 2000, p. 332). Goals can be conceived of as long-range goals for student learning (Clark & Peterson, 1986) and as short-term mental structures that arise in interaction with events in the classroom (Saxe, 1991). For the purpose of this study we designate the former as goals and the latter as intentions to engage in certain behaviour. Fazio (1990) argues that under conditions of high motivation and sufficient cognitive ability, people can put an effort in constructing their beliefs related to certain goals and intentions. Without those two conditions beliefs related to a goal or intention only occur when they are routinely activated. Routine activation of such beliefs only takes place when the goal or intention raises a strongly tied attitude toward (or evaluation of) that goal or intention. Therefore, strong beliefs are generally more resistant to change than weak beliefs.

7.3.2 Constructed teacher beliefs and behaviour

As the nature of teacher beliefs hardly enables accurate assessment, it is not possible to show directly the correspondence between teachers' beliefs and behaviour. The clusters of beliefs that guide teacher behaviour are context-related and exist predominantly as tacit knowledge that cannot be easily articulated. Implicit beliefs can only be studied once they are first made explicit and we assume that teacher beliefs as described in their portfolios are constructed representations of their beliefs, that is the beliefs that teachers claim to have themselves.

7.3.3 Assessed teacher beliefs and behaviour

Based on judgment models developed in social cognitive psychology (Jones & Davis, 1965; Carlston, 1992, 1994), Van der Schaaf, Stokking & Verloop (2005a, 2005b) showed that raters use schemata to judge, foresee and comprehend the beliefs that teachers construct (see chapter 6). These schemata are comparable to personal constructs (Kelly, 1955), content categories used to organize and simplify information. Interpersonal filters in these schemata induce raters to look selectively at the information about the teachers and to interpret the information according to their own constructs. As a result, raters' representations about teachers' beliefs are not necessarily consistent with the ones that teachers hold themselves.
On the other hand, raters' mental representations are shaped by their prior knowledge and experience, e.g. when they have knowledge of the working context of the teachers assessed (internal raters, working in the same educational organization as the teachers), have teaching experience themselves, and have experience in scoring portfolios. Research showed that it is possible to influence rater judgment by certain training procedures (Lievens, 2001; Woehr & Huffcutt, 1994). Depending on the form and content of the training, discrepancies between raters' representations and teachers' constructed beliefs may decrease.
Irrespective of the amount of discrepancy between raters' representations and teachers' constructions, the raters' representations are worthwhile in themselves as they enable raters to explain the meaning of teachers' behaviour. When a rater says that a teacher has a certain belief this means that the teacher is behaving in a manner consistent with having such a belief. Therefore, we distinguish between (teacher) constructed and (rater) assessed teacher beliefs.

7.4 Method

7.4.1 Participants

Sample of teachers. Within a sample of 115 upper secondary schools in the Netherlands (20% of the population) we approached the heads of the economics, geography, and history departments with information about the study and a request for participation. Twenty-one teachers from twenty-one schools were willing to participate. This low response can be explained by the demanding character of the study in terms of the required motivation, time, and experience in teaching students research skills. Three teachers participated in a pilot study. Eighteen teachers participated in the main study on which we report in this paper (9 history, 6 economics, 3 geography). The average age of the eighteen teachers was 44 years and they had an average of 15 years teaching experience.
Sample of students. Each participating teacher randomly selected a classroom of which the students rated their teacher in a questionnaire. In total 317 students were involved (on average 18.6 students per teacher). The number of students varied per teacher, depending on how many students had chosen the teacher's subject for examination.
Sample of raters. Since having a teaching background themselves helps raters assess teachers' constructed beliefs and behaviour (Pula & Huot, 1993), we selected raters with teaching experience. We chose external raters (not working in the same school) to avoid the biasing of perceptions or judgments on episodic memories. The raters also participated in earlier studies in which content standards, performance standards, and a portfolio assessment procedure for teaching students research skills were developed (Van der Schaaf et al., 2003, 2005c). All raters were teachers in social science subjects. Three of them were also teacher trainers and one of them was also a school principal. All raters received financial compensation for their participation. None of the raters had assessed teacher portfolios before. All raters supported the standards and the assessment procedure used in this study.

7.4.2 Teacher portfolio evidence and standards

The assessment procedure had been developed and empirically tested in an earlier study (Van der Schaaf et al., 2005c, see chapter 5). In their portfolio teachers collected material about teaching students research skills in several phases of the teaching process (pre-active, interactive, and post-active). The portfolios contained at least seven elements (see Figure 5.2), including a number of productions (documents especially prepared for the portfolio), one reproduction (a registration of teacher's regular work captured on behalf of the portfolio), and several artefacts (regular products of teacher's work). The content of the elements focused on parts of the Theory of Planned Behavior (see chapter 5): background factors (e.g. years of teaching experience, subject, class, and courses followed in teaching students research skills), teachers' beliefs in teaching students research skills (e.g. their goals for engaging students in research), subjective norms (e.g. arrangements with colleagues), control beliefs (e.g. perceived facilities for student research, perceived difficulties students have in doing research), and behaviour (e.g. the way teachers instruct and coach their students).
In one previous study we studied teachers' tasks needed to develop students' research skills and the teacher competences (knowledge, skills, and attitudes) needed to perform those tasks. We developed content standards, most of them made up of several indicators, that describe what teachers should know and be able to do in teaching students research skills during the three teaching phases (pre-active, interactive, post-active) (see chapter 3).

7.4.3 Rater training

The raters were thoroughly trained in assessing teacher portfolios. They studied a manual describing in detail the objectives, planning, and procedures of the assessment, the content standards, the anchor points, and the portfolio materials. The raters first individually studied an example portfolio. They then participated in a training session (four hours plenary) in which they used judgment forms (see chapter 6). After the training, the raters individually rated three portfolios in a pilot (one randomly selected portfolio per subject), intended as a try-out of the rating procedure and to give the raters feedback in order to improve their judgments. After the pilot, the raters received feedback about their scoring and suggestions for improving their judgments. Suggestions included the accurate use of the standards, the assessment procedure, and points of special interest concerning the interpretation.

7.4.4 Data collection

7.4.4.1 Constructed teacher beliefs and behaviour

Since the 18 teachers in our study did not have any experience in preparing a portfolio we structured the content of the elements, while teachers were free in the form of the material and also free to add other material (Shapley & Bush, 1999). The teachers received instructions including a description of the aims of the study, the content and performance standards, and the assessment procedure. Teachers were provided help in developing their portfolio when needed, for instance by making videos and by typing spoken texts (Freidus, 1998; Seldin, 1997; Tillema, 1998). All 18 portfolios were constructed within a few months time span. All teachers confirmed that their portfolios represented their beliefs and behaviour (see chapter 5), so we suppose that their portfolios are quite authentic. Afterwards, the teachers received written feedback concerning their weaknesses and strengths in view of the content standards and performance standards and the comments made by the raters. The teachers filled out a questionnaire with questions on a 5-point scale (1 = I strongly disagree; 5 = I strongly agree) to evaluate the relevance of the feedback given, based on raters' assessments.

7.4.4.2 Assessed teacher beliefs and behaviour

Students' assessments. Based on the content standards we developed a questionnaire containing 15 statements on how their teachers instruct and coach research skills (see Appendix 5). The students answered on a four-point scale from 1 'not at all like my teacher' to 4 'very much like my teacher'. We successfully tested the questionnaire in a pilot study with 45 students of 3 teachers (Cronbach's alpha .84).
Raters' assessments. The 18 portfolios were rated, each one independently by two raters. We organized the rater pairs according to their major expertise. In a 9-month period, we assigned the 18 portfolios in the order they became available. The raters used judgment forms to rate the portfolios. Firstly, raters illustrated each content standard with significant portfolio material. Secondly, they described their interpretations of the material. Finally, raters assigned a score on the portfolio for each content standard, using a 5-point scale with anchor points (analytic score) (for an example, see Appendix 1). They added an overall (holistic) judgment on a five-point scale, considering the total of the evidence from the separate scores (holistic score). The rating process produced two types of scores: (unweighted) mean analytic scores focusing on the means of the scores for the eight content standards and holistic scores being overall judgments of the teachers' portfolios.

7.4.5 Data analysis

7.4.5.1 Constructed teacher beliefs and behaviour

Teachers' interactive behaviour was coded from videotapes of two lessons per teacher. The teachers recorded two periods of instruction and coaching of a maximum of 30 minutes each. The videos were fully transcribed. We split every video in segments per research step (according to the ten steps described earlier). The researcher (the author) and an assistant researcher analysed the videos in terms of the content (the ten steps), the structuredness of the learning environment, the setting (e.g. classroom, multimedia centre), and the students' grade. Per segment we counted the sub skills taught and also computed frequencies of the forms of interaction between teacher and students (teacher central, interaction, students central).
As to the structuredness of the learning environment here we focus only on the teacher's behaviour and not on the research assignment or other tools that will also structure students' learning environment. Ratings of the learning environment were done on a scale from 1 (very structured) to 3 (very low structured), based on a detailed description of teacher behaviour related to the structure of a learning environment (Vermunt & Verloop, 1999; Van der Schaaf, 2000). High scores were given when teachers encouraged and gave students opportunities to work autonomously (e.g. by offering students choices and refraining from giving unnecessary help). Low scores were given when the teacher controlled all aspects of research activities.
Data concerning background factors, teachers' goals in teaching students research skills, their subjective norms, and their control beliefs were analysed descriptively (frequencies, means).

7.4.5.2 Assessed teacher beliefs and behaviour

Students' assessments. We analysed the scalability of the content standards in the questionnaire by computing Cronbach's alpha. We calculated frequencies and checked for differences between teachers using oneway Anova on their students' ratings of their classroom behaviour.
Raters' assessments. We used the percentage exact agreement in holistic scoring and the Cronbach's alpha coefficient (jury alpha) in analytic scoring (on the content standards) per portfolio within the rater pairs to estimate the interrater reliability.

7.4.5.3 The correspondence between teacher beliefs and behaviour as constructed and as assessed

The data from the teacher questionnaire to evaluate the relevance of the feedback given, based on raters' assessments, was analysed descriptively.
Since we had nested data (students within teachers), to analyse the correspondence between teacher beliefs and behaviour as constructed by teachers in their portfolios and as assessed by external raters and by their students we used linear multilevel analysis. We used the following (sets of) predictors: 1) teachers' beliefs concerning the goals of teaching students research skills; 2) teachers' behaviour (research steps), teacher centrality, structuredness of the learning environment); 3) students' assessments of their teachers' classroom behaviour. The criterion variables were raters' mean analytic scores.
To further explore the structure of the scores given we searched for differences between the teachers. To see whether teachers could be grouped according to their scores, we used exploratory Q-principal component analysis on the teachers whose portfolios showed sufficient interrater reliabilities and we used independent samples T-tests to analyse the groups of teachers found for differences in the raters' assessments on the separate content standards (per teacher per content standard taking the mean of the scores of both raters) and also in their students' assessments of their classroom behaviour.

7.5 Results

7.5.1 Constructed teacher beliefs and behaviour

Teacher beliefs. Teachers particularly pursued the following goals: teach students the research subskills (n = 16) and let students practice independent learning (n = 13). A third of the teachers aimed at developing students' subject knowledge (n = 5).
As to control beliefs, according to all teachers their schools provided sufficient facilities for student research. According to 16 of the 18 teachers students had some difficulties with research assignments. The main problems concerned the feasibility of the research assignments in view of the available time (n = 9) and the data gathering (n = 6). Nine teachers gave suggestions to solve these problems. Offering more coaching and formative feedback was mentioned most (n = 5). Two teachers declared to prefer no intervention by the teacher at all, because "doing research concerns student's own learning process". Half of the teachers had followed a one-day or two-day training in teaching students research skills.
About teachers' subjective norms, only half of the teachers said to make clear arrangements with their colleagues about the research assignments, and it was mainly done with regard to the content and the time schedule.
Teacher behaviour. The two video registrations per teacher (n = 36) showed teachers coaching students in class (n = 11), in small groups (n = 11), or individually (n = 3). Eleven registrations included instructional activities (mostly explaining the assignment). Five of the lessons took place in a multimedia centre, the other lessons in regular classrooms.
The teaching and coaching as showed mainly concerned the following research steps or subskills: identify and formulate a problem using subject-specific concepts; formulate a research question; gather and select data (all in two thirds of the videos); report and present the research (in half of the videos). In a third of the lessons attention was given to data analysis, and in only two registrations teachers (also) paid attention to drawing conclusions. Attention to the students' subskill 'evaluating the research' was not shown in the registrations. This result may have been caused by the selection of lessons (most teachers chose lessons and coaching sessions in the first half of the research process of their students). The learning environments the teachers created, differed significantly in their degree of structuredness (range 1.30-3.93, sd .70) ($F = 3.11$; df 35; $p = .019$). On average the learning environment was very structured, due to teachers' centrality as shown in their talking during 76% of the total video registrated time. However, the teachers varied substantially in this percentage (range 43%-95%; sd .13) ($F = 7.79$; df 35; $p < .001$).
In the interviews after video registration seventeen teachers confirmed they had reached their initial goals. Consequently, they were quite satisfied with their lessons. Ten teachers said to base this conclusion on the attitude, comments and questions of the students. However, no teacher explicitly checked whether the students had understood the lesson.

7.5.2 Assessed teacher beliefs and behaviour

Students' assessments. The items in the student questionnaire formed a reliable scale (n = 317; 15 items, mean 3,13, variance .04; Cronbach's alpha .81). On average teachers scored best on the items "My teacher is enthusiastic" (mean 3.43, sd .77) and "My teacher is willing to explain something for a second time" (3.38; .79). The teachers got the lowest scores on the items "My teacher verifies whether we understand the assignment (2.74; .96) and "My teacher veri-

fies whether we are actually working on the assignment" (2.89; .95). The questionnaire results confirmed the results of the interviews after the video registration that showed that teachers hardly monitor students' research activities in an explicit way. One-way Anova showed significant differences between the classes (teachers) ($F = 10.19$; df 17; $p < .001$).
Raters' assessments. The jury alphas for the separate analytic scores were satisfactory for 12 rater pairs, ranging from .39 to .76. The jury alphas were low or even negative for six pairs, ranging from -.80 to .22. The rater pairs appointed exactly the same holistic scores on the five-point scale to 35% of the judgments, a difference of half a point showed up in 12% of the judgments, a difference of one point in 47% of the judgments, and a difference of 1.5 points in 6% of the judgments (one rater pair). The content standards formed a reliable scale (Cronbach's alpha .76).

7.5.3 The correspondence between teacher beliefs and behaviour as constructed and as assessed

All teachers in the pilot (n=3) and the main study (n=18) had returned the questionnaire concerning the relevance of the feedback given, based on the raters' assessments. The items in this questionnaire formed three reliable scales. The first scale contained questions about the relevance of the written feedback (4 items, n = 19), e.g. 'I recognize my teaching in the feedback given' (Cronbach's alpha .84; mean 3.68; sd .19). The second scale was about the transparency of the assessment (3 items, n = 20), e.g. 'It is clear on which standards the assessment of my portfolio is based' (Cronbach's alpha .84; mean 4.0; sd .78). The third scale concerned the way in which the teachers got motivated by the feedback given (3 items, n = 21), e.g. 'The feedback gives information to improve my teaching' (Cronbach's alpha .81; mean 3.8; sd .98). The results show that the teachers were quite satisfied by the feedback given and evaluated the raters' assessments as being relevant.
However, the teachers' beliefs and behaviour as constructed in their portfolios could not significantly predict the raters' mean analytic scores (the beliefs and behaviour as assessed). The linear multilevel analyses showed also that the raters' mean analytic scores could be significantly predicted from the students' assessments of their teachers' classroom behaviour (about how their teachers instruct and coach research skills) ($r = .21$, $p < .05$).
A Q-principal component analysis on the 12 teachers whose portfolios had shown sufficiently interrater reliabilities to take the average of the ratings of the two raters per content standard, using .60 as a minimum for a significant loading, resulted in two groups of teachers (see *Table 7.1*). The components accounted for 34% and 22% of the variance, respectively.
Independent samples T-tests showed that the teachers in the first group scored on average significantly higher on one of the content standards, namely THINK (the choice of and argumentation for teaching strategies that meet students' knowledge, abilities and experience) than the teachers in the second group ($t = 2.78$; df 9; $p = 0.02$). That is, according to the raters, the teachers in the first group are better in explicating the rationales behind their behaviour (mean 2.13 on a three-point scale; sd .22) than the teachers in the second group (mean 1.55; sd .14). The two groups of teachers differed also in their students' assessments of their classroom behaviour. An independent samples T-test showed that the students' assessments in the first group (n = 102; mean 3.21; sd .49) were on average significantly higher than those in the second group (n = 109; mean 2.93; sd .53) ($t = 4.05$; df 208; $p < .001$). The two groups did not significantly differ on their overall portfolio ratings (mean analytic score and holistic score).

Table 7.1 Results of an exploratory Q-principal component analysis (N=12)

Teacher	Component 1	Component 2
1	-0.02	-0.23
2	0.01	**0.62**
3	-0.45	0.28
4	**0.61**	0.51
5	**0.93**	0.04
6	**0.75**	-0.02
7	**0.81**	-0.13
8	-0.07	**0.64**
9	-0.16	**0.92**
10	0.11	**0.65**
11	**0.93**	0.04
12	**0.71**	0.46

7.6 Conclusions

The purpose of our study was to research into the relation between teacher beliefs and teacher behaviour in teacher portfolio assessment, focussing on the interactive phase of teaching. Despite the widespread idea that teacher assessment should take account of teacher beliefs in relation to their behaviour, the role of teacher beliefs in teacher portfolio assessments is far from clear. This is due to the context-bound and tacit nature of teacher beliefs and the fact that raters' interpretations of teacher beliefs do not necessarily correspond with the beliefs the teachers hold themselves.

We distinguished between teachers' beliefs and behaviour as constructed (by teachers in their portfolios) and as assessed (by students observing their teachers and by raters assessing the portfolios). A valid portfolio assessment assumes a correspondence between beliefs (and behaviour) as constructed and beliefs (and behaviour) as assessed, and such a correspondence is needed to make a comparison between beliefs and behaviour as assessed meaningful.

We investigated teachers' constructed beliefs and behaviour by analysing 18 original portfolios (including video registrations) about teaching students research skills. The video registrations showed that the teaching of research skills mainly concerned the research steps 'identify and formulate a problem using subject-specific concepts', 'formulate a research question' and 'gather and select data'. In general teachers highly structured the learning environments, on average they talked 76% of the time, and they barely monitored students' progress. The teachers particularly aimed to teach their students research subskills and to let them practice independent learning. They believed that they have sufficient facilities to teach student research skills and most teachers were satisfied with the way they teach their students. The teachers mainly perceived problems as to the feasibility in view of the available time.

To investigate teachers' assessed beliefs and behaviour, we used both their own students' and external raters' assessments. Students' assessments of their teachers' classroom behaviour showed that they were most positive about their teacher's enthusiasm and least positive about the degree to which their teacher verifies whether they comprehend the assignment. According to the external raters assessing teachers' beliefs and behaviour as described by the teachers in their portfolios, the teachers score best on the content standard 'Choosing an appropriate assignment' and worst on the content standard 'Reflecting on the program and on actions in teaching students research skills'.

The evidence about the correspondence between teachers' beliefs (and behaviour) as constructed and their beliefs (and behaviour) as assessed is mixed. The teachers evaluated the raters' assessments, based on their portfolios, as relevant and their feedback as satisfying. However, the teachers' beliefs and behaviour as constructed in their portfolios could not significantly predict the raters' mean analytic scores (the beliefs and behaviour as assessed).
Also, the evidence about the correspondence between teachers' beliefs and behaviour (as assessed) is mixed. We could not clearly demonstrate a correspondence between the teachers' beliefs and behaviour. This is not an unknown phenomenon and it is often explained by the nature of the constructs assessed (teacher beliefs are mainly tacit and teacher behaviours are mainly automatic). Further, as our sample of teachers was small and also their beliefs and behaviour could be represented in only a small number of portfolio elements, we should not be surprised by this result. Nevertheless, the external raters' assessments of the teachers' beliefs and behaviour (as constructed in their portfolios) on the content standards showed to be significantly related to the students' assessments of their teachers' classroom behaviour. Also, one of the two groups of teachers which could be distinguished using Q-principal component analysis was, according to the external raters, better than the other group in explicating the rationales behind their behaviour. Also, the classroom behaviour of the same group of teachers was more highly evaluated by their own students. These results could be interpreted as indicating at least some correspondence between teachers' beliefs (cognitions) and their behaviour.
As the meaning of these results depends on the degree to which the raters have based their assessments on the content standards as trained, we checked for this (this was done in an earlier study, see chapter 5). We coded the content of raters' judgments written down in their judgment forms and compared this with the full descriptions of the content standards and indicators as used in our project. The results showed that raters mainly base their ratings of content standard THINK on two indicators in this content standard, namely a) the choice of and argumentation for teaching strategies that meet students' knowledge, abilities and experience and d) the alignment with assessment. The indicators 'the teaching strategies fit teachers' goals' (b) and 'the teaching strategies fit the assignment' (c) were less referred to. This could mean that raters in teacher portfolio assessment are focussed on the consistency between teacher cognitions and behaviour.

The participating teachers in the study were all experienced Dutch social science teachers in upper secondary education who voluntarily participated in the study. As a result, the sample will not be representative of the population of all social science teachers, let alone teachers in other subjects. Due to the voluntary participation it is possible and even plausible that the sample consisted especially of teachers who had some experience in teaching research skills. Moreover, only teachers' constructed beliefs and behaviour in their portfolio concerning the interactive phase of teaching students research skills were involved in our study. The results cannot be generalized to other types of contexts (e.g. situations in which other types of teacher assessment instruments are used) or other categories of teacher beliefs and behaviour (e.g. teachers' tacit knowledge or their teaching in the pre-active and post-active teaching phases).
The results of this study give way to further questions addressing the role of teacher cognitions in developing a portfolio. Since we assume that teachers' constructed beliefs and behaviour are partly activated by the assessment instrument and the context they react on, re-

search is needed into the nature of teacher's reactions and the conditions that influence their responses. Such research is likely to be complex, as teachers differ on relevant personality characteristics, experiences, competences, and working context. Nevertheless, we think that quasi-experimental studies with explicit variation in the structuredness of the portfolios applied in different settings, combined with verbal protocols of teachers registrated while they construct a portfolio can reveal more insight in relevant teacher cognitions.

8 Summary of results and general discussion

8.1 Introduction

In the first chapter we introduced three questions that we researched in Chapters 3 to 7. These questions were: 1) How can teacher competences concerning teaching students research skills be validly assessed? 2) How can rater quality in teacher assessment be improved? 3) What is the correspondence between teacher beliefs and behaviour as constructed by teachers in their portfolios and as assessed by their students and by external raters? In the first section we will summarize the main research findings that follow from the studies in this thesis. Next, we will discuss some general limitations of these studies. Then, we recommend suggestions for future research to expand our knowledge about the intersection between cognition and assessment. Finally, some practical recommendations for portfolio assessment are provided. Whereas specific limitations and implications of the studies have been addressed to in each independent chapter, in the next sections we focus on some general issues about teacher portfolio assessment in this thesis as a whole.

8.2 Summary of research findings

In Chapters 3 to 5 we addressed the first question and engaged in procedures to improve the valid assessment for summative and formative assessment purposes of teacher competences.

In Chapter 3 we searched into a Delphi method as a promising option to develop adequate teaching content standards. We tested the method in a small-scale standard setting study. We focused on the sub questions: 1a) How can a Delphi method be used and optimised as a procedure for developing content standards for teacher assessment in the context of educational reform? 1b) Which content standards can be set regarding teaching students research skills? For the development of content standards on teaching students research skills, a Delphi method based on stakeholders' judgments has been designed and tested. We made a first selection of relevant teaching tasks based on a literature study into teaching students research skills and on an empirical study into leading edge teachers' practices regarding developing students' research skills. We made an initial description per task, which we converted into a questionnaire that formed the input to the Delphi study. Next, we selected a Delphi panel

by asking groups with a considerable stake in the results of the study which persons or organizations should participate in the process of developing content standards. Most of them mentioned practising teachers and external experts. Subsequently, from a random sample of 115 schools we approached the heads of the geography, history and economics departments and the school principles with, amongst others, a detailed intake form and a request for participation. 21 stakeholders were selected to participate in the Delphi study.
In three monthly cycles the stakeholders judged the initial descriptions of content standards based on a 4-point Likert scale on the aspects: content relevance, thoroughness of formulation, clarity of formulation and correspondence of the standards with everyday teaching practices. They also answered open questions about the proposal as a whole. To improve the content standards, in each round the stakeholders were asked for suggestions. After each round the standards were revised, based on their judgments and comments, after analyses by two researchers.
The study shows that the Delphi method makes it possible to gradually improve content standards by an iterative process. The method resulted in nine content standards that are considered by 21 stakeholders to be highly relevant, thoroughly and clearly formulated and corresponding well with everyday teaching practice. The results show satisfactory interrater and interitem consistencies. We argued that although the ratings became more positive in the course of the rounds, it is not fully clear what caused this change.
The content standards, which could be set, were the following. In teaching students research skills teachers should know and be able to: 1) select goals for students; 2) choosing an appropriate assignment: 3) prepare and manage students to work on their assignments; 4) plan and select teaching strategies that support the development of students' research skills; 5) teach students research skills; 6) create a positive pedagogical climate; 7) adequately assess the research skills of students; 8) reflect on the program and on actions in teaching students research skills; 9) collaborate with colleagues (chapter 3).
The next issue concerned the degree to which teachers have to meet the content standards for a satisfying result. We decided not to include content standard 9 (collaborate with colleagues) in the performance standard setting study, because this issue highly depends on a teacher's workplace and is barely put into practice in many schools (Stokking & Van der Schaaf, 2000; Van der Schaaf, Veenhoven & Stokking, 2003; Rijborz, 2003). In chapter 4 we examined a Policy capturing procedure to develop teacher performance standards that can be used for performance assessment. We searched into the sub questions: 1c) How can a Policy capturing method be used and optimised as a procedure for developing performance standards for teacher assessment? 1d) Which performance standards can be set regarding teaching students research skills?
We reasoned that methods based on Policy capturing are particularly interesting because they can be used both for the development of performance standards and to gain insight into a fitting judgmental model (compensatory or conjunctive). Moreover, they can support the development of weights to be attached to the content standards (Jaeger, 1995), and they can stimulate the judgmental process of the panellists. This method has also disadvantages, being its difficulty for the judges and the fact that it is not clear how to foster consensus. To counter these difficulties we took the following measures: a careful selection of panellists representing different opinions and having sufficient knowledge and motivation for their task; construction of the judgmental based on eight content standards about which panellists in an earlier study had reached consensus; a training process; interim feedback; a group discussion using

the nominal group technique (Delbecq, Van de Ven & Gustafson, 1975); and judgment of profiles in two rounds.
In two rounds, nine stakeholders (selected from the Delphi panel in chapter 3) individually judged 64 respectively 27 fictitious teaching profiles containing scores on eight criteria on 3 anchor points (the teacher satisfies the content standard completely, reasonably, or to a small degree). Between the rounds, the panellists held a structured group discussion in which they discussed the underlying arguments for their judgments.
The procedure resulted in significant and consistent results, in the form of regression models. The panellists' judgments in the second round were more consistent than in the first round. It is not clear precisely which intervention or combination of interventions contributed to this growth in consistency. Further, the panellists agreed rather strongly on the performance standards accompanying the eight content standards. However, the panellists did not agree on the weights to be attached to the content standards in the final judgment, referring to their own school context and personal experiences. We could not fully trace what has caused these results.
The purpose of chapter 5 was to develop a teacher portfolio design for summative and formative assessment purposes. We focused on two links between relevant components in a portfolio development process, namely the link between content standards and the portfolio design and the link between the content standards and the raters' scoring. The main sub questions were: 1e) How is the initial portfolio design related to the content standards? 1f) How is raters' portfolio scoring related to the content standards?
Eight experts participated in a panel study. Most of them also participated in the studies reported in chapters 3 and 4. The experts judged an initial portfolio design based on the content standards and performance standards as formulated in Chapters 3 and 4. The portfolio design was also based on the Theory of Planned Behavior (Ajzen, 1985; Ajzen & Fishbein, 2005). They used a Policy capturing method (comparable to the one in Chapter 4) to weight the portfolio elements. Also a pilot study, in which teachers using the initial design voluntarily developed a portfolio, was included. The results show that the experts consistently agreed with the initial design and that the experts and the teachers thought the design to be satisfactory related to the content standards.
Subsequently, 18 teachers developed a portfolio based on the final design and 6 trained raters (selected from the eight experts of the previous Policy capturing study) assessed the teacher portfolios according to eight content standards. The results show that both the experts and the teachers preferred artefacts and reproductions as forms of portfolio elements above productions. Nevertheless, we argued that for the sake of a valid portfolio design it is important to include productions that represent teacher cognitions. The interrater reliability of 12 portfolios was satisfactory. The results indicate that raters quite strongly based their assessments on the content standards as trained.

The second question was: How can rater quality in teacher assessment be improved? In chapter 6 we reported on a study into the sub questions: 2a) What is the reliability of the assessments? 2b) How do raters assess teachers' portfolios? 2c) How are raters' cognitive processes related to their judgments? We researched into this question by training six raters (the same raters as in the study reported in chapter 5) who subsequently systematically assessed 18 portfolios, of which each one was independently rated by two raters, and by searching into their cognitions using verbal protocols, judgment forms and interviews.

In a training session raters practised in scoring teacher portfolios on judgment forms developed by the researcher and in discussing the arguments underlying their judgments given. In a pilot six raters individually scored three teacher portfolios on a judgment form. During the pilot all raters also retrospectively verbalized their judgments of one portfolio during a two-hour session with the researcher. The verbalisations were fully transcribed and coded by two researchers. The coding scheme was primarily based on the Associated Systems Theory (Carlston, 1992; 1994). The categories of the coding system could satisfactorily be distinguished in the verbal protocols and the judgment forms of raters' scoring of three teacher portfolios in the pilot.
After the pilot in a 9-month period, the six rater pairs judged 18 portfolios. Each portfolio was rated independently by two raters on judgment forms. The raters retrospectively verbalized their rating process of two randomly chosen portfolios (after scoring) from the portfolios they had judged. The judgment forms and protocols were fully transcribed and coded.
The interrater reliability of 12 out of 18 portfolios was satisfactory. Variance analyses showed only slight rater effects. Raters' holistic scores in all but one case differed by no more than one point (on a 5-point scale).
The results show that the raters used cognitive representations on the dimensions concrete-abstract (e.g. visual manifestations) and positive–negative (positive and negative comments), and on the categories the situation in which teachers operate their own judgment process, and the judgment procedure. Whereas we used the judgment forms completed during the judgment sessions as cues for producing the retrospective verbal protocols, the comments in the verbal protocols were generally less concrete than those in the judgment forms. The cognitive representations in the judgment forms were significantly related to the judgments given. The differences in cognitive representations within rater pairs were significantly related to the interrater reliabilities.
Although the raters were instructed and trained in weighting the scores according to the policy as developed in the performance standard setting study reported in chapter 4, the raters based their holistic judgments of the portfolios on the unweighted means of the analytic scores given. Further, retrospective interviews revealed that the raters wanted to build a coherent image of the teacher they were assessing. The raters' experience was that the teacher images were influenced by previously rated portfolio material and the content standards, as well as by the raters' own experiences as a teacher. Next, we discovered that the judgments given could significantly be predicted by the raters' representations as categorized on the basis of their comments in the judgment forms on the dimensions concrete–abstract and positive–negative. The higher the correspondence within rater pairs in these representations, the higher the interrater reliability. We argued that this result is significant because it indicates the existence of a relation between raters' cognitions and their assessments given.

The third question was: What is the correspondence between teacher beliefs and behaviour as constructed by teachers in their portfolios and as assessed by their students and by external raters? We researched into this question in chapter 7 by comparing plain portfolio data (including self descriptions and video registrations of teachers' lessons) with assessments by students and raters. We distinguished between beliefs constructed by the teachers, i.e., beliefs teachers hold themselves and explicate in their portfolios, and raters' assessed teacher beliefs, i.e., raters' representations about teachers' beliefs as explicated in their portfolios.
Since a prerequisite for a valid portfolio assessment is that teachers' constructed and assessed

beliefs are related, we qualitatively analysed the beliefs and behaviour of 18 teachers as constructed in their portfolios. Teachers' interactive behaviour was coded from videotapes of two lessons per teacher which were fully transcribed. We analysed the videos in terms of the content (research steps), the structuredness of the learning environment, the setting, and the students' grade. Ratings of the learning environment were done on a 3-point scale from 'very structured' to 'very low structured'. Data concerning background factors, and teachers' goals and beliefs were analysed descriptively.
Further, the 18 portfolios were rated on eight content standards (developed in chapter 3), each one independently by two trained raters. According to the raters, the teachers scored best on the content standard 'Choosing an appropriate assignment' and worst on the content standard 'Reflecting on the program and on actions in teaching students research skills'. The teachers' behaviour was also assessed by their own students in a questionnaire concerning how their teacher instruct and coach research skills (total n = 317). The results show that students were most positive about their teacher's enthusiasm and least positive about the degree to which their teacher verifies whether they comprehend the assignment.
Linear multilevel analysis showed that students' assessments of the teachers' classroom behaviour could be significantly predicted by the raters' assessments of the teachers' beliefs and behaviour as described in their portfolios. However, the raters' assessments themselves could not be significantly predicted from the beliefs and behaviour as described in the portfolios. We searched for differences between the teachers by using exploratory Q-principal component analysis on those raters' assessments that had a reasonable to good degree of inter rater reliability and we tested the two groups of teachers found for differences in the students' assessments of their classroom behaviour. Teachers with high rater judgments on the content standard THINK ("the choice of and argumentation for teaching strategies that meet students' knowledge, abilities and experience") had significantly higher student assessments than teachers who were judged low on this content standard.

8.3 Study limitations

In chapter 2 we argued that teacher competence-based portfolio assessment must be subjected to the same critical review as other types of assessments (Linn, Dunbar & Baker, 1991; Linn, 1994; Messick, 1994; Haertel, 1999; Stokking & Voeten, 2000; Stokking, Van der Schaaf, Jaspers & Erkens, 2004). Therefore, reconsidering the limitations of the studies reported in this thesis we are leaded back to the quality requirements established in chapter 2 concerning valid and reliable measurement and adequate assessment, i.e., acceptable and practical useful.
Portfolio assessment suffers from inherent difficulties that are also into play in the studies in this thesis. Probably the most pressing concern in obtaining valid measurements is how best to turn complex teaching and teacher development into teacher assessment. Contemporary teacher assessments generally suffer from a lack of theoretical understanding about teacher competences and the way teacher competences develop over time. As teachers' professional development is context dependent and can be described as an irregular process within which sudden improvements occur (Huberman, 1995; Day, 1999), most teacher assessment models are not suitable to assess teachers' development. For that reason, we based the performance standards on minimally required mastery levels, instead of on acquired insight into the development in teachers' knowledge and behaviour over time. Obviously, research into teachers'

professional development in context should be a major focus of research (Kwakman, 1999; Bakkenes, Vermunt & Wubbels, 2004).

We used stakeholders' judgments in a Delphi study and in a Policy capturing study to further develop content and performance standards. A difficulty of such judgment studies is to explain what causes the judgments of the panellists. For example, in chapter 3 in three Delphi rounds the ratings of an initial set of content standards became more positive in the course of the rounds. The change may alternatively be attributed to the iterative development of the content standards using the critical remarks of the panellists, to the feedback given, to the panellists' familiarization with the content standards (a learning effect), or even to the fact that some panel members have simply confirmed the majority judgment (Greatorex & Dexter, 2000). The policy capturing study to develop performance standards in chapter 4 showed that the panellists agreed rather strongly on the performance standards accompanying the eight content standards, but did not agree on the weights to be attached to the content standards in the final judgments. We could not fully trace what had caused these results.

Another major concern in obtaining valid and reliable measurements is that teacher portfolio assessment involves complex interactions between teacher competences, the portfolio, the standards used, rater characteristics, and raters' interpretations. Consequently, portfolio assessments usually show a lack of generalizability, i.e., a low degree to which the results are valid for a broader range of tasks, conditions and populations. We now discuss the main restrictions with respect to validity and reliability.

Firstly, since portfolio assessments take place in social settings and are context dependent, the assessment results reported are relevant for a particular population and specific situations. As only a small teacher sample was involved, all teachers being experienced social science teachers in upper secondary education and participating voluntarily, the sample is not representative of the population of Dutch social science teachers. Nevertheless, the literature shows that the methods used in this thesis (e.g. the Delphi method and the Policy capturing method) can also be employed in other situations (De Loe, 1995; Linstone & Turoff, 1975; Jaeger, 1995).

Secondly, in order to provide a valid assessment, we think that teacher assessment has to be general enough to accommodate the broad variation in teachers' choices, performances, and products. Likely, limitations are related to the number of portfolio elements in our studies. Teachers used four types of artefacts, three types of productions, and one type of reproduction to show their competences in teaching students research skills. Further, the portfolio elements concentrated on specific teacher behaviour and cognitions related to teaching students research skills.

Thirdly, we did not use parallel tests to research into convergent and divergent validity because such tests simply do not exist and the contextual and unique character of teaching would make such parallel tests less valuable. Therefore, we followed another course to find support for the validity of the assessments in this thesis. For example, we developed content and performance standards in repeated rounds, and researched into rater cognitions and raters' use of content standards while using verbal protocols and judgment forms. Further research is needed to ascertain the exact value of these methods to validate portfolio assessment.

Finally, in chapter 6 we reported on an analysis of variance between teachers and between raters on the scores given to estimate whether the score variability was due to rater effects. Because of the structure of the available data we could not check for possible interaction ef-

fects between raters, teachers and other facets of the portfolio assessment. Though we could make a stronger case when we could indicate the generalizability of the scoring, we did make several explicit attempts to enhance the validity of the studies in this thesis. For instance, we conscientiously developed content standards based on a literature review and empirical research into teaching students research skills (Stokking & Van der Schaaf, 2000), and we used structured panel studies to develop content standards, performance standards and a portfolio design. Furthermore, we thoroughly trained the raters in our study and we structured the assessment process by using a rater manual and judgment forms. We analysed the judgment processes and we checked whether the portfolios were sufficiently authentic and whether the ratings were recognizable according to the assessed teachers. All in all we think we gathered enough support to claim that the assessments reported in this thesis are sufficiently valid and reliable. However, for teacher portfolio assessment to be used for summative assessment purposes generalizability studies are urgent, focussing on the facets raters and teaching situations (Straetmans & Sanders, 2001).

As to the content standards of adequate assessment (acceptability and practical utility), the quality of portfolio assessment is affected by many factors. These factors include: the time and amount of money available to develop valid and reliable portfolio assessments; the degree to which a teacher's portfolio relates to his or her working context; the support and time a teacher receives to prepare a portfolio; and the training and support of the raters. As a result, to conduct satisfying portfolio assessments is a costly and time-consuming activity and it is likely that in practice not all these prerequisites can be met.

Although we partly based the content standards on empirical research, because authentic data was unavailable we had to use fictitious data in the Policy capturing studies. Of course, the development of performance standards and the establishment of the weights of the portfolio elements would ideally be based on empirical cases representing authentic teaching situations. The use of fictitious data may limit the conclusions we can draw. To make a stronger case for performance standards set by judgment studies, we suggest to compare the results of the judgmental policy capturing study with the results of other methods to identify how good teachers perform on the content standards, for example the contrasting groups method (Berk, 1976).

Further, to line up with teachers' abilities and possibilities to construct a portfolio and to make the portfolio assessment process more feasible, we as researchers structured the content standards and the portfolio elements. One could argue that the structure we put in may have masked or even caused misrepresentation of teachers' personal context-bound knowledge and beliefs in teaching students research skills (Kagan, 1990). However, on the other hand we assume that open content standards and a totally unstructured portfolio design would not guarantee that a teacher's highly personal cognitions are represented in his or her portfolio. Instead, we assume it is a prerequisite for valid assessments that the portfolio design, the content standards, the scoring model, the teachers' and raters' possibilities, and the aim and context of the assessment are aligned.

Because of their practical utility portfolio assessments might play an effective role in teacher development. Teacher portfolio assessments are often used for formative assessment purposes, e.g. to facilitate the teachers' ability to achieve insight into teaching difficulties and to give teachers meaningful feedback (Lyons, 1998). These effects of assessment represent important contributions to its consequential validity (Cronbach, 1988; Messick, 1975, 1989; Dierick & Dochy, 2001). Though in general the teachers in our study said they had learned from con-

structing a portfolio and they were quite satisfied with the feedback they had received from the raters (chapter 7), teacher questionnaires are self-evaluations and do not show the exact influence of portfolio assessment on teacher development. Future research should search for effects of portfolio assessments on teacher development.

8.4 Implications for future research

In this thesis we view construct validation as the key issue of portfolio assessment. Construct validation concerns the theoretical and conceptual definition of the constructs measured and the empirical evidence that supports the adequacy of these measurements. In this view, strong theoretical frameworks are important prerequisites to construct a teacher portfolio design and to interpret the empirical results of portfolio assessments (Messick, 1989). Such theoretical frameworks should at least cover two fields. Firstly, the constructs measured i.e., the conceptualisation of teacher competences and teacher competence development. Secondly, panellists' and raters' cognitions while developing content standards, performance standards, and a portfolio design, and while assessing portfolios. In this thesis, we primarily aimed to contribute to the second field. Therefore, this field is central here.

It is not yet clear which theories are needed in portfolio assessment studies, let alone that they have been integrated into one single theory. In this thesis we have build upon theories in different traditions within social cognitive psychology that we assume to be useful in portfolio assessment studies. Firstly, we used the Theory of Planned Behavior (TPB; Ajzen, 1985; Ajzen & Fishbein, 2005) to develop and analyse teacher portfolios (based on teachers' constructed beliefs and behaviour) in chapters 5 and 7. The TPB is a leading theory in social psychological research into the relationship between human cognitions and behaviour. Secondly, we used the Associated Systems Theory (AST; Carlston, 1992; 1994) to capture raters' cognitive processes while rating, reported about in chapter 6. The AST focuses on several forms of human cognitive representations that operate simultaneously when forming impressions of other people and expands on previous research in social psychology and neurology. Furthermore, the Policy capturing studies conducted in chapters 4 and 5 are based on theoretical approaches to judgments and decision making as described in the Cognitive Continuum Theory (CCT) (Hammond, 1980). The CCT focuses explicitly on the tension between intuitive and analytic thinking in human judgment processes. As mentioned before (see chapter 2), Brunswik (1952) conceptualised human judgment in a 'lens model'. This model parallels the thought of a distal cause in the environment that scatters its effects that humans re-combine when they see an object. In teacher portfolio assessment raters also have to 'recollect' various portfolio elements (i.e., effects) developed by a teacher (i.e., distal cause in the perception of the rater).

The theories mentioned describe distinct facets that can be important in different phases of portfolio assessment studies. For example, the TPB can be useful in the phase of developing a portfolio design (chapter 5), and the AST showed to be important in analyzing raters' cognitive representations while scoring (chapter 6). Although the theories have been useful in the studies in this thesis, on a more general level they seem to differ in their suitability to describe teacher and rater cognitions. A restriction of especially the TPB and the methodology used in Policy capturing is that they cannot fully explain tacit cognitive processes. After all, the TPB focuses only on explicated cognitions and the multiple regression models used in the Policy capturing studies produce quite static representations of raters' judgments, 'averaged' over

profiles. Though Policy capturing proved to be useful in our study, especially to trigger panellists to explicate their underlying cognitions in their judgment processes (chapter 4), in fact it may not work well to fully explain rater cognitions because the multiple regression models used are based on the idea that the meaning of the data in the models is similar for all panellists. For example, regression analysis implies the idea that raters interpret scale points in the same way and that there are no grey areas between scale points. In practice, this is not per se the case, and some scholars even found evidence that raters vary in the meaning they attach to the same scale-points (e.g. Smith, 1993). We think the AST may be more convenient to describe raters' cognitive processes because this theory explicitly distinguishes self-referenced representations involved in raters' personal reactions to the assessees, such as tendencies or predispositions to respond to a person in a particular manner (orientations) and affective responses. In general, these personal reactions are of a more tacit nature.

Despite their differences in origin, scope and convenience, we recognize some ways in which the theories potentially meet. Fazio (1994) summarizes parallels between literature on person perception (e.g. AST) and research on attitudes (e.g. TPB), regarding, amongst others: a great concern with the integration of information into a summary judgment, e.g. regarding the impressions of a person (AST) or beliefs towards a behaviour (TPB); the recognition that these summary judgments influence memory processes (e.g. as stated in chapter 6, rating teacher portfolios is influenced by portfolios rated before); and a recognition of the influence of situational conditions that prompt persons to form such summary judgments. Furthermore, the AST and the TPB correspond with literature on the CCT in the sense that all theories recognize different modes of cognitions that can be portrayed on a continuum from intuitive tacit cognitions to more analytical explicit cognitions.

We believe that synthesizing different useful theories can be a first step to develop a stronger theoretical basis for portfolio assessments. Such synthesizing efforts should focus on differences in perceptions of the same content standards, performance standards and/or constructed portfolios by different judges, the nature of these perceptions, and the question how these perceptions are (interactively) constructed.

The development of such a theoretical framework requires research designs that are responsive to raters' cognitions. Accordingly, in the studies in this thesis we explored the possibility of coupling panellists' and raters' qualitative arguments (e.g. as shown in verbal protocols and judgment forms) to quantitative judgment outcomes (e.g. scores given to teacher portfolios) by adding phases to the judgment process wherein panellists and raters generate motives for their judgments. For example, in chapter 3 we searched into panellists' comments on their judgments of content standards, in chapters 4 and 5 we explored panellists' reasons for their judgments given in Policy capturing studies, and in chapter 6 we qualitatively researched raters' cognitions and related these cognitions to the quantitative judgments given. By using a mixed quantitative and qualitative approach to study rater cognitions and judgment outcomes, we managed to provide some evidence for a relation between raters' cognitions and their judgments. We conclude that this relation needs further exploration, for example by studies into possible changes in the content of panellists' cognitive representations of the content standards during a standard setting process. Verbal and written data collection can provide more insight into the rating process, for example by using verbal protocols and written judgment forms. Next, gathering verbal and written data by novice and expert raters may be fruitful to further our understanding of what constitutes a 'good' rater (Huot, 1993).

To improve the validity of portfolio assessment, according to the chain model (see chapter

5), at least four salient components are into play and interacting with each other, namely: the teacher model, the task model, the scoring model and the assessment model. In chapters 2 and 5 we reasoned that certain characteristics of the constructs, the context and the aim of the assessment may impact the kind of evidence for the interactions between the components. For example, a portfolio (task model) can be more or less structured, the scoring model can be more or less analytic or holistic, and the aim of the assessment can be more or less summative or formative. Insight into how characteristics of these assessment process components influence portfolio assessment is vital for improving the quality of portfolio assessments. Quasi-experimental research designs will be constructive for this purpose. Such designs should take account of the complex situations in which teacher assessment takes place and can be responsive to the personal data collected. Recently we started such a quasi-experimental study comparing different portfolio designs and their impact on students' competence development in pre-university education (Stokking & Van der Schaaf, 2004).
A final suggestion for follow-up research, also pursued in the study just mentioned, is to integrate theory and research into teacher competences and teacher development with the domain of learning and instruction. For example, standards teachers should meet, in beliefs and behaviour, correspond to standards their teaching should meet. Some of the most important standards deal directly with the coordination and regulation of the interaction with and between students, and teachers while interacting with their students bring in their professional competences and develop them further and students' feedback is highly potentially formative in this process, just as the other way round (teacher's feedback being formative for students).

8.5 Suggestions for assessment practice

Although it is common to view teacher portfolio assessment as a tool for improving teacher development (Lyons, 1998), there is less evidence of its appropriateness for summative assessment purposes. This thesis aimed to provide some insight into this. In this last section we review some suggestions for assessment practice, more or less directly based upon our research. The suggestions are arranged to the issues in the assessment of teacher competences mentioned in chapter 1: the development of content and performance standards, the quality of teacher portfolio assessment, and the relevant quality requirements.

Developing teaching content and performance standards

- The meaning of content and performance standards strictly spoken can only be defined relative to the underlying perceptions of the stakeholders that participate in standard setting studies. Content standards are not at random chosen samples of teacher's work, rather they are constructions of a group of stakeholders about what teachers should know and be able to do. Therefore it is important to involve teachers in a standard setting procedure. When using the standards (for example in assessment situations) the precise meanings of content and performance standards become important and the users of the standards (e.g. assessors and assessees) at least should endorse these meanings (chapters 3 and 4).
- The Delphi method and the Policy capturing method as used in our studies are useful to develop content standards and performance standards in a structured way. Communicating and discussing about the content and performance standards in organized group discussions might clarify the meanings underlying the standards and might be helpful to

achieve a shared understanding of these meanings. Disagreement among competent panellists should be taken serious and is an indication of the complexity of the content and performance standard setting process rather than evidence of error variance in the judgment processes (Smith, 1993; chapters 3 and 4).

- As the meanings of the content and performance standards are relative to the participating panellists, a careful selection of panellists becomes essential, for example by an intake focussing on their motivation, their ideas about teaching, their own experiences in teaching, and their qualities in teacher assessment. In our studies the numbers of 21 stakeholders in a Delphi procedure and 9 experts in a Policy capturing study were enough to produce stable results (Linstone & Turoff, 1975; Berk, 1996; chapters 3 and 4).
- As the panellists' understanding of the standard setting method is critical to the final standards produced, the panellists should be thoroughly trained in the standard-setting methods. It is advisable to give panellists the opportunity to discuss and improve their first set of judgments. A second set of judgments will often represent standards of a better quality (Hambleton, 1997; chapter 4).
- As raters tend to join content standards (chapter 4 and 6) it is wise to limit the number of content standards. Though using a number of eight content standards was feasible in our studies, others advise to cut down the number to five (Arthur, Woehr & Maldegen, 2000; Kolk, 2001).
- High-stake teacher assessment, such as for certification or merit pay, must result in a single outcome. To combine the scores on the content standards into a final judgment a compensatory model seems reasonable, because it is realistic to assume that teachers have their weak sides and that they still have room to develop further. When the content standards in a compensatory model are weighted, it is advisable to be very explicit in the attributed weights because raters prefer to simply average the scores on the content standards in order to get and overall judgment (chapter 6).

The quality of teacher portfolio assessment

- The purpose and context of portfolio assessment impose constraints on the assessment design. When using a single portfolio assessment design for multiple purposes (e.g. summative and formative assessment purposes), the designers and users (i.e., teachers) of the assessment design should recognize the tensions this probably entails. A prerequisite for combining purposes is to clearly communicate to the teachers which portfolio evidence will be used for what assessment purposes.
- Portfolio assessment should not be used for teacher selection, because the predictive validity of portfolio assessment will be low due to its context specific and personal character and its low generalizability. The process of construct validation, searching for evidence about the meaning of the assessment scores given, should not be confused with the analysis of predictive validity.
- Designing portfolio assessments is a complex and iterative process in which numerous components are involved and developmental decisions are revised and refined. A chain model of construct validation can be helpful in this process, because it points to important elements that should be explicitly linked. The process of developing a portfolio assessment design should be conscientiously documented using such a model (chapter 5).
- Raters' representations on the dimensions concrete-abstract (e.g. visual manifestations) and positive-negative (positive and negative comments) are related to the quality of their

portfolio assessments in this thesis. It is suggested to include these dimensions in rater training (chapter 6). For example, raters can be trained to alternate concrete with abstract representations in their judgments. Raters can also be familiarized with giving positive as well as negative comments in the judgment forms according to their judgments. This implies that rater training should support raters to increase their understandings of the content and performance standards and the training should teach raters to be consistent in their judgments and their evaluative comments.

The relevant quality requirements

- The practical relevance of the psychometric quality criteria concerning reliable, valid and adequate assessments (see chapter 2) deserves critical reflection and differentiation (Stokking et al., 2004). One point of view is that raters (e.g. schoolheads) cannot reasonably be required to meet the quality criteria (in view of their competences, time and facilities). However, in our view, the quality requirements by which those practices are to be judged depend on the purpose of the assessment: whether it is formative or summative, for the teachers, and/or for others. In all cases validity should have priority (cf. Linn et al., 1991; Linn, 1994), because assessment is all about the intended teacher competences. Transparency, lack of bias and functionality should always be satisfactory. In formative assessment, reliability, objectivity, and equality might get less priority, because the results only indicate possibilities for improvement. Workability and efficiency might get a high priority, so that feedback to teachers can be frequently given. In summative assessment reliability, objectivity and equality are also important.
- For high-stake purposes, special measures for quality assurance should be routine. Such measures might include using standards in the form of model products or model answers on each performance level (e.g. a good and a bad example); using more than one assessor (controlling for consistency); and using information from more than one source, for example the scores of a portfolio assessment, of an assessment centre, and of a knowledge test (contributing to generalizability) (Stokking et al., 2004).

References

Aguirre, J., & Speer, N. M. (2000). Examining the relationship between beliefs and goals in teacher practice. *Journal of Mathematical Behavior, 18* (3), pp. 327-356.

Ajzen, I. (2002). Perceived behavioral control, self-efficacy, locus of control, and the Theory of Planned Behavior. *Journal of Applied Social Psychology, 32*, 665-683.

Ajzen, I. (1985). From intentions to actions: a Theory of Planned Behavior. In J. Kuhl & J. Beckman (Eds.), *Action-control: from cognition to behavior* (pp. 11-39). Heidelberg, Germany: Springer.

Ajzen, I., & Fishbein, M. (2005). The influence of attitudes on behavior. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 173-221). Mahwah, NJ: Erlbaum.

Amelang, M., & Borkenau, P. (1986). The trait concept: Current theoretical considerations, empirical facts, and implications for Personality Inventory Construction. In A. Angleitner & J. S. Wiggins (Eds.), *Personality assessment via questionnaire* (pp. 7-24). London: Plenum Press.

Anderson, J.R. (1982). Acquisition of cognitive skill. *Psychological Review 89*, 369-406.

Andrews, T.E., & Barnes, S. (1990). Assessment of teaching. In W.R. Houston (Ed.), *Handbook of research on teacher education* (pp. 569-598). New York: Macmillan.

Angoff, W.H. (1974). Criterion-referencing, norm-referencing and the SAT. *College Board Review*, 92 (summer), 2-5, 21.

Arter, J. (1993). *Designing scoring rubrics for performance assessments: the heart of the matter*. Portland, OR: Northwest Regional Educational Laboratory, Test Center.

Arthur, W., Woehr, D., & Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: a conceptual and empirical re-examination of the assessment center construct-related validity paradox. *Journal of Management, 26* (4), 813-835.

Artzt, A.F., & Armour-Thomas, E. (1998). Mathematics teaching as problem solving: a framework for studying teacher metacognition underlying instructional practice in mathematics. *Instructional Science, 26*, 5-25.

Baker, E.L., Abedi, R.L. Linn, R.L., & Niemi, D. (1995). Dimensionality and generalizability of domain-independent performance assessment. *Journal of Educational Research 89* (4), 197-205.

Bakkenes, I., Vermunt, J., & Wubbels, T. (2004). *Leren van docenten in de beroepspraktijk vanuit een theoretisch perspectief.* Paper presented at the Onderwijs research dagen, June 9-11 2004 Utrecht University, The Netherlands.

Bandura, A. (1998). Health promotion from the perspective of social cognitive theory. *Psychology and health, 13*, 623-649.

Barton, J., & Collins, A. (1993). Portfolios in teacher education. *Journal of Teacher Education, 44* (4), 200-210.

Baxter, J.A., & Lederman, N.G. (1999). Assessment and measurement of pedagogical content knowledge. In J. Gess-Newsome & N.G. Lederman (Eds.), *Examining pedagogical content knowledge. The construct and its implications for Science education,* (pp. 147-161). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Bechtold, H. (1959). Construct validity: a critique. *American Psychologist, 14,* 619-629.

Beijaard, D., & Verloop, N. (1996). Assessing teachers' practical knowledge. *Studies in Educational Evaluation, 22* (3), 275-286.

Berk, R.A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education, 45*, 4-9.

Berk, R.A. (Ed.) (1980). *Criterion-referenced measurement. The state of the art*. Baltimore, London: The Johns Hopkins University Press.

Berk, R.A. (1996). Standard setting: the next generation (where few psychometricians have gone before!), *Applied Measurement in Education*, 9 (3), 215-235.

Berliner, D.C. (1994). Expertise: The wonder of exemplary performances. In J. Mangieri, & C. Block (Eds.). *Creating powerful thinking in teachers and students: Diverse perspectives* (pp. 141-186). Fort Worth, TX: Harcourt Brace.

Berliner, D.C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research, 35,* 436-482.

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

Bird, T. (1990). The schoolteacher's portfolio: an essay on possibilities. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: assessing elementary and secondary school teachers* (pp. 241-256). Newbury Park, CA: Corwin Press.

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

Blumenfeld, P.C., Soloway, E., Marx, R.W., Krajcik, J.S., Guzdial, M., & Palincsar, A. (1991) Motivating Project-Based learning: Sustaining the doing, supporting the learning. *Educational Psychologist, 26*, 369-398.

Boekaerts, M. (1991) Subjective competence, appraisals and self-assessment. *Learning and Instruction 1*, 1-17.

Bond, L., Smith, T.W., Baker, W.K., & Hattie, J.A. (2000). *The certification system of the National Board for Professional Teaching Standards: a construct and consequential validity study*. Greensboro: Centre for educational research and evaluation, University of North Carolina.

Borko, H., & Livingston, C. (1989). Cognitions and improvisation: differences in mathematics instruction by experts and novice teachers. *American Educational Research Journal, 26* (4), 473-498.

Borko, H., & Putnam, R.T. (1996). Learning to teach. In D.C. Berliner & R.C. Calfee (Eds.), *Handbook of educational psychology* (pp. 673-708). New York: Macmillan.

Borman, W.C. (1977). Consistency of rating accuracy and rating error in the judgment of human performance. *Organizational Behaviour and Human Performance, 20*, 238-252.

Borsboom, D., Mellenbergh, G.J., Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111* (4), 1061-1071.

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

Bromme, R. & Tillema, H. (1995). Fusing experience and theory: The structure of professional knowledge. *Learning and Instruction, 5*, 261-269.

Brookhart, S.M.(1999). Response to Delandshere and Petrosky's 'Assessment of complex performances: limitations of key measurement assumptions. *Educational Researcher, 28* (3), 25-28.

Brown, A.L. (1992). Design experiments: theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences, 2,* 141-178.

Brown, C.R., Moore, J.L., Silkstone, B.E., & Botton, C. (1996). The construct validity and context dependency of teacher assessment of practical skills in some pre-university level science examination. *Assessment in Education, 3* (3), 377-391.

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

Bruner, J.S. (1957). Going beyond the information given. In H. Gruber, K. Hammond & R. Jessor (Eds.), *Contemporary approaches to cognition* (pp. 41-69). Cambridge, MA: Harvard University Press.

Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago: University of Chicago Press.

Bullough, R.V. (1997). Becoming a teacher: self and the social location of teacher education. In B.J. Biddle, T.L. Good & I.F. Goodson (Eds), *The international handbook of teacher and teaching* (pp. 79-134). Dordrecht, The Netherlands: Kluwer.

Burns, C.W. (1999). Teaching portfolios and the evaluation of teaching in higher education: Confident claims, questionable research support. *Studies in Educational Evaluation, 25*, 131-142.

Calderhead, J. (1996). Teachers: beliefs and knowledge. In D.C. Berliner & R.C. Calfee (Eds), *Handbook of educational psychology*, (pp. 709-725.) New York: Macmillan.

Carlston, D. (1992). Impression formation and the modular mind: the associated systems theory. In L.L. Martin & A. Tesser (Eds). *The construction of social judgments* (pp. 301-341). Hillsdale, NJ: Erlbaum.

Carlston, D. (1994). Associated systems theory: A systematic approach to cognitive representations of persons. *Advances in Social Cognition, 7*, 1-78.

Carter, K. (1990). Teachers' knowledge and learning to teach. In W.R. Houston (Ed.), *Handbook of research on teacher education* (pp. 291-310). New York: Macmillan.

Chi, M., Glaser, R., & Farr, M.J. (1988). *The nature of expertise*. Hillsdale, NJ: Elrbaum.

Clark, C.M., & Peterson, P.L. (1986). Teachers' thought processes. In M.C. Wittrock (Ed.), *Handbook of research on teaching, 3rd ed.* (pp. 255-296). New York: Macmillan.

Clark, C.M., & Yinger, R.J. (1979). Teachers' thinking. In: P.L. Peterson & H.J. Walberg (Eds.), *Research on teaching; concepts, findings, and implications* (pp. 231-263). Berkeley, CA: McCutchan.

Clayton, M.J. (1997). Delphi: a technique to harness expert opinion for critical decision-making tasks in education. *Educational Psychology, 17* (4), 373-386.

Cochran-Smith, M., & Fries, M.K. (2001). Sticks, stones, and ideology: the discourse of reform in teacher education. *Educational Researcher, 30*, 3-15.

Cochran-Smith, M., & Lytle, S.L. (1999). Relationships of knowledge and practice: teacher learning in communities. In a. Iran-Nejad & C.D. Pearson (Eds.), *Review of Research in Education* (Vol. 24, pp. 249-305). Washington DC: American Educational Research Association.

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

Collins, A. (1991). Portfolios for biology teacher assessment. *Journal of Personnel Evaluation in Education, 5*, 147-168.

Collins, A. (1992). Towards a design science of education. In E. Scanlon & T. O'Shea (Eds.), *New directions in educational technology* (pp. 15-22). Berlin: Springer.

Collins, A., Brown, J.S. and Newman, S.E. (1989) Cognitive apprenticeship: teaching the crafts of reading, writing, and mathematics. In: L.B. Resnick (Ed.), *Knowing, learning, and instruction: essays in honor of Robert Glaser* (pp. 453-494). Hillsdale: Erlbaum.

Commissie Toekomst Leraarschap (1993). *Een beroep met perspectief. De toekomst van het leraarschap* [A profession with prospects. The future of the teacher profession]. Amsterdam, The Netherlands: Commissie toekomst leraarschap.

Conway, M.A. (1990). Associations between autobiographical memories and concepts. *Journal of Experimental Psychology: Learning, Memory and Cognition, 16*, 799-812.

Cooksey, R.W. (1996). *Judgmental analysis: theory, methods, and application.* San Diego: Academic Press.

Crawley, F. E. (1990). Intentions of science teachers to use investigative teaching methods: A test of the Theory of Planned Behavior. *Journal of Research in Science Teaching, 27*, 685-697.

Crocker, L. (1997). Assessing content representativeness of performance assessment exercises. *Applied Measurement in Education, 10* (1), 83-95.

Cronbach, L. J., & Meehl P. I. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

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

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

Crooks, T.J. & Kane, M.T. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice, 3* (3), 265-285.

Cross, L. H., Impara, J. C., Frary, R .B. & Jaeger, R .M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher examination, *Journal of Educational Measurement, 21*, 113-130.

Darling-Hammond, L. (2002). Research and rhetoric on teacher certification: a response to "teacher certification reconsidered". *Educational Policy Analysis Archives, 10* (36). Retrieved from http: //epaa.asu.edu/epaa/v10nr36.html.

Darling-Hammond, L., Berry, B., & Thoreson, A. (2001). Does teacher certification matter? Evaluating the evidence. *Educational Evaluation and Policy Analysis, 23* (1), 57-77.

Day, C. (1999). *Developing teachers: the challenges of lifelong learning.* London: Falmer Press.

Day, D.V., & Sulsky, L.M. (1995). Effects of frame-of-reference training and information configuration on memory organization and rating accuracy. *Journal of Applied Psychology, 80*, 158-167.

De Corte, E. (2000). Marrying theory building and the improvement of school practice: a permanent challenge for instructional psychology. *Learning and Instruction, 10* (3), 249-266.

De Groot, A.D. (1961). *Methodologie* [Methodology]. *Grondslagen van onderzoek en denken in de gedragswetenschappen*. 's-Gravenhage: Mouton & co.

De Groot, A.D. (1970). Some badly needed non-statistical concepts in applied psychometrics. *Nederlands tijdschrift voor de psychologie, 25*, 360-376.

De Loe, R.C. (1995). Exploring complex policy questions using policy Delphi. A multi-round, interactive survey method. *Applied Geography, 15* (1), 53-68.

Delandshere, G. (1994). The assessment of teachers in the United States. *Assessment in education, 1* (1), 95-113.

Delandshere, G., & Arens, S.A. (2001). Representations of teaching and standards-based reform: are we closing the debate about teacher education? *Teaching and Teacher Education, 17*, 57-566.

Delandshere, G., & Petrosky, A. (1994). Capturing teachers' knowledge: Performance assessment (a) and post-structuralist epistemology, (b) from a post-structuralist perspective, (c) and post-structuralism, (d) none of the above. *Educational Researcher, 23*, 11-18.

Delbecq, A.L., Van de Ven, A.H., & Gustafson, D.H. (1975). *Group techniques for program planning: a guide to nominal group and Delphi processes*. Glenview: Foresman and Company.

Den Brok, P. (2001). *Teaching and student outcomes. A study on teachers' thoughts and actions from an interpersonal and a learning activities perspective.* Doctoral dissertation. Utrecht, The Netherlands: IVLOS.

DeNisi, A.S., Cafferty, T.P., & Meglino, B.M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Performance, 33*, 360-396.

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

Dochy, F., & Segers, M, (1999), Innovatieve toetstvormen als gevolg van constructiegericht onderwijs: op weg naar een assessment-cultuur [innovative assessment caused by construct based education: towards an assessmentculture]. In: M. Lacante & P. De Boeck (red.). *Meer kansen creëren voor het Hoger Onderwijs. Handboek Leerlingenbegeleiding*, Dordrecht: Kluwer, pp. 181-206.

Dolk, M. (1997). *Onmiddellijk onderwijsgedrag; over denken en handelen van leraren in onmiddellijke onderwijssituaties*. Doctoral dissertation. Utrecht, The Netherlands: WCC.

Drenth, P.J.D. & Sijtsma, K. (1990). *Testtheorie. Inleiding in de theorie van psychologische test en zijn toepassingen.* Houten/Diegem: Bohn Stafleu Van Loghum.

Dreyfus, H.L., & Dreyfus, S.E. (1986). *Mind over machine: the power of human intuition and expertise in the era of the computer*. Oxford: Basil Blackwell.

Duggan, S. & Gott, R. (1995). The place of investigations in practical work in the UK National Curriculum for Science. *International Journal of Science Education*, *17* (2), 137-147.

Dunbar, S.B., Koretz, D., & Hoover, H.D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*, 289-304.

Dwyer, C.A. (1994). Content standards for performance-based teacher assessment: validity, standards, and issues. *Journal of personnel evaluation in education, 8,* 135-150.

Dylan, W. (1996). *Construct-referenced assessment of authentic tasks: alternatives to norms and content standards*. Paper presented at the 24th Annual Conference of the International Association of Educational Assessment Testing and Evaluation: Confronting the Challenges of Rapid Social Change, Barbados, may 1998.

Eisner, E.W. (1991). *The enlightened eye. Qualitative inquiry and the enhancement of educational practice.* New York: Macmillan Publishing Company.

Elander, J. (2004). Student assessment from a psychological perspective. *Psychology Learning and Teaching, 3* (2), 34-41.

Embretson, S. (1993). Psychometric models of learning and cognitive processes, In N. Frederiksen, R.J. Mislevy, & I.I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 125-150). Hillsdale: Lawrence Erlbaum Associates.

Eraut, M. (2000). Non-formal learning and tacit knowledge in professional work. *British Journal of Educational Psychology, 70*, 113-136.

Eraut, M.E. (1994). *Developing professional knowledge and competence.* London: Falmer Press.

Erffmeyer, R.C., Erffmeyer, E.S., & Lane, I.M. (1986). The Delphi Technique: an empirical evaluation of the optimal number of rounds. *Group & Organization Studies, 11* (1-2), 120-128.

Ericsson, K.A., & Charness, N. (1994). Expert peformance: its structure and acquisition. *American psychologist, 49*, 725-747.

Ericsson, K.A., & Simon, H.A. (1980). Verbal reports as data. *Psychological review*, *87*, 215-251.

Ericsson, K.A., & Simon, H.A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Ericsson, K.A., & Smith, J. (1991). *Toward a general theory of expertise: prospects and limits*. New York: Cambridge University Press.

Fazio, R.H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 23, pp. 75-109). San Diego, CA: Academic Press.

Fazio, R.H. (1994). Attitudes in associated systems theory, In R.S. Wyer (Ed.), *Associated systems theory: a systematic approach to cognitive representations of persons. Advances in Social Cognitions,* (Vol. 7, pp. 157-167). Hillsdale, NJ: Lawrence Erlbaum Associates.

Fazio, R.H., & Olson, M.A. (2003). Implicit measures in social cognition research: their meaning and use. *Annual Review of Psychology*, 54: 297-327.

Feldman, J.M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66*, 127-148.

Fenstermacher, G.D. & Richardson, V. (2000). *On making determinations of quality in teaching.* A paper prepared at the request of the Board on International Comparative Studies in Education of the National Academy of Sciences, University of Michigan, Ann Arbor.

Fishbein, M.,Triandis, H.C., Kanfer, F.H., Becker, M., Middlestadt, S.E., & Eichler, A. (2001) Factors influencing behavior and behavior change. In A. Baum, T.A. Revenson & J.E. Singer (Eds.), *Handbook of health psychology* (pp. 3-17). Mahwah, NJ: Lawrence Erlbaum Associates.

Fiske, S.T. (1992). Thinking is for doing: portraits of social cognition from daguerreotype to laserphoto. *Journal of Personality and Social Psychology, 63*, 877-889.

Fitts, P.M., & Posner, M.I. (1967). *Human performance.* Belmont, CA: Brooks/Cole.

Fitzpatrick, A. R. (1989) Social influences in standard setting: the effects of social interaction on group judgments, *Review of Educational Research, 59* (3), pp. 315-328.

Fradd, S.H., & Lee, O. (1999) Teachers' roles in promoting science inquiry with students from diverse language backgrounds. *Educational Researcher, 28*, 6, 14-20.

Freidus, H. (1998). Mentoring portfolio development. In N. Lyons (Ed.), *With portfolio in hand. Validating the new teacher professionalism* (pp. 51-68). New York: Teachers College Press.

Gee, B. and Clackson, S.G. (1992). The origin of practical work in the English school science curriculum. *SSR, 73* (265), 79-83.

Gentile, C. (1992). *Exploring new methods for collecting students' school-based writing: NAEP's 1990 Portfolio Study*. Washington, DC: Office of Educational Research and Improvement.

Gilbert, D.T. (1989). Thinking lightly about others: Automatic components of the social inference process. In J.S. Uleman & J.A. Bargh (Eds.), *Unintended thought* (pp. 189-211). New York: Guilford.

Glaser, R., & Silver, E. (1994). *Assessment, testing and instruction: Retrospect and prospect.* (Center for the Study of Evaluation Tech. Rep. No. 379). Los Angeles : National Center for Research on Evaluation, Standards, and Student Testing.

Goldhaber, D.D. & Brewer, D.J. (2000). Does teacher certification matter? High school certification status and student achievement. *Educational evaluation and policy analysis, 22*, 129-145.

Gonczi, A. (1994). Competency based assessment in the professions in Australia. *Assessment in Education 1* (1), 27-45.

Gott, R., & Duggan, S. (1996). Practical work: its role in the understanding of evidence in science. *International Journal of Sciene Education 18* (7), 791-806.

Greatorex, J. & Dexter, T. (2000). An accessible analytic approach for investigating what happens between the rounds of a Delphi study. *Journal of Advanced Nursing, 32* (4), 1016-1024.

Grossman, P.L. (1990). *The making of a teacher. Teacher knowledge and teacher education.* New York: Teachers College Press.

Grossman, P.L. (1995). Teachers' knowledge. In Anderson, L.W. (Ed.), *International Encyclopaedia of teaching and teacher education. 2nd ed* (pp. 20-24). Oxford: Pergamon.

Haertel, E.H. (1999). Performance assessment and educational reform. *Phi Delta Kapan, 80* (9), 662-666.

Hager, P. & Gonczi, A. (1994). General issues about assessment of competence. *Assessment & Evaluation in Higher Education, 19* (1) , 3-16.

Hager, P., Gonczi, A., & Athanasou, J. (1994). General issues about assessment of competence. *Assessment and evaluation in higher education, 19* (1), 3-16.

Hambleton, R. K. & Plake, B. S. (1995) Using an extended Angoff procedure for setting standards on complex performance assessments, *Applied Measurement in Education, 8,* pp. 41-55.

Hambleton, R.K. (1997). Standard setting in criterion-referenced tests, In J.P. Keeves (Ed.), *Educational research, methodology, and measurement: an international handbook, 2nd ed.* (pp. 798-857). Oxford: Elsevier Science Ltd.

Hammond, K.R. (1980). *The integration of research in judgment and decision theory (Report No. 226).* Boulder: Center for research on judgment and policy, University of Colorado.

Haney, J. J., Czerniak, C. M., & Lumpe, A. T. (1996). Teacher beliefs and intentions regarding the implementation of science education reform strands. *Journal of Research in Science Teaching, 33*, 971-993.

Harvey, R.J., & Wilson, M.A. (2000). Yes Virginia, there is an objective reality in job analysis. *Journal of Organizational Behavior*, 829-854.

Hodson, D. (1992). Assessment of practical work. *Science & Education 1*, 115-144.

Hogan R.T. (1991). Personality and personality measurement. In M.D. Dunnette & L.M. Hough LM, *Handbook of industrial and organizational psychology, 2nd ed*, (Vol 2, pp. 873-919). Palo Alto: Consulting Psychologists Press.

Huberman, M. (1995). *Professional careers and professional development: some intersections*. In T.R. Guskey & M. Huberman (Eds.), Professional development in education: new paradigms and practices (pp. 193-224). New York: Teacher College Press.

Huot, B.A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M.M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment. Theoretical and empirical foundations* (pp. 206-232). Cresskill, NJ: Hampton Press.

Ingvarson, L. (1998). Teaching standards: foundations for professional development reform. In A. Hargreaves, A. Lieberman, M. Fullan, & D. Hopkins (Eds.), *International handbook of educational change.* Part two (pp. 1006-1031). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Ingvarson, L. (2002). Strengthening the profession? A comparison of recent reforms in the UK and the ISA. *Policy Briefs, 2,* 1-16. Camberwell, Australia: Australian council for educational research.

Jaeger, R. M. (1990) Setting standards on teacher certification tests, in: J. Millman & L. Darling-Hammond (Eds), *The new handbook of teacher evaluation assessing elementary and secondary school teachers* (pp. 295-321). Newbury Park: Sage Publications.

Jaeger, R. M. (1995) Setting performance standards through two-stage judgmental Policy capturing, *Applied Measurement in Education, 8* (1), pp. 15-40.

Johnson, R.L., McDaniel, F., & Willeke, M.J. (2000). Using portfolios in program evaluation: An investigation of interrater reliability. *The American Journal of Evaluation, 21*, 65-80.

Jones, E.E., & Davis, K.E. (1965). From acts to dispositions: the attribution process in person perception. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology, 2* (pp. 219-266). New York: Academic Press.

Kagan, D.M. (1990). Ways of evaluating teacher cognition: inferences concerning the Goldilocks Principle. *Review of Educational Research, 60* (3), 419-469.

Kagan, D.M. (1992). Professional growth among preservice and beginning teachers. *Review of Educational Research, 62* (2), 129-169.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin 112* (3), 527-535.

Kane, M.T. (1994) Validating the performance standards associated with passing scores, *Review of Educational Research, 64* (3), pp. 425-461.

Kanselaar, G., Galen, F., van, Beemer, H., Erkens, G., & Gravemeijer, K. (1999). *Grafieken leren met de computer* [learning graphics by computer]. Utrecht, The Netherlands: ICO-ISOR, Utrecht University.

Katz, L.G., & Raths, J.D. (1985). Dispositions as goals for teacher education. *Teaching & Teacher Education,* 1 (4), 301-307.

Keeves, J.P. (1994). *Methods of assessment in schools*. In T. Husén & N. Postlethwaite (Eds), International encyclopedia of education, 2nd ed. (pp. 362-370). New York: Pergamon.

Kelly, A.E. (2003). Research as design. *Educational researcher, 32* (1), 3-4.

Kelly, G.A. (1955). *The psychology of personal constructs*. New York: Norton.

Kirschner, P.A. (1992). Epistemology, practical work and academic skills in science education. *Science & Education, 1*, 273-299.

Klarus, R. (1998). *Competenties erkennen: een studie naar modellen en procedures voor leerwegonafhankelijke beoordeling van beroepscompetenties* [Accreditation of competence: a study on outcomes-based models and procedures for accreditation of prior learning]. Doctoral dissertation. Nijmegen, The Netherlands: Nijmegen University.

Kolk, N.J. (2001). *Assessment centers. Understanding and improving construct-related validity*. Doctoral dissertation. Amsterdam, The Netherlands: Kurt Lewin Instituut.

Koretz, D., Klein, S., McCaffrey, D., & Stecher, B. (1992). *The reliability of scores from the 1992 Vermont portfolio assessment program*. Washington, DC: RAND Institute on Education & Teaching.

Korthagen, F. (1998). *Leraren leren leren* [Teachers learning to learn]. *Realistisch opleidingsonderwijs, geïnspireerd door Ph. A. Kohnstamm.* Inaugural lecture. Amsterdam:Vossiuspers AVP.

Kwakman, K. (1999). *Leren van docenten tijdens de beroepsloopbaan* [Teacher learning during their careers]. Doctoral dissertation. Nijmegen, The Netherlands: Nijmegen University.

Landy, F.J., & Farr, J.L. (1980). Performance rating. *Psychological Bulletin, 87*, 72-107.

Lee, O., & Fradd, S. H. (1998). Science for all, including students from non-English language backgrounds. *Educational Researcher, 27* (3), 1-10.

Leinhardt, G., (1988). Situated knowledge and expertise in teaching. In J. Calderhead (Ed.), *Teachers' professional learning* (pp. 146-168). London: The Falmer Press.

LeMahieu, P., Gitomer, D., & Eresh, J. (1995). Portfolios in large-scale assessment: Difficult but not impossible. *Educational Measurement: Issues and Practice, 14*, 11-16, 25-28.

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology*, *86*, 255-264.

Linn, R.L. (1994). Performance assessment. Policy promises and technical measurement standards. *Educational Researcher, 23* (9), 4-14.

Linn, R.L., Baker, E.L. & Dunbar, S.B. (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher, 20* (8), 15-21.

Linstone, H.A. & Turoff, M. (Eds.) (1975). *The Delphi method: techniques and applications,* (pp. 37-71). London: Addison-Wesley Publishing Company.

Little, J. (1990) The persistence of privacy. Autonomy and initiative in teachers' professional relations. *Teachers College Record, 91* (4), 509-536.

Lock, R. (1989). Assessment of practical skills. Part 1. The relationships between component skills. *Research in Science & Technological Education, 7* (2), 221-233.

Lock, R. (1990a). Assessment of practical skills. Part 2. Context dependency and construct validity. *Research in Science & Technological Education, 8* (1), 35-52.

Lock, R. (1990b) Open-ended, problem-solving investigations. What do we mean and how can we use them? *SSR, 71* (256), 63-72.

Long, D.L., & Bourgh, T. (1996). Thinking aloud: telling a story about a story. Commentary. *Discourse Processes, 21*, 329-339.

Loughran, J. & Corrigan, D. (1995). Teaching portfolios: A strategy for developing learning and teaching in preservice education. *Teaching and Teacher Education, 11*(6), pp. 565-577.

Lyons, N. (Ed.), (1998). *With portfolio in hand. Validating the new teacher professionalism*. New York: Teachers College Press.

Martindale, C. (1991). *Cognitive psychology: A neural-network approach*. Pacific Grove, CA: Brooks/Cole.

Meester, M.A.M., & Kirschner, P.A. (1995). Practical work at the open university of the Netherlands. *Journal of Science Education and Technology, 4* (2), 127-140.

Mehrens, W. A. (1990) Combining evaluation data from multiple sources. In: J. Millman & L. Darling-Hammond (Eds.), *The new handbook of Teacher Evaluation. Assessing elementary and secondary school teachers* (pp. 322-334). Newbury Park: Sage Publications.

Meijer, P.C. (1999). *Teachers' practical knowledge: teaching reading comprehension in secondary education.* Doctoral dissertation. Leiden, The Netherlands: Leiden University.

Messick, S. (1975). The standard problem: meaning and values in measurement and evaluation. *American Psychologist, 30*, 955-966.

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

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

Messick, S. (1995). Validity of psychological assessment. Validation of inferences form persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50 (9), 741-749.*

Millman, J. & Greene, J. (1989). The specification and development of tests of achievement and ability. In R.L. Linn (Ed.), *Educational Measurement 3rd ed.* (pp. 335-366). New York: Macmillan.

Ministerie van Onderwijs en Wetenschappen (1993). *Vitaal leraarschap* [A vigorous teaching profession]. *Beleidsreactie naar aanleiding van het rapport 'Een beroep met perspectief' van de Commissie Toekomst Leraarschap* Zoetermeer, The Netherlands: Ministerie van onderwijs en wetenschappen.

Ministerie van Onderwijs, Cultuur en Wetenschap (1998). *Verder met vitaal leraarschap* [Continuing a vigorous teaching profession]. Zoetermeer, The Netherlands: Ministerie van onderwijs en wetenschappen.

Mislevy, R.J., Steinberg, L.S, & Almond, R.G. (1999). *On the Roles of Task Model Variables in Assessment Design. CSE Technical Report 500.* Princeton, New Jersey: Educational Testing Service.

Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62* (3), 229-258.

Moss, P.A. (1994). Can there be validity without reliability? *Educational Researcher, 23* (2), 5-12.

Murphy, K.R., & De Shon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*, 873-900.

Nedermeijer, J., & Pilot, A. (2000). *Beroepscompetenties en academische vorming in het hoger onderwijs [Vocational competences and academic education in higher education].* Groningen, The Netherlands: Wolters-Noordhoff.

Newell, A., & Simon, H.A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Norcini, J.J., & Shea, J.A. (1997). The credibility and comparability of standards. *Review of Research in Education, 17*, 3-29.

Norcini, J.J., Lipner, R.S., Langdon, L.O., & Strecker, C.A. (1987). A comparison of three variations on a standard-setting method. *Journal of Educational Measurement, 24* (1), 56-64.

Norman, D.A. (1985). Human information processing: the conventional view. In A.M. Aitkenhead & J.M. Slack (Eds.), *Issues in cognitive modelling* (pp. 309-336). Hillsdale, NJ: Erlbaum.

Novak, J.R., Herman, J.L. & Gearhart, M. (1996). Establishing validity for performance-based assessments: An illustration for collections of student writing. *The Journal of Educational Research 89* (4), 220-233.

Nunally, J. (1978). *Psychometric theory. 2nd ed*. New York: McGraw-Hill.

Pajares, M.R. (1992). Teachers' beliefs and educational research: clearing up a messy construct. *Review of educational research, 62* (3), 307-332.

Paulson, F.L., & Paulson, P.B. (1994). *Assessing portfolios using the constructivist paradigm.* Paper presented at the annual meeting of the American Educational Research Association. New Orleans, LA. April 4-8. (ERIC Document Reproduction Service No. 376209).

Peterson, K.D. (1995). *Teacher evaluation: a comprehensive guide to new directions and practices.* Thousand oaks, CA: Corwin Press.

Plake, B.S., Melican, G.J., & Mills, C.N. (1991). Factors influencing intrajudge consistency during standard-setting. *Educational Measurement: Issues and Practice, 10* (2), 15-16, 22, 25.

Popping, R. (1983). *Overeenstemmingsmaten voor nominale data [Computing agreement on nominal data]*. Doctoral dissertation. Groningen, The Netherlands: Rijksuniversiteit Groningen.

Pula, J.J., & Huot, B.A. (1993). A model of background influences on holistic raters. In M.M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment. Theoretical and empirical foundations* (pp. 237-265). Cresshill, NJ: Hampton Press.

Putnam, R.T., & Borko, H. (1997). Teacher learning: implications of new views of cognition. In B.J. Biddle, T.L. Good, & I.F. Goodson (Eds.), *International handbook of teachers and teaching, 2,* (pp. 1223-1296). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Putnam, S.E., Pence, P. & Jaeger, R.M. (1995). A multi-stage dominant profile method for setting standards on complex performance assessments. *Applied Measurement in Education, 8* (1), 57-83.

Reckase, M.D. (1995). Portfolio assessment: A theoretical estimate of score reliability. *Educational measurement: Issues and Practice, 14* (1), 12-14.

Reeves, T.C., & Okey, J.R. (1996). Alternative assessment for constructivist learning environments. In B.G. Wilson (Ed.), *Constructivist learning environments: Case studies in instructional design* (pp. 191-202). Englewood Cliffs, NJ: Educational Technology Publications.

Reynolds, A. (1992). Getting to the core of the apple: a theoretical view of the knowledge base of teaching. Journal of Personnel Evaluation in Education, 6, 41-55.

Riggs, W.E. (1983). The Delphi Technique. An experimental evaluation. Technological forecasting and social change, 23, 89-94.

Rijborz, D. (2003). *Leren onderzoeken* [Learning how to do research]. Doctoral dissertation. Amsterdam, The Netherlands: Onderwijscentrum Vrije Universiteit Amsterdam.

Roehrig, G.H. and Luft, J.A. (2004) Constraints experienced by beginning secondary science teachers in implementing scientifc inquiry lessons. *International Journal of Science Education 26* (1), 3-24.

Roth, W.M, & Roychoudhury, A. (1993). The development of science process skills in authentic contexts. *Journal of Research in Science Teaching, 30,* 2, 127-152.

Rowe, K.J. & Hill, P.W. (1996). Assessing, recording and reporting students' educational progress: the case for 'subject profiles'. *Assessment in Education 3* (3), 309-352.

Russo, J.E., Johnson, E.J., & Stephens, D.L. (1989). The validity of verbal protocols. *Memory & Cognition, 17*, 759-769.

Ryan, J.M. & Kuhs, T.M. (1993). Assessment of preservice teachers and the use of portfolios. Theory into Practice, 32, 75-81.

Sackman, H. (1975). Delphi Critique. Expert opinion, forecasting, and group processes. Massachusetts: Lexington Books.

Sandberg, J. (1994). Human competence at work – an interpretative approach. Doctoral dissertation. Gotenberg, Sweden.

Saroyan, A., & Amundsen, C. (2001). Evaluating university teaching: Time to take stock. *Assessment and Evaluation in Higher Education,* 26 (4), 337-349.

Saxe, G. B. (1991). Culture and cognitive development: Studies in mathematical understandings. Hillsdale, NJ: Lawrence Erlbaum.

Scheele, D.S. (1975). Reality construction as a product of Delphi interaction. In H.A. Linstone & M. Turoff (Eds.), The Delphi method: techniques and applications, (pp. 37-71). London: Addison-Wesley Publishing Company.

Schleicher, D.J., & Day, D.V. (1998). A cognitive evaluation of frame-of-reference rater training: Content and process issues. *Organizational Behavior and Human Decision Processes, 73*, 76-101.

Schoenfeld, A.H. (1998). Toward a theory of teaching-in-context. Issues in education, 4 (1), 1-94.

Scriven, M. (1988). Evaluating teachers as professionals. (ERIC Document Reproduction Service No. ED300882).

Secretary of state for education and employment (1998). *Teachers: meeting the challenge of change.* London: Department for education and employment.

Seldin, P. (1997). *The teaching portfolio: a practical guide to improved performance and promotion/tenure decisions.* Bolton, MA: Anker.

Shapley, K.S., & Bush, M.J. (1999) Developing a valid and reliable portfolio assessment in the primary grades: building on practical experience. *Applied Measurement in Education, 12,* 111-132.

Shavelson, R.J., Baxter, G.P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215-232.

Shuell, T.J. (1990). Phases of meaningful learning. *Review of Educational Research*, *60*, 531-547.

Shulman, L.S. (1986). Those who understand: knowledge growth in teaching. Educational Researcher, February, 4-14.

Shulman, L.S. (1987). Knowledge and teaching: foundations of the new reform. Harvard Educational Review, 57 (1), 1-22.

Shulman, L.S. (1996). Paradigms and research programs in the study of teaching: a contemporary perspective. In M.C. Wittrock (Ed.), *Handbook of research on teaching 3rd ed.* (pp. 3-36). New York: Macmillan.

Smith, K., & Tillema, H. (2001). Long-term influences of portfolios on professional development. *Scandinavian Journal of Educational Research, 45* (2), 183-203.

Smith, K.S., & Simpson, R.D. (1995). Validating teaching competences for faculty members in higher education: a national study using the Delphi method. Innovative Higher Education, 19 (3), 223-234.

Smith, W.L. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement technique. In Williamson, M. M., & Huot, B.A. (1993). *Validating holistic scoring for writing assessment. Theoretical and empirical foundations.* (pp. 142-205). Cresskill, NJ: Hampton Press.

Smits, T.J.M. (2003). *Werken aan kwaliteitsverbetering van leerlingonderzoek* [Improving the quality of student research]. *Een studie naar de ontwikkeling en het resultaat van een scholing voor docenten*. Doctoral dissertation. Nijmegen, The Netherlands: Nijmegen University.

Snow, R.E. and Lohman, D.F. (1989) Implications of cognitive psychology for educational measurement. In: R.L. Linn (Ed.), *Educational measurement* (pp. 263-331). New York: Macmillan.

Snyder, J., Lippincott, A., & Bower, D. (1998). Portfolios in teacher eduction: technical or transformational. In N. Lyons (Ed.), *With portfolio in hand* (pp. 123-142) Columbia: Teachers College, Columbia University

Steenstra, C. (2003). *Geography's contribution to general education. An international comparative study.* Doctoral dissertation. Delft, The Netherlands: Eburon.

Sternberg, R.J. (1996). Costs of expertise. In K.A. Ericsson (Ed.), *The road to excellence* (pp. 347-354). Mahwah, NJ : Lawrence Erlbaum Associates.

Sternberg, R.J. (1999). Intelligence as developing expertise. *Contemporary Educational Psychology, 24*, 359-375.

Sternberg, R.J., & Hovarth, J.A. (1995). A prototype view of expert teaching. *Educational Researcher, 24* (6), 9-17.

Stokking, K.M. (2005). *Course material for a course on innovative assessment on the ICO International Summer School 2005 for PhD-students.* Cyprus: University of Nicosia.

Stokking, K.M. & Van der Schaaf, M.F. (2000). *Assessing teacher competences related to teaching students research skills.* Rewarded research proposal Dutch organisatisation for scientific research, NWO-PROO 411-21-204. Utrecht, The Netherlands: Utrecht University, Department of Educational Sciences.

Stokking, K.M. & Van der Schaaf, M.F. (2000). *Ontwikkeling en beoordeling van onderzoeksvaardigheden* [Development and assessment of research skills]. Utrecht, The Netherlands: Utrecht University, Department of Educational Sciences.

Stokking, K.M. & Van der Schaaf, M.F. (2004). *Development and assessment of cross-curricular skills using a student portfolio.* Rewarded research proposal Dutch organisation for scientific research, NWO PROO 411-03-119. Utrecht, The Netherlands: Utrecht University, Department of Educational Sciences.

Stokking, K.M., Van der Schaaf, M.F., Jaspers, J., & Erkens, G., (2004). *Teachers' assessment of students' research skills. British educational research journal, 30* (1), 93-116.

Stokking, K., & Voeten, R. (2000). Valid classroom assessment of complex skills. In R. Simons, J.L van der Linden and T. Duffy (Eds.), *New Learning* (pp. 101-118). Boston: Kluwer Academic Publishers.

Straetmans, J.J.M. & Sanders, P.F. (2001). *Beoordelen van competenties van docenten* [Assessing teacher competences]. Utrecht, The Netherlands: Programmamanagement Educatief Partnerschap.

Swanborn, P.G. (1999). *Evalueren. Het ontwerpen, begeleiden en evalueren van interventies: een methodische basis voor evaluatie-onderzoek* [Evaluation. Designing, coaching and evaluating interventions: a methodological basis for evaluation-research]. Amsterdam, The Netherlands: Boom.

Thair, M., & Treagust, D.F. (1997). A review of teacher development reforms in Indonesian secondary science: The effectiveness of practical work in biology. *Research in Science Education, 27* (4), 581-597.

Tillema, H.H. (1998). Assessment of potential, from assessment centres to development centres. *International Journal of Selection and Assessment, 6* (3), 185-191.

Tomlinson, P. (1995). Can competence-profiling work for effective teacher preparation? Part II: pitfalls and principles. *Oxford Review of Education, 21* (3), 299-314.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization and memory* (pp. 381-403). New York: Academic Press.

Tulving, E. (1983). *Elements of episodic memory*. Oxford: Clarendon Press.

Turoff, M. (1975). The Policy Delphi. In H.A. Linstone, & M. Turoff (Eds.), The Delphi method: techniques and applications, (pp. 84-100). London: Addison-Wesley Publishing Company.

Uhlenbeck, A. (2002). *The development of an assessment procedure for beginning teacher of English as a foreign language.* Doctoral dissertation. Leiden, The Netherlands: ICLON Leiden University.

Uhlenbeck, A., Beijaard, D., & Verloop, N. (2003). Requirements for an assessment procedure for beginning teachers: implications from recent theories on teaching and assessment. *Teachers College Record. 104* (2), 242-272.

Van Driel, J.H., Verloop, N., & De Vos, W. (1998). Developing science teachers' pedagogical content knowledge. *Journal of research in science teaching, 35* (6), 673-695.

Van de Pligt, J. (1997). Judgment and decision making, in: G.R. Semin & K. Fiedler (Eds), *Applied social psychology*, (pp. 30-64). London: Sage Publications.

Van den Akker, J. (1999). Principles and methods of development research. In J. van den Akker, R.B. Branch, K. Gustafson, N.M. Nieveen, & T. Plomp (Eds.), *Design approaches and tools in education and training* (pp. 1-14). Dordrecht, The Netherlands: Kluwer.

Van der Schaaf, M., Veenhoven, J. & Stokking, K. (2003). *Ontwikkelen van onderzoeksvaardigheden. Studies in de gammavakken van de tweede fase havo/vwo naar het ontwikkelen van onderzoeksvaardigheden bij leerlingen door docenten* [Developing research skills]. Paper Vereniging voor Onderwijsresearch conference, November 26, 2003, Nijmegen, The Netherlands.

Van der Schaaf, M.F. (2000). *Vaardig in het begeleiden van zelfstandig leren? Monitoringsinstrument* [Competent in coaching self-regulated learning? Monitoringinstrument]. Utrecht, The Netherlands: Utrecht University, Department of Educational Sciences.

Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (2003). Developing performance standards for teacher assessment by Policy capturing. *Assessment & Evaluation in Higher Education, 28*, 395-410.

Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (2005a). Cognitive representations in raters' assessment of teacher portfolios. *Studies in Educational Evaluation, 31*, 27-55.

Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (2005b). De invloed van cognitieve representaties van beoordelaars op hun beoordeling van docentportfolio's [The influence of raters' cognitive representations in assessing teacher portfolios]. *Pedagogische Studiën, 8 (1)*, 7-26.

Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (2005c). Developing content standards for teaching students research skills using a Delphi method. Manuscript submitted for publication.

Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (2005d). Designing teacher portfolios using a chain model of construct validation. Manuscript submitted for publication.

Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (2005e). Teacher beliefs and teacher behaviour in portfolio assessment. Manuscript submitted for publication.

Van Rens, E.M.M., & Dekkers, P.J.J.M. (2000). Leren onderzoeken - de rol van de docent [Learning to inquire – the role of the teacher]. *Tijdschrift voor Didactiek der Beta-wetenschappen 17* (1), 76-94.

Van Tartwijk, J, Pilot, A. & Wubbels, T. (2000). Naar een digitaal portfolio [Heading towards a digital portfolio]. In. R. Bosker (Ed.). *Onderwijskundig Lexicon editie III, Evalueren in het onderwijs* (pp. 41-58). Alphen aan den Rijn, The Netherlands: Samson.

Van Tilburg, P.A. & Verloop, N. (2000). Kennis van en opvattingen over het onderwijzen van onderzoeksvaardigheden [Knowledge of and opinions on the teaching of inquiry skills]. *Tijdschrift voor Didactiek der Beta-wetenschappen 17* (1), 60-75.

Vankan, L., van Nijnatten, H., & Ankoné, H. (1999) Model voor het maken van praktische opdrachten [Model for developing practical assignments]. *Geografie educatief, 2*, 42-44.

Veenhoven, J. (2004). *Begeleiden en beoordelen van leerlingonderzoek [Teaching and assessing students' research]. Een interventiestudie naar het leren ontwerpen van onderzoek in de tweede fase bij aardrijkskunde.* Doctoral dissertation. Utrecht, The Netherlands: ICO.

Velde, C. (1997). Crossing borders: an alternative conception of competence and implications of professional practice in the workplace. 27th Annual SCUTREA conference proceedings. (pp. 460-464), Queensland, Australia: Queensland university of technology.

Verloop, N. (1992). Praktijkkennis van docenten: een blinde vlek in de onderwijskunde [Teachers' practical knowledge: A blind spot in educational theory]. *Pedagogische Studiën, 69* (6), 410-423.

Verloop, N. (1995). De leraar [The teacher]. In J. Lowyck, & N. Verloop (Eds.), *Onderwijskunde: een kennisbasis voor professionals* (pp. 108-150). Groningen, The Netherlands: Wolters-Noordhoff.

Verloop, N., Van Driel, J., & Meijer, P. (2001). Teacher knowledge and the knowledge base of teaching. *International Journal of Educational Research, 35* (5), 441-461.

Vermetten, Y, Daniels, J., & Ruijs, L. (2000). *Inzet van assessment: waarom, wat , hoe, wanneer en door wie?* [Using assessment: why, what, how, when, by who?]. *Beslismodel voor een beargumenteerde keuze van assessmentvormen in onderwijs en opleiding.* Heerlen, The Netherlands: Onderwijstechnolgisch expertisecentrum, Open Universiteit.

Vermunt, J.D., & Verloop, N. (1999). Congruence and friction between learning and teaching. *Learning and instruction, 9*, 257-280.

Wade R. C., & Yarbrough, D.B. (1996). Portfolios: A tool for reflective thinking in teacher education? *Teaching & Teacher Education 12* (1), 63-79.

Walsh, K. (2001). *Teacher certification reconsidered: stumbling for quality*. Baltimore, MD: Abell Foundation.

Westera, W. (2001). Competences in education: a confusion of tongues. *Journal of Curriculum Studies, 33* (1), 75-88.

Wheeler, P., & Haertel, G. D. (1993). *Resource handbook on performance assessment and measurement: a tool for students, practitioners, and policymakers.* Berkeley: Owl Press.

White, B.Y., & Frederiksen, J.R. (1998). Inquiry, modeling , and metacognition: making science accessible to all students. *Cognition and Instruction 16* (1), 3-118.

Wiggins, G. (1989). A true test: toward more authentic and equitable assessment. *Phi Delta Kappan 70* (9), 703-713.

Wiggins, G. (1993). Assessment: authenticity, context, and validity. *Phi Delta Kappan 74* (november), 200-214.

Woehr, D.J., & Huffcutt, A.I. (1994). *Rater training for performance appraisal: A quantitative review. Journal of Occupational and Organizational Psychology, 67*, 189-205.

Wolf, A. (1995). *Competence-based assessment.* Milton Keynes: Open university press.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assess-

ment. *Review of Educational Research 17*, 31-74.

Wolf, K., & Dietz, M. (1998). Teaching portfolios: purposes and possibilities. *Teacher Education Quarterly, 25*, 9-22.

Wolfe, E.W. (1996). A report on the reliability of a large-scale portfolio assessment of language arts, mathematics and science. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY. ERIC ED 399 285.

Wolming, S. (1999). Validity issues in higher education selection: a Swedish example. *Studies in Educational Evaluation, 25*, 335-351.

Zwaan, R.A., & Brown, C.M. (1996). The influence of language proficiency and comprehension skill on situation-model construction. *Discourse Processes, 21*, 289-327.

Appendices

Appendix 1

Rater manual, an example

Content standard 5 (TEACH): Teaching students research skills

Score 5, the teacher satisfies the content standard completely: the teacher uses multiple teaching strategies that support the development of student understanding in conducting investigations, and challenges students to accept and share responsibility for their own learning (see content standard 4). He or she uses the strategies in an adequate manner. For example, if students collaborate, he or she recognizes and responds to students' diversity and encourage all students to participate fully in fulfilling the assignment. He or she encourages each student to share his or her ideas with the other members. The strategies he or she uses suit the phases of the investigations the students are working at. For example, in the phase of formulating a research question, he or she coaches students in brainstorming; in the phase of gathering information, he or she gives the students keywords for searching the internet; in the phase of data processing, he or she directs the students to work in groups. The teacher monitors students' understanding and the amount of direction given in the teaching/learning process. In order to gather data, he or she can observe students, listen to students, read their logs. He or she uses the data to improve his or her teaching practice.
Score 4, the teacher satisfies the content standard to a high degree.
Score 3, the teacher satisfies the content standard to a reasonable degree: the teacher uses, to a certain extent, teaching strategies that support the development of student understanding in conducting investigations, or challenges students to accept and share responsibility for their own learning. He or she uses the strategies in a reasonable manner. The strategies he or she uses may suit the phases of the investigations the students are working at. The teacher sometimes monitors students' understanding and the amount of direction given in the teaching/learning process.
Score 2, the teacher satisfies the content standard to a small degree.
Score 1, the teacher does not satisfy the content standard: the teacher does not use teaching strategies that support the development of student understanding in conducting investigations, and does not challenge students to accept and share responsibility for their own learning. The strategies he or she uses do not suit the phases of the investigations the students are working at. The teacher does not monitor students' understanding and the amount of direction given in the teaching/learning process.

Portfolio elements

The results of two interviews concerning teachers' practical knowledge and intentions regarding instructing and coaching students research skills.
Two video recordings of lessons in which the teacher instructs and coaches students doing research.
A research assignment that is central to the portfolio, including the learning goals of the assignment and the teacher's reasons for the assignment content and form.Student evaluations of the teacher.

Assessment instruction

1. Illustrate the content standard with significant portfolio evidence in the portfolio elements mentioned above.
2. Describe your interpretation of each kind of evidence.
3. Summarize your interpretation in a score on a 5-point scale using the anchor points above giving evidence from the video most weight.

Indication for weighting evidence in portfolio elements content standard TEACH

Appendix 2

Portfolio elements and content standards in the initial portfolio

Content standards Portfolio elements	GOAL	ASSIGN	MAN.	THINK	TEACH	CLIM.	ASSESS	REFL
1.Self-description	X							X
2.Series assignments	X							
3. Interviews			X	X	X	X		X
4. Videos			X	X	X	X		
5. Assignment	X	X	X		X		X	
6. Assessment							X	
7. Reflection								X

Appendix 3

Psychometric reports of the scales in the panellist questionnaire

The panellists answered on a five point scale from 1 = I strongly disagree to 5 = I strongly agree

	mean	sd	Rir
Scale 'Relevance of the portfolio elements'			
1. The portfolio elements are relevant for teaching students research skills	4,13	.35	.65
2. The portfolio elements are in congruence with the content standards	4,00	.76	.55
3. The portfolio elements reflect teachers thinking and doing in teaching students research skills	4,25	.46	.78
4. The portfolio elements fit the content standards	3,75	.89	.75
5. The portfolio reflects the salient tasks of teachers when teaching students research skills	4,37	.52	.65
6. Teachers who are less competent in teaching students research skills will score lower on portfolio assessment portfolio than teachers who are more competent	4,13	1.13	.46
7. The portfolio shows what a teacher is competent of teaching students research skills	4,38	.52	.82
8. The portfolio shows what a teachers does in teaching students research skills	4,25	.71	.52
Scale 'Usability of the portfolio'			
1. The information given by a portfolio is sufficient for raters to judge teacher's competences in teaching students research skills	4,00	.76	.57
2. The information given by a portfolio is sufficient for raters to give teachers feedback on their strong and weaker points in instructing, coaching and assessing research	4,38	.52	.87
3. The portfolio elements suit teacher's everyday practice	4,38	.52	.87
4. It is clear to teachers how to develop a portfolio	4,38	.52	.87
5. It is clear to teachers what is the content of the portfolio	4,25	.71	.92
6. It is clear to teachers what the assessment standards are	4,25	1.04	.69
7. It is clear to teachers how the assessment results are used	3,25	1.04	.50
Scale 'Fairness of the portfolio'			
1. The portfolio causes teachers to be judged incorrectly (recoded)	4,13	.64	.71
2. It is fair to use the portfolio model for teacher assessment	3,75	1.04	.85
3. It is fair to give teachers feedback on their teaching based on the portfolio elements	4,25	.71	.77
4. The portfolio does not put certain groups of teacher at a disadvantage	4,00	.92	.96

Statistics for the scales

	'Relevance of the portfolio elements'	'Usability of the portfolio'	'Fairness of the portfolio'
N (panellists)	8	8	8
n items	8	7	4
Mean	4,20	4,13	4,03
Variance	.24	.35	.56
Sd	.49	.59	.75
Skewness	.55	.17	-0.33
Kurtosis	-.27	-1.07	-1.64
Cronbach's alpha	.85	.89	.91

Appendix 4

Psychometric reports of the scales in the teacher questionnaire

The teachers answered on a five point scale from 1 = I strongly disagree to 5 = I strongly agree

	mean	sd	Rir
Scale 'Usability of the portfolio'			
1. Constructing the portfolio following the manual is doable	3,71	1.01	.82
2. Constructing the portfolio following the manual does not cost too much time	2,95	1.28	.85
3. Constructing the portfolio following the manual is difficult (recoded)	3,38	1.20	.78
4. A teacher portfolio is a time-consuming assessment instrument (recoded)	3,38	1.40	.67
5. The manual gives clear instructions for constructing a portfolio	4,10	.77	.53
Scale 'Representativeness of the portfolio'			
1. My portfolio is representative for the way I teach students research skills	4,24	.62	.51
2. The portfolio reflects what I think and do when teaching students research skills	4,29	.56	.65
3. The portfolio shows how I teach students research skills	3,95	.86	.40
Scale 'Process of preparing a portfolio'			
1. In the portfolio I am able to show how I teach students research skills	3,7	.86	.77
2. I prefer to modify my portfolio	3,8	1.32	.39
4. If the content of the portfolio was more structured, I would be more able to construct a portfolio that reflects my practice (recoded)	4,1	.91	.68
5. There are better methods to show my competences in teaching students research skills than a portfolio (recoded)	3,45	1.39	.64
6. All portfolio elements are needed to show the way I teach students research skills	3,6	1.02	.70
7. The portfolio reflects the salient tasks in teaching students research skills	4,1	.79	.67
8. The manual and coaching of the researcher have obstructed me to show what I think and do when teaching students research skills (recoded)	4,8	.41	.43

Statistics for the scales

	'Usability of the portfolio'	'Representativeness of the portfolio'	'Process of preparing a portfolio'
N (teachers)	20	21	20
n items	5	3	7
Mean	3,51	4,20	3,94
Variance	.90	.30	.51
Sd	.95	.54	.71
Skewness	.05	-.11	.015
Kurtosis	-.53	-.30	-.99
Cronbach's alpha	.88	.68	.83

Appendix 5

Psychometric report of the scale in the student questionnaire

The students answered on a four point scale: 1 'not, at all'; 2 'to a little degree'; 3 'to a reasonable degree'; 4 'to a high degree'.

items	mean	sd	Rir
1. The aim of the research assignment is clear	3,32	.77	.46
2. We have sufficient facilities to do the assignment	3,27	.76	.38
3. The assessment content standards are clear	3,26	.85	.32
4. The assignment is difficult (recoded)	2,88	.90	.25
5. My teacher gives practical instructions	3,11	.78	.52
6. The teacher regularly verifies whether we understand the assignment	2,74	.96	.53
7. My teacher gives clear instructions	3,24	.78	.55
8. My teacher is enthusiastic	3,43	.77	.52
9. My teacher is willing to explain something a second time	3,38	.79	.50
10. My teacher has faith in us	3,17	.85	.56
11. My teacher has high expectations of us	2,93	.79	.28
12. My teacher is interested in us	3,20	.78	.64
13. My teacher shows interest in our ideas	3,16	.83	.64
14. My teacher takes notice of contacts between students	2,91	1.91	.27
15. My teacher verifies whether we are actually working on the assignment	2,88	.95	.42

Statistics for the scale

N	=	317
n items	=	15
Mean	=	3,13
Variance	=	.79
Sd	=	.89
Skewness	=	-.26
Kurtosis	=	.97
Cronbach's alpha	=	.81

Appendix 6

Categories of comments

Dimension concrete-abstract (encoding per line)

1. Visual appearance (e.g. "She says …")
2. Categorizations (e.g. "That means …")
3. Personality traits (e.g. "She is an organized person.")

Dimension type of comment (encoding per line)

1. Negative (all sorts of negative comments)
2. Neutral (neutral or balanced feedback; e.g. "On the one hand [followed by a negative comment]; on the other hand [followed by a positive comment]")
3. Positive (all sorts of positive comments)
4. Tips and suggestions (e.g. "He rather should …")

Situation in which teachers operate (encoding per segment)

1. To place oneself in the teacher's situation proceeding from his/her own experiences and background*
2. Comments regarding the material and method used by the teacher. Taken into consideration while rating:
3. Students' characteristics
4. Time available to the teacher
5. Teacher's experience in teaching students research skills
6. Teacher's collaboration with colleagues*
7. The school policy
8. General expectations towards teachers
9. The possibility of bias due to the researcher's presence in the school context*

Raters' own judgment process (encoding per segment)

1. References to the manual while rating (e.g. "the manual says …")
2. Notifications about lack of information for giving an adequate assessment (e.g. "She says she gives feedback to her students, but in her portfolio I do not find any information about that.")
3. Explaining own judgment process (e.g. "I am rating in this way, because …")
4. Pitfalls and uncertainties while rating (e.g. "The difficulty is …")
5. Questioning own assessment process (to the researcher) (e.g. "Am I doing it right now?")*
6. Notifications and corrections of mistakes (e.g. "What I am doing now is not correct.")
7. References to earlier assessed portfolio material
8. References to earlier assessed portfolio(s)
9. Other metacognitive comments

Judgment procedure (encoding per segment)

Remarks (positive and negative) about:
1. Judgment criteria and standards
2. Portfolio material
3. Assessment procedure
4. Weighting of the content standards
5. Other

* This category does not occur in the judgment forms

Appendix 7

An example of a content standard description

Content standard 4 (THINK): Previously thinking of and selecting teaching strategies that support the development of students' research skills

Score 5, the teacher satisfies the content standard completely: the teacher chooses multiple teaching strategies that support the development of student understanding in conducting investigations, and challenges students to accept and share responsibility for their own learning.
The strategies he or she chooses suit the phases of the investigations the students are working at. For example, in the phase of formulating a research question, he or she chooses strategies to coach students in brainstorming; in the phase of gathering information, he or she chooses strategies that give the students keywords for searching the internet; in the phase of data processing, he or she chooses strategies that direct the students to work in groups.
The teaching strategies fit teachers' goals, the research assignment and the assessment method
Score 4, the teacher satisfies the content standard to a high degree.
Score 3, the teacher satisfies the content standard to a reasonable degree: the teacher chooses a few teaching strategies that support the development of student understanding in conducting investigations, and this challenges students to accept and share responsibility for their own learning.
The strategies he or she chooses may suit the phases of the investigations the students are working at.
The teaching strategies may fit teachers' goals, the research assignment and the assessment method in a reasonable manner.
Score 2, the teacher satisfies the content standard to a small degree.
Score 1, the teacher does not satisfy the content standard: the teacher does not choose teaching strategies that support the development of student understanding in conducting investigations, and does not choose strategies that challenge students to accept and share responsibility for their own learning. The strategies he or she chooses neither suit the phases of the investigations the students are working at, nor the teachers' goals, the research assignment or the assessment method.

Samenvatting

De toegenomen aandacht voor kwaliteitsbewaking en verantwoording in het onderwijs en voor docenten als beroepsgroep van professionals, gaat internationaal gepaard met onderzoek naar beoordeling van competenties van docenten (competenties opgevat als kennis, vaardigheden en houdingen die nodig zijn voor adequaat functioneren in een professie). Daarbij wordt gebruik gemaakt van nieuwe vormen van beoordelen (competence-based assessment), zoals met behulp van portfolio's. Tegelijkertijd wordt van docenten in alle onderwijssectoren een rolverandering gevraagd. Naast kennisoverdracht worden ook het begeleiden van meer zelfstandig lerende leerlingen en meer authentieke vormen van beoordelen belangrijker. Docentbeoordeling is al geruime tijd onderwerp van internationaal debat. Dit debat sluit aan bij ten minste vier actuele discussies, die onderling samenhangen.

Een eerste punt van discussie betreft de beoordelingscriteria (content standaarden) waarop docenten kunnen worden beoordeeld. Kernvragen zijn hoe deze beoordelingscriteria kunnen worden vastgesteld, wat deze beoordelingscriteria zijn en welke minimumeisen (performance standaarden) daarbij moeten gelden voor een voldoende beoordeling.

Ten tweede hebben ontwikkelingen in onderzoek naar docentcognities en expertise mede geleid tot het uitwerken van nieuwe vormen van assessment (competence-based), zoals portfoliobeoordeling. Een belangrijke vraag is hoe portfoliobeoordeling kan worden ingericht en via welke procedures een dergelijke inrichting kan worden ontwikkeld.

Ten derde zijn naast nieuwe vormen van assessment ook de kwaliteitscriteria waaraan deze moeten voldoen verder uitgewerkt. Er bestaat discussie over de vraag in hoeverre bij nieuwe vormen van beoordelen traditionele kwaliteitscriteria moeten worden losgelaten. De klassieke criteria betrouwbaarheid en validiteit zijn inmiddels meer gedifferentieerd en aangevuld met criteria op het vlak van aanvaardbaarheid en praktische bruikbaarheid.

Ten vierde hebben nieuwe opvattingen over leren, waarin het zelf actief construeren van kennis wordt benadrukt, grote gevolgen voor de rol van de docent. De docent moet meer dan voorheen leerlingen instrueren, begeleiden en beoordelen rond opdrachten met een open karakter, waaraan leerlingen zelfstandig en in samenwerking werken. Dit vergt van de docent deels nieuwe competenties, in termen van meer of minder contextgebonden cognities (kennis, vaardigheden en houdingen) en gedrag. Recent is daarbij veel aandacht ontstaan voor het belang van de kennisbasis, met name de praktijkkennis van docenten.

Dit proefschrift draagt primair bij aan de eerste en de tweede en secundair aan de vierde

discussie, mede door daarbij expliciet inzichten uit de derde discussie te betrekken. In het proefschrift worden methodieken en procedures ontwikkeld en uitgetest voor het beoordelen van (competenties van) docenten, nodig voor het instrueren, begeleiden en beoordelen van zelfstandig onderzoek door leerlingen (een toespitsing van de vierde discussie). Deze worden getoetst aan een uitgewerkte set psychometrische criteria (ontwikkeld op basis van de derde discussie). De discussies zijn verweven waar het gaat om de functie van de beoordeling (tweede discussie), de keuze van te beoordelen competenties (vierde discussie), en de te bereiken validiteit en betrouwbaarheid (derde discussie).

In dit proefschrift wordt verslag gedaan van een serie onderzoeken gericht op de beoordeling van competenties van ervaren docenten in de tweede fase voortgezet onderwijs met behulp van een docentportfolio, toegespitst op het instrueren, begeleiden en beoordelen van onderzoek van leerlingen in de gammavakken. Het proefschrift heeft als hoofddoel het ontwikkelen en beproeven van heldere procedures voor het beoordelen van competenties van docenten en is primair gericht op de summatieve beoordeling en secundair op de formatieve beoordeling van ervaren docenten in de tweede fase voortgezet onderwijs bij de gammavakken.
De onderzoeksvragen zijn:

1. Hoe kunnen de competenties van docenten die nodig zijn voor het instrueren, begeleiden en beoordelen van zelfstandig onderzoek door leerlingen worden beoordeeld?
2. Hoe kan de kwaliteit van zulke beoordelingen worden verbeterd?
3. Wat is de relatie tussen opvattingen en gedrag zoals geconstrueerd door docenten in hun portfolio en zoals beoordeeld door leerlingen en externe beoordelaars?

Het onderzoek is instrumenteel-nomologisch van aard (De Groot, 1961) en betreft een ontwikkelingsonderzoek (Van den Akker, 1999) gericht op de ontwikkeling van procedures voor de beoordeling van docentcompetenties met behulp van docentportfolio's. In de **hoofdstukken 3 en 4** worden beoordelingscriteria ontwikkeld met behulp van een Delphi methode (Linstone & Turoff, 1975) en een Policy capturing methode (Jaeger, 1995). In **Hoofdstuk 5** ontwikkelen we een portfolio format met behulp van een zogenaamd 'chain model' voor construct validering (Mislevy, et al., 1999). Tevens is het onderzoek descriptief van aard waar het gaat om de analyse van cognitieve representaties van beoordelaars in **Hoofdstuk 6**, en de analyse van docentopvattingen en docentgedrag in **Hoofdstuk 7**.

De **hoofdstukken 3 tot en met 5** geven een antwoord op de eerste onderzoeksvraag. Om te beginnen onderzochten we in een literatuurstudie welke taken docenten zouden moeten vervullen bij het aanleren van onderzoeksvaardigheden en over welke competenties ze daartoe zouden moeten beschikken. Vervolgens zijn we in een vragenlijstonderzoek nagegaan hoe de dagelijkse praktijk van het aanleren van onderzoeksvaardigheden er volgens docenten uit ziet (zie **Hoofdstuk 2**). Op basis daarvan ontwikkelden we een voorlopig overzicht van beoordelingscriteria, die we gebruikten als input voor het Delphi onderzoek in **Hoofdstuk 3**.
We gebruikten de Delphi methode bij het vaststellen van beoordelingscriteria, omdat deze methode een gestructureerde en iteratieve paneldiscussie mogelijk maakt en betrokkenen een belangrijke stem geeft. De selectie van deelnemers voor dit panel vond plaats via een gestructureerde intake. Het Delphi onderzoek is individueel en schriftelijk uitgevoerd. In drie ronden beoordeelden en herzagen 21 geselecteerde panelleden (vertegenwoordigers van de diverse belangengroepen) conceptvoorstellen van de beoordelingscriteria. Na elke

ronde kregen zij feedback van de onderzoekers over hun resultaten. Met deze methode ontwikkelden we met een hoge mate van support en consensus negen beoordelingscriteria voor het beoordelen van docentcompetenties bij het instrueren, begeleiden en beoordelen van onderzoek door leerlingen. De beoordeling door de panelleden van deze beoordelingscriteria werd in de opeenvolgende ronden steeds positiever. We hebben niet exact kunnen verklaren wat daaraan ten grondslag lag. Mogelijk heeft het iteratieve karakter met tussentijdse feedback ertoe bijgedragen dat panelleden het beoordelen van de beoordelingscriteria zich steeds meer eigen maakten. We concludeerden dat dit verschijnsel nog nader onderzoek vereist.
De beoordelingscriteria die zijn ontwikkeld luiden als volgt: 1. Lange termijn doelen hanteren (DOEL); 2. Geschikte onderzoeksopdracht kiezen (OPDR); 3. Het organiseren van het werken door leerlingen aan de onderzoeksopdracht (ORGA); 4. Nadenken over instructie en begeleiding (DENK); 5. Zelfstandig onderzoek door leerlingen instrueren en begeleiden (INSTR); 6. Pedagogisch klimaat scheppen (KLIM); 7. Beoordelen van het zelfstandig onderzoek van leerlingen (BEO); 8. Reflecteren op het onderwijsprogramma en het eigen handelen (REFL); 9. Samenwerken met collega's (SAMEN).
Een volgende (deel)vraag was hoe goed docenten aan de beoordelingscriteria moeten voldoen voor een voldoende beoordeling. Daartoe ontwikkelden 9 geselecteerde panelleden (de meeste van hen participeerden ook in het hiervoor beschreven Delphi onderzoek) performance standaarden met behulp van een Policy capturing studie (**Hoofdstuk 4**). We besloten beoordelingscriterium 9 (SAMEN) niet mee te nemen bij het ontwikkelen van performance standaarden, omdat dit criterium te sterk afhankelijk is van de werkcontext van de docent en in veel scholen samenwerken nog onvoldoende in de praktijk wordt gebracht (Stokking & Van der Schaaf, 2000; Van der Schaaf, Veenhoven & Stokking, 2003; Rijborz, 2003).
We redeneerden dat met name methodes gebaseerd op Policy capturing interessant zouden zijn voor competence-based assessments, omdat ze zowel gebruikt kunnen worden voor de ontwikkeling van performance standaarden als voor het verkrijgen van inzicht in een passend beoordelingsmodel (compensatorisch of conjunctief). Bovendien kunnen dergelijke methodes het bepalen van wegingen van beoordelingscriteria ondersteunen en het beoordelingsproces van panelleden stimuleren (Jaeger, 1995).
Na een training ontwikkelden de panelleden in twee ronden met tussenliggende groepsdiscussie met succes performance standaarden (met ankerpunten op een vijfpuntschaal). Hierbij moest ook worden aangegeven in welke mate de afzonderlijke beoordelingscriteria moeten meewegen in de totaalbeoordeling. De procedure bleek bruikbaar en resulteerde in significante en consistente regressiemodellen. De beoordelingen door de panelleden in de tweede ronde waren consistenter dan die in de eerste ronde. Hoewel de panelleden het eens waren over de performance standaarden behorende bij de acht beoordelingscriteria, verschilden ze van mening over de gewichten die aan de beoordelingscriteria zouden moeten worden toegekend. We konden niet precies achterhalen waarom op laatstgenoemde punt geen consensus kon worden bereikt.
Het doel van **Hoofdstuk 5** was om een ontwerp voor een docentportfolio te ontwikkelen, voor zowel summatieve als formatieve beoordelingsdoeleinden. Daarbij richtten we ons op twee verbanden tussen belangrijke componenten bij het ontwikkelen van een portfolio. Ten eerste het verband tussen de beoordelingscriteria en het portfolio-ontwerp. Ten tweede het verband tussen de beoordelingscriteria en het scoren door de beoordelaars.
Acht experts, die grotendeels ook participeerden in de studies in hoofdstukken 3 en 4

(waarbij ze beoordelingscriteria en performance standaarden ontwikkelden), namen deel aan het onderzoek in Hoofdstuk 5. De experts beoordeelden een voorlopig portfolio-ontwerp gebaseerd op de acht beoordelingscriteria en performance standaarden zoals geformuleerd in hoofdstukken 3 en 4. Het voorlopige portfolio-ontwerp was ook gebaseerd op de Theory of Planned Behavior (Ajzen, 1985; Ajzen & Fishbein, 2005). De experts gebruikten een Policy capturing methode (vergelijkbaar met Hoofdstuk 4) om de portfolio-onderdelen te wegen. We voerden een pilot uit waarin drie docenten een voorlopig portfolio-ontwerp uitprobeerden. De resultaten laten zien dat de experts het met elkaar eens waren over het voorlopige portfolio-ontwerp en dat zowel de experts als de docenten dachten dat het ontwerp voldoende gerelateerd was aan de beoordelingscriteria.
Vervolgens ontwikkelden 18 docenten een portfolio gebaseerd op het uiteindelijke portfolio-ontwerp en beoordeelden zes getrainde beoordelaars de docentportfolio's aan de hand van de acht beoordelingscriteria. Deze zes getrainde beoordelaars waren afkomstig uit de eerder genoemde groep van acht experts die betrokken waren bij de ontwikkeling van het portfolio-ontwerp. De resultaten laten zien dat zowel de experts als de docenten een voorkeur hadden voor artefacten en reproducties als portfolio-onderdelen en producties minder aantrekkelijk vonden. Desalniettemin vinden wij dat het, omwille van een valide portfolio-ontwerp, belangrijk is om producties op te nemen in een docentportfolio, omdat die mogelijk docentcognities kunnen representeren.
In het onderzoek beoordeelden zes beoordelaars paarsgewijs 18 portfolio's. Elk portfolio werd door twee beoordelaars onafhankelijk van elkaar beoordeeld. Bij 12 van deze portfolio's was de interbeoordelaarsbetrouwbaarheid redelijk tot goed. Het onderzoek wijst uit dat de beoordelaars hun beoordelingen redelijk sterk op de beoordelingscriteria baseerden. Echter, voor een meer theoretisch begrip van de gegeven beoordelingen is nader onderzoek gewenst naar docentcognities bij het samenstellen van een portfolio en naar de cognities van beoordelaars bij het scoren van docentportfolio's.

In **Hoofdstuk 6** behandelden we de vraag hoe de kwaliteit van docentportfoliobeoordelingen kan worden verbeterd. In dat hoofdstuk stelden we een heikel punt bij het gebruik van portfoliobeoordelingen centraal, namelijk de betrouwbaarheid van dergelijke beoordelingen. Portfoliobeoordelingen worden beïnvloed door cognitieve representaties van de beoordelaars. Inzicht daarin is cruciaal om de betrouwbaarheid van portfoliobeoordelingen te kunnen verbeteren. In Hoofdstuk 6 onderzochten we zowel de betrouwbaarheid van beoordelingen als de cognitieve representaties van beoordelaars zoals die tot uitdrukking komen in retrospectieve hardopdenkprotocollen en beoordelingsformulieren. Door beide aan elkaar te relateren konden we nagaan in hoeverre beoordelingen kunnen worden verklaard vanuit cognitieve representaties van beoordelaars. Zoals eerder aangegeven beoordeelden zes beoordelaars paarsgewijs 18 portfolio's en was de interbeoordelaarsbetrouwbaarheid bij 12 portfolio's redelijk tot goed. Variantie-analyse wees nauwelijks op beoordelaarseffecten. We gebruikten de 'Associated Systems Theory' (Carlston, 1992, 1994) en de 'Correspondent Inference Theory' (Jones & Davis, 1965) om de inhoud van de hardopdenkprotocollen en beoordelingsformulieren te analyseren. De beoordelingen konden grotendeels worden verklaard door categorisering van de cognitieve representaties van de beoordelaars op twee dimensies: 'abstracte versus concrete opmerkingen' en 'positieve versus negatieve evaluatie'. Dat is een belangrijk resultaat omdat het wijst op een aanwezige relatie tussen cognities van beoordelaars en (de kwaliteit van) hun beoordelingen. We concludeerden onder meer

dat de kwaliteit van docentportfoliobeoordelingen mogelijk kan worden verbeterd door in beoordelaarstrainingen explicieter aandacht te geven aan de genoemde dimensies.

In **Hoofdstuk 7** onderzochten we de vraag: wat is de relatie tussen docentopvattingen en gedrag zoals geconstrueerd door docenten in hun portfolio en zoals beoordeeld door leerlingen en externe beoordelaars? Om deze vraag te beantwoorden vergeleken we de inhoud van de verschillende docentportfolio's (inclusief zelfbeschrijvingen en videoregistraties van lessen) met de beoordelingen van leerlingen en beoordelaars. We maakten een onderscheid tussen gedrag en opvattingen zoals geconstrueerd door docenten (de opvattingen die ze expliciteren in hun portfolio's), en de beoordelingen van docentgedrag en docentopvattingen door de beoordelaars. We analyseerden de opvattingen en het gedrag van 18 docenten in hun portfolio's met behulp van de Theory of Planned Behavior (Ajzen, 1985; Ajzen & Fishbein, 2005), waarin geëxpliciteerde opvattingen en gedrag in een bepaalde context met elkaar worden verbonden. Verder werd elk portfolio beoordeeld door twee onafhankelijke beoordelaars en werd het gedrag van docenten in de klassensituatie beoordeeld door hun leerlingen in een vragenlijst (n = 317). Deze vragenlijst betrof vragen over hoe leerlingen vinden dat hun docenten onderzoeksvaardigheden instrueren en begeleiden.
Analyse van de inhouden van de 18 docentportfolio's toont op geaggregeerd niveau geen relatie tussen geëxpliciteerde cognities van docenten en hun gedrag. Wel konden de beoordelingen door de leerlingen van het gedrag van hun docent significant worden voorspeld vanuit de beoordelingen door de externe beoordelaars. Echter, de beoordelingen door de externe beoordelaars konden op zichzelf niet significant worden voorspeld vanuit de docentopvattingen en het docentgedrag zoals door ons geanalyseerd in de portfolio's.
We zochten naar verschillen tussen docenten met behulp van exploratieve Q-principale componenten-analyse in de twaalf portfoliobeoordelingen die voldoende betrouwbaar zijn beoordeeld. We vonden dat de docenten ingedeeld konden worden in twee groepen die significant van elkaar verschilden in de leerlingbeoordelingen volgens de eerder genoemde leerlingvragenlijst. De twee groepen verschilden ook in de beoordelingen van externe beoordelaars op het beoordelingscriterium 4 "Nadenken over instructie en begeleiding" (DENK). Docenten die, volgens de beoordelingen van de beoordelaars, nadenken over hun eigen handelen, hebben significant hogere scores op de leerlingvragenlijsten dan docenten met lage beoordelingen op dat criterium.
Hoewel we, gezien het beperkte aantal deelnemers aan het onderzoek, voorzichtig moeten zijn met het interpreteren van deze resultaten, genereren ze op zijn minst vervolgvragen ten aanzien van de rol van docentcognities bij docentportfoliobeoordeling.

Een ander belangrijk aandachtspunt bij docentportfoliobeoordeling is dat een dergelijke beoordeling gepaard gaat met complexe interacties tussen docentcompetenties, het portfolio, de gebruikte beoordelingscriteria, kenmerken van beoordelaars, en interpretaties van de beoordelaars. Als gevolg daarvan hebben portfoliobeoordelingen doorgaans een beperkte generaliseerbaarheid. Met het oog op het gebruik van docentportfoliobeoordelingen zijn generaliseerbaarheidstudies essentieel (Straetmans & Sanders, 2001). Deze zouden ten minste gericht moeten zijn op de facetten 'beoordelaars' en 'onderwijssituaties'.
Een constructvalide docentportfoliobeoordeling kan worden ondersteund door nadere theorievorming. Dergelijke theorievorming zou ten eerste betrekking moeten hebben op de te meten constructen, dat wil zeggen de conceptualisering van docentcompetenties en

de ontwikkeling van docentcompetenties in hun werkcontext. Onderzoek naar de aard en ontwikkeling van docentcompetenties in hun werkcontext is dan ook essentieel (cf. Kwakman, 1999; Bakkenes, Vermunt & Wubbels, 2004).
Tevens is nadere theorievorming nodig op het vlak van de cognities van panelleden en beoordelaars tijdens het beoordelen van portfolio's. In dit proefschrift hebben we gebruik gemaakt van verschillende theorieën die afkomstig zijn uit de sociaal cognitieve psychologie en die bruikbaar waren in de context van onze onderzoeken (bijvoorbeeld de Theory of Planned Behavior en de Associated Systems Theory). Een mogelijke eerste stap op weg naar het ontwikkelen van een sterkere theoretische fundering voor (onderzoek naar) portfoliobeoordelingen, is het synthetiseren van verschillende bruikbare theorieën in één overkoepelend theoretisch raamwerk.
Verder gaven we aan dat om portfoliobeoordelingen te verbeteren inzicht nodig is in hoe verschillende kenmerken van portfoliobeoordelingen (bijvoorbeeld het werken met een meer of minder gestructureerd portfolio, of met een meer of minder analytisch of holistisch scoringmodel) van invloed zijn op de kwaliteit van portfolio-beoordelingen. Quasi-experimentele designs waarin wordt gevarieerd op verschillende kenmerken zijn voor dit doel zinvol. Recent startten we een dergelijke quasi-experimenteel onderzoek naar het gebruik van leerlingportfolio's in de bovenbouw van het Vwo (Stokking & Van der Schaaf, 2004).

Curriculum vitae

Marieke van der Schaaf, geboren op 5 oktober 1974 te Hoorn, behaalde in 1993 het VWO diploma aan de Kalsbeekscholengemeenschap in Woerden. In hetzelfde jaar startte ze een studie onderwijskunde aan de Universiteit Utrecht. In 1994 behaalde ze het propaedeutisch examen onderwijskunde (cum laude). In 1997 behaalde ze het doctoraal examen onderwijskunde (cum laude). Gedurende haar studie werkte ze parttime bij verschillende werkgevers waarvoor ze hoofdzakelijk onderzoekswerk verrichtte (onder meer Procesmanagement Voortgezet Onderwijs te Den Haag en Inspectie van het Onderwijs te De Meern).
Vanaf 1998 is ze als onderwijskundig onderzoeker werkzaam bij de Afdeling Onderwijskunde aan de Universiteit Utrecht. Sinds 2002 is ze bij dezelfde organisatie ook werkzaam als universitair docent. In de periode 1998-2001 deed ze een zestal derde geldstroom onderzoeken (grotendeels kortlopend onderwijsonderzoek aangevraagd door het onderwijsveld), voornamelijk gericht op didactiek en beoordeling van algemene vaardigheden van leerlingen in het studiehuis (tweede fase voortgezet onderwijs). Ook voerde ze in die periode een NWO-PROO onderzoek uit naar het ontwikkelen en beoordelen van onderzoeksvaardigheden van leerlingen (nr. A575-35-007). In 1999 verwierf ze samen met prof. dr. Karel Stokking een NWO-PROO onderzoek waarover in dit proefschrift wordt gerapporteerd (nr. 411-21-204). In 2003 schreef ze, wederom samen met prof. dr. Karel Stokking, een gehonoreerd NWO-PROO voorstel naar de ontwikkeling van algemene vaardigheden bij leerlingen bij het werken met leerlingportfolio's (nr. 411-03-119). Momenteel is ze als post-doc onderzoeker op laatstgenoemd NWO-project en als gekwalificeerd universitair docent fulltime werkzaam aan de Afdeling Onderwijskunde van de Universiteit Utrecht.

Marieke van der Schaaf was born on October 5th 1974 in Hoorn, The Netherlands. She completed her pre-university education at the Kalsbeekscholen-gemeenschap in Woerden (1993). Next, she started to study Educational Sciences at Utrecht University, in which domain she graduated in 1997 (cum laude). During her study, she also worked as an educational research assistant for several Dutch educational organizations. Since 1998 she works as an educational researcher at the Department of Educational Sciences at Utrecht University. Since 2002 she is also a teacher at the same department.

List of publications and awarded research proposals

*) Partly or as a whole included in this thesis

* Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (2005). Cognitive Representations in Raters' Assessment of Teacher Portfolios. *Studies in Educational Evaluation, 31*, 27-55.

* Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (2005). De invloed van cognitieve representaties van beoordelaars op hun beoordeling van docentportfolio's [The influence of raters' cognitive representations in assessing teacher portfolios]. *Pedagogische Studiën, 8* (1), 7-26.

Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (2005). Assessing teacher beliefs in teacher portfolios. Paper Onderwijs Research Dagen, May 30 - June the 1st, 2005 Gent, Belgium.

* Stokking, K.M., Van der Schaaf, M.F., Jaspers, J., & Erkens, G. (2004). Teachers' assessment of students' research skills. *British Educational Research Journal, 30*(1), 93-116.

Van der Schaaf, M., Stokking, K., & Verloop, N. (2004). Raters' cognitive representations in assessing teacher portfolios. Paper Onderwijs Research Dagen June 9-11, 2004 Utrecht, The Netherlands.

Stokking, K.M., Van der Schaaf, M.F., Leenders, F., & De Jong, J (2004). Meten van reflectie bij studenten [Measurement of students' reflection]. Paper Onderwijs Research Dagen June 9-11, 2004 Utrecht, The Netherlands.

* Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (2003). Developing performance standards for teacher assessment by Policy capturing. *Assessment and Evaluation in Higher Education, 28*(4), 395-410.

Van der Schaaf, M., Veenhoven, J., & Stokking, K. (2003). Ontwikkelen van onderzoeksvaardigheden [Developing research skills]. Paper Conference Vereniging voor Onderwijsresearch, November 26, 2003 Nijmegen, The Netherlands.

Van der Schaaf, M.F. (2002). Naar een nieuwe didactiek [Heading towards new didactics]. *Van twaalf tot achttien, 2,* 14-15.

Van der Schaaf, M.F. (2002). Oppasser in het studiehuis. [Baby-sitter in the 'studyhouse']. *Van twaalf tot achttien, 1*, 20-21.

Van der Schaaf, M.F. (2002). Zelfstandig leren, wat doe je eraan? [Self-regulated learning, what to do about it?] In J. Ahlers, R. Diephuis, T. Hoogbergen & P. Leenheer (Eds.), 6. *Vormgeving van het onderwijs. Handboek studiehuis tweede fase*, (pp. 1-22). Alphen a/d Rijn: Kluwer.

Van der Schaaf, M.F. (2001). De juiste leerling op de juiste plaats [The appropriate student in the right place]. *Didaktief & School, 7*(31), 26-27.

Van der Schaaf, M.F. (2001). *Maatregelen bij zelfstandig leren* [Measures for self-regulated learning]. Utrecht, The Netherlands: ICO-ISOR, Utrecht University. (Isbn 90-6709-050-6)

Van der Schaaf, M.F. (2001). Valse start in het studiehuis [A false start in the 'studyhouse']. *Van twaalf tot achttien, 10*, 14-15.

Van der Schaaf, M.F. (2000). *Evaluatie VSB Mediatheken project* [Evaluation of the VSB Multimedia centre project]. Utrecht, The Netherlands: ICO-ISOR, Utrecht University. (Isbn 90-6709-045-x)

Van der Schaaf, M.F. (2000). Innoveren naar meer zelfstandig leren [Innovating towards more self-regulated learning]. In K. Stokking, G. Erkens, B. Versloot & L. van Wessum (Eds.), *Van onderwijs naar leren. Tussen het aanbieden van kennis en het faciliteren van leerprocessen.* (pp. 233-246). Leuven, Apeldoorn: Garant.

Van der Schaaf, M.F. (2000). Leren van jezelf als begeleider [Teaching yourself as a coach]. *Didaktief & School, 6*(30), 22-23.

* Stokking, K.M., & Van der Schaaf, M.F. (2000). *Ontwikkeling en beoordeling van onderzoeksvaardigheden* [Development and assessment of research skills]. Verslag van een onderzoek in de Tweede Fase VO. Utrecht, The Netherlands: ICO-ISOR, Utrecht University. (Isbn 90-6709-047-6)

Van der Schaaf, M.F., Goris, M., & Stokking, K.M. (2000). *Praktische opdrachten bij Grieks en Latijn* [Practical assignments in Classical languages (Greek and Latin)]. Utrecht, The Netherlands: ICO-ISOR, Utrecht University. (Isbn 90-6709-040-9)

Van der Schaaf, M.F. (2000). *Samenvatting evaluatie VSB-mediatheken project* [Summary evaluation VSB Multimedia centre project]. Utrecht, The Netherlands: ICO-ISOR, Utrecht University.

Van der Schaaf, M.F. (2000). *Vaardig in het begeleiden van zelfstandig leren? Monitoring-instrument* [Competent in coaching self-regulated learning? A monitoringinstrument]. Utrecht, The Netherlands: ICO-ISOR, Utrecht University. (Isbn 90-6709-041-7)

Van der Schaaf, M.F. & Balvers, H.C.M. (2000). *Vaardig in het begeleiden van zelfstandig leren? Ontwikkeling van een reflectie-instrument.* [Competent in coaching self-regulated learning? Development of a reflection-instrument] Paper Onderwijs Research Dagen, May 24-26, 2000 Leiden, The Netherlands.

Stokking, K.M., & Van der Schaaf, M.F. (1999). Beoordeling van onderzoeksvaardigheden van leerlingen [Assessing students' research skills] Richtlijnen, alternatieven en achtergronden. Studiehuisreeks, 27. Tilburg: Mesoconsult. (Isbn 90-1384-264-1)

Van der Schaaf, M.F., & Stokking, K.M. (1999). *Didactiek en organisatie bij zelfstandig leren [Didactics and organization of self-regulated learning]. Een inventarisatie van veranderingen en ervaringen rond de invoering van het Studiehuis.* Utrecht, The Netherlands: ICO-ISOR, Utrecht University.

Stokking, K.M., & Van der Schaaf, M.F. (1999). Onderzoeksvaardigheden: waaraan moet een goede toets voldoen? [Research skills: requirements for an adequate assessment] *Didaktief & School,* 1/2, 36-37.

Van der Schaaf, M.F., & Stokking, K.M. (1999). *Vaardig in het begeleiden van zelfstandig leren? Constructie van een monitoringinstrument ten behoeve van het begeleiden van zelfstandig werken en leren van leerlingen door docenten* [Competent in coaching self-regulated learning? Construction of a monitoring instrument]. Utrecht, The Netherlands: ICO-ISOR, Utrecht University.

Harskamp, E., Van der Schaaf, M.F., & Stokking, K.M. (1999). *Zelfstandig leren in Studiehuis en Beroepseducatie* [Self-regulated learning in pre-university and vocational education]. Groningen/Utrecht, The Netherlands: GION/ICO-ISOR. (Isbn 90-6690-460-7)

Van der Schaaf, M.F., & Stokking, K.M. (1999). *Zelfstandig werken en leren in het Studiehuis. Resultaten vragenlijst Oosterlicht College* [Self-regulated learning in the 'studyhouse'. Results of a survey at Oosterlicht College]. Utrecht, The Netherlands: ICO-ISOR, Utrecht University.

Stokking, K.M., & Van der Schaaf, M.F. (1998). *Zelfstandig werken en leren in het studiehuis* [Self-regulated learning in the 'studyhouse']. Utrecht, The Netherlands: ICO-ISOR, Utrecht University.

Van der Schaaf, M.F. (1996). Netwerken, een modern fenomeen [Networking, a modern phenomenon]. In J. Ahlers (Ed.), *Handboek basisvorming/Tweede Fase*, (pp. 1-22). Alphen a/d Rijn: Samson H.D. Tjeenk Willink.

Submitted for publication

* Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (2005). Developing content standards for teaching students research skills using a Delphi method. Manuscript submitted for publication.

* Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (2005). Designing teacher portfolios using a chain model of construct validation. Manuscript submitted for publication.

* Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (2005). Teacher beliefs and behaviour in portfolio assessment. Manuscript submitted for publication.

Stokking, K.M., & Van der Schaaf (2005). Views of research and teaching students how to do it. Manuscript in preparation.

Awarded research proposals by the Dutch Organization for Scientific Research

Stokking, K.M., & Van der Schaaf, M.F. (2004). *Development and assessment of cross-curricular skills using a student portfolio.* Utrecht, The Netherlands: ICO-ISOR, Utrecht University. (Grant NWO-PROO 411-03-119)

Stokking, K.M., & Van der Schaaf, M.F. (2000). *Assessing teacher competences related to teaching students research skills.* Utrecht, The Netherlands: ICO-ISOR, Utrecht University. (Grant NWO-PROO 411-21-204)

Dankwoord

Ik heb het voorrecht gehad te mogen werken aan een onderzoek dat ik vanaf de aanvraag tot en met de afronding naar eigen inzichten kon invullen. Dat was echter onmogelijk geweest zonder de inzet van vele mensen. Een aantal van hen wil ik nadrukkelijk bedanken.
De basis voor dit proefschrift is een NWO-PROO onderzoeksaanvraag. Ik dank de beoordelaars van NWO-PROO voor hun toekenning.
Bij de uitvoering van het onderzoek is een cruciale rol vervuld door de ruim veertig 'stakeholders' die zijn geïnterviewd of hebben deelgenomen aan het delphi-onderzoek, en de ruim twintig docenten en hun leerlingen die speciaal voor het onderzoek een portfolio hebben samengesteld en vragenlijsten hebben ingevuld. Sinus Barendregt, Marcel van de Borgt, Harrie Mennen, Ron Hendriks, Ruud Visser en Hans Wessels zijn in de rollen van panelleden en beoordelaars gedurende een paar jaar de steunpilaren van het onderzoek geweest. Dank voor de prettige samenwerking en het vele en goede werk dat jullie hebben geleverd.
Speciale dank gaat uit naar mijn promotoren Karel Stokking en Nico Verloop. Karel, dank voor je inspiratie, vertrouwen, coaching en collegialiteit bij de tot standkoming van dit proefschrift, de andere afgesloten en lopende onderzoeksprojecten en het universitaire onderwijs waarin ik met veel plezier als docent participeer. Ik hoop nog veel van je te leren en de passie voor het veelzijdige vak te blijven delen. Nico, dank voor al je expertise op de gebieden van 'teachers' en 'evaluation', je goede adviezen en kritische commentaar.
Veel dank gaat uit naar de onderzoeksassistenten Annick de la Rive Box en mijn moeder Kitty van der Schaaf die zich zeer nauwkeurig en volhardend over een groot deel van de dataverwerking en –analyse hebben ontfermd.
Ronny Wierstra dank ik voor het inhoudelijk becommentariëren van een aantal hoofdstukken uit het proefschrift. Piet van der Lei en Kitty van der Schaaf, bedankt voor het verbeteren van de Nederlandstalige samenvatting. Evelien van de Velde, bedankt voor het corrigeren van het Engels. Roy Sanders dank ik voor het op professionele wijze verzorgen van de lay-out van het proefschrift. Brian Golsteijn, dank voor al je creativiteit, inzet en vrijgemaakte tijd voor de omslag en de uitnodigingen. Anne van Dam, bedankt voor je logistieke inzicht en inzet.
Dank gaat uit naar mijn collega's van de Afdeling Onderwijskunde in Utrecht voor het creëren van een prettige en stimulerende werkomgeving. Dat geldt in het bijzonder voor mijn kamergenote Frieda Leenders. Dank voor je belangstelling, je adviezen, en het luisteren naar mijn verhalen.

De collega's van de ICLON-onderzoeksgroep, de werkgroep competenties en de onderzoeksgroep OZON dank ik voor de stimulerende discussies en kritische uitwisseling over aan het onderzoek gerelateerde onderwerpen.
Kitty, Piet, Tilly en Wim, dank voor al jullie goede zorgen en belangstelling voor mijn onderzoek.
Irene, lief zusje, hoe kan ik trotser zijn om op dezelfde dag als jij in Utrecht te mogen promoveren en ook nog eens elkaars paranimf te zijn!
Mijn allerliefste Arjan, dank je voor je onvoorwaardelijke liefde en support en al het goede van afgelopen decennium en voor de toekomst. Je betekent alles voor mij en dat is meer dan ik in woorden kan uitdrukken.
De laatste woorden van dank gaan uit naar al die mensen die een bijdrage hebben geleverd, maar hier niet met naam zijn genoemd.
Dit proefschrift draag ik met liefde op aan Hein en Feikje van der Schaaf.

Marieke
Harmelen, augustus 2005

Sister, you've been on my mind
Sister, we are two of a kind
So, sister, I'm keeping my eyes on you
(the color purple)