Validity in teacher assessment. An exploration of the judgement processes of assessors

Nijveldt, M.J.

Validity in Teacher Assessment

An exploration of the judgement processes of assessors

ICLON
Leiden University Graduate School of Teaching

ico
This research was carried out in the context of the Dutch Interuniversity Center for Educational Research.

NWO
Netherlands Organisation for Scientific Research

This research was funded by the Netherlands Organization for Scientific Research (NWO) (Project no. 411-21-205).

Title: Validity in Teacher Assessment: an exploration of the judgement processes of assessors
Titel: Validiteit in de beoordeling van docentcompetenties: een verkenning van beoordelingsprocessen van assessoren

Print: Gildeprint, Enschede
Cover design: Francien Frommé (www.grootzus.nl)
ISBN 978-90-9022467-1

Validity in Teacher Assessment: An exploration of the judgement processes of assessors

Proefschrift

ter verkrijging van

de graad van Doctor aan de Universiteit Leiden,

op gezag van de Rector Magnificus prof. mr. P.F. Van der Heijden

volgens besluit van het College voor Promoties

te verdedigen op woensdag 16 januari 2008

klokke 16.15 uur

door

Mirjam Jeannette Nijveldt

geboren te Gouda

in 1979

Promotiecommissie

Promotores

Prof. dr. N. Verloop

Prof. dr. Th. Wubbels

Copromotores

Prof. dr. D. Beijaard

Prof. dr. J.M.G. Brekelmans

Overige leden

Prof. dr. K. Smith (referent)

Prof. dr. J.H. van Driel

Dr. J.W.F. van Tartwijk

Contents

1

Introduction

The research presented in this dissertation concerns validity in teacher assessment. As the assessment task requires considerable judgement on the part of assessors, the overall validity of the assessment rests heavily on the nature and quality of assessors' judgement processes. Although several validity criteria tailored for the assessment of complex performances like teaching have been suggested, it remains unclear just how these criteria are to be met in practice. Central to this dissertation are three studies of the validity of the assessment of student teachers, which were focused particularly on the judgement processes of assessors. All three studies addressed the judgement processes of assessors who evaluated the interpersonal competence of a student teacher, using a specific assessment procedure. Interpersonal competence refers to the ability of teachers to develop satisfying interpersonal relationships with students, and thereby a satisfying classroom atmosphere. This first chapter deals with the background to the three studies, the general problem definition, and the context of the studies. The chapter concludes with an overview of the chapters that follow.

1.1 Background to the study

Assessment of teacher competence

Assessment of teaching is currently becoming more and more common practice. The growing use of teacher assessment clearly coincided with the acknowledgement of the central role of the teacher in the educational process: teacher competence is nowadays widely seen as the most critical factor in influencing student learning (Darling-Hammond & McLaughlin, 1999; Rivkin, Hanushek, & Kain, 2005). Professional teachers are expected to continuously improve

their competences throughout their careers and to adjust these competences to new teaching demands. In many western countries, teaching competences which are required for effective performance in specific teaching contexts have been specified, and procedures to assess the extent to which teachers master these competences are being developed and used. Both the specification and the assessment of teaching competences are believed to support the professional development of teachers as they encourage reflection and meaningful dialogue or debate among practitioners on the essence of good teaching (Darling-Hammond & Snyder, 2000; Delandshere & Arens, 2003; Dwyer & Stufflebeam, 1996). In the Netherlands, recent legislation by the 'Professions in Education Act' (Wet op de beroepen in het onderwijs, Stb. 2004, 344; Stb. 2005, 672) led to the formulation of national professional standards for teachers, assisting staff members, and school managers. These standards are described by teachers and other stakeholders in the field of education, coordinated by the Association for the Professional Qualities of Teachers and other Educational Personnel (SBL). The basic competences for teachers in both primary and secondary education include interpersonal competence; pedagogical competence; subject-matter knowledge and methodological competence; organizational competence; collaboration with colleagues; collaboration with the working environ-ment; and reflection and development.

In the Netherlands, teacher assessment is used, for instance, to admit candidates to alternative teacher certification programmes. In order to enter such alternative programmes for teacher education, prospective alternative route candidates ('zij-instromers') — who have a non-teaching background, but have acquired, in jobs other than teaching, competences which are considered relevant to teaching — need to demonstrate that they meet the initial standards of competence. Also, teacher assessment is increasingly used in the context of regular teacher education, both to stimulate reflection and to guide further professional development of student teachers (i.e., assessment for formative purposes), and to inform decisions about the certification of these prospective teachers (i.e., assessment for summative purposes). Although teacher education institutes are required to use the standards formulated by SBL as a guideline in their educational programmes, they have great liberty in deciding how to align their programmes and assessment practices with the standards.

Requirements for teacher assessment

In order to obtain a viable, comprehensive view of teacher competence and to successfully promote teacher professional development, teacher assessment for both formative and sum-

mative purposes should clearly reflect that teaching is complex, context-specific, and person-bound. Teaching is *complex* because much happens in a classroom simultaneously and tensions may occur among competing goals of the teacher (e.g., giving a specific student personal attention versus managing the class as a whole). Because the classroom setting makes constant demands on the teacher there is, moreover, little time to reflect (Doyle, 1986; Uhlenbeck, Verloop, & Beijaard, 2002). Teaching is *context-specific* as, instead of applying set routines, teachers need to adapt their teaching to the needs of the particular students and to particular situations (Cochran-Smith, 2003). These situations are defined by the nature of the subject matter, the goals of instruction, the individual characteristics of the students, and the settings in which teaching and learning take place (Darling-Hammond & Snyder, 2000). Adapting teaching to such particularities requires a wide range of teaching strategies and styles, and the ability to evaluate teaching situations and select those strategies and approaches that fit these particular situations best. Lastly, teaching is also *person-bound*. Teachers' decision-making inevitably depends on personal goals in teaching, personal styles, personal theories about teaching and learning, and personal interpretations of particular teaching situations (Uhlenbeck et al., 2002). Against the background of these characteristics of teaching, a number of requirements have been suggested for valid teacher assessment which also constitutes valuable learning experiences for teachers (Darling-Hammond & Snyder, 2000; Dwyer, 1998; Uhlenbeck et al., 2002). Assessment procedures should, for instance, have a high level of authenticity: actual teaching activities should be assessed in the context of a teacher's actual work and classroom setting. In connection with this, assessments should not be limited to teacher behaviour, but should also include opportunities for teachers to elaborate on underlying cognitions, i.e., knowledge and beliefs (Calderhead, 1996; Clark & Peterson, 1986). The candidate could, for example, be invited to elaborate on what he or she aimed to achieve in considered teaching situations, the way he or she tried to achieve these goals, and how the specific context influenced his or her decisions. The evidence necessary to acquire a comprehensive view of teaching competence should be addressed using a variety of methods. Using multiple sources of evidence is generally assumed to result in a more valid assessment, as no method in itself can lead to an accurate overall judgement of teacher competence. In addition to standardized material, non-standardized, qualitative material should be collected, like videotaped lessons, lesson plans, or written reflections on performance. This material should be evaluated by individuals with relevant expertise, against criteria which are meaningful for performance in the field. It is important that such assessment criteria have the right level of specificity: if criteria

are too general or vague, they are difficult to apply fairly, and if they are too specific, they do not leave room for different teaching styles or different teaching contexts (Dwyer, 1995). Finally, assessors should analyze the evidence obtained in a holistic manner; pieces of performance should be analyzed, interpreted, and evaluated in the context of the whole performance. Although some aspects of the performance may be especially praised, or noted as needing improvement, the focus in holistic scoring is on the quality of the overall performance, paying attention to the contributions to the whole made by different aspects. The underlying assumption is that the whole is greater than the sum of its parts (Mabry, 1999).

Validity and reliability

Validity and reliability are traditionally seen as the primary quality criteria for any form of assessment. Although these criteria have been operationalized around the use of standardized forms of assessment, they are widely considered just as important for the authentic assessment of complex performance (Mabry, 1999; Messick, 1989). Reliability concerns the accuracy of the assessment, that is, the extent to which an assessment procedure would produce a similar score on different occasions or if given by different assessors. In practice, reliability is generally defined as agreement between assessors on the same task (i.e., inter-rater reliability). The traditional definition of validity is the extent to which a test measures what it was designed to measure. In line with this view, the early literature about validity emphasized three types of validity: construct, content, and criterion validity. Construct validity refers to the extent to which the test is an adequate measure of the construct (i.e., the underlying skill or competence being assessed). Important in this respect is a clear definition of the construct which addresses all of its relevant aspects (Gipps, 1994). Content validity concerns the coverage of appropriate and necessary content, and refers to the extent to which the test covers the skills or competences necessary for a good performance. Criterion validity refers to the extent to which test scores predict future performance and correlate with the results of another test of the same skill or competence. Following Cronbach (1988) and Messick (1989), validity can be seen as a broad concept that unifies these three types of validity, with construct as the central theme. Validity is no longer considered a property of the test as such, but is related to the soundness of the user's *interpretation* of the test scores: "Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment…[V]alidity is an inductive summary of both the existing evidence for and the potential consequences of score

interpretations and use" (Messick, 1989, p.13). Messick's definition thus clearly stresses 'consequential validity' of assessments in terms of their potential impact on learning in addition to the soundness of inferences. This 'unified conception' of validity is widely accepted at present.

Although validity has been widely acknowledged as the most important criterion for high-quality assessment (Gipps, 1994), in research and in the practice of authentic assessment, validity has often received less attention than reliability (Pitts, Coles, & Thomas, 1999). A *first reason* for this is that authentic forms of teacher assessment — based on sources of data which are qualitative, context-specific, and person-bound — are generally perceived as highly valid. Validity is, therefore, often taken for granted, whereas reliability is often found to be unsatisfactory, as unambiguous objective assessment simply does not exist. In this respect, it is important to note that although inter-rater reliability is a feasible and, therefore, common quality criterion, there are still a number of problems to be considered in relation to the interpretation of such consistency estimates. A lack of consistency between assessors' judgements, for instance, is not necessarily an indicator of an unsound interpretation, but can be caused by differences in the expertise of the assessors, which might be complementary and, therefore, meaningful (Delandshere & Petrosky, 1998; Dierick, Dochy, & Van de Watering, 2001). A lack of consistency could also be the result of a number of unforeseeable difficult issues: low reliability of an initial judgement of competence of a specific teacher may, for instance, indicate that the performance of this specific teacher needs additional attention as it raises a set of issues that has not been considered yet in the assessment framework (Moss, Schutz, & Collins, 1998). It is also possible that assessors do agree on the level of competence of the candidate but ground this decision on different arguments (e.g., Moss & Arens, 2001). The meaning of such a consensus should be questioned. Thus, inter-rater reliability does not necessarily indicate whether assessors are making sound judgements, and some authors advocate that reliability needs to be redefined and that other criteria to ensure that the judgement process proceeds fairly and responsibly should be used. For instance, the quality of assessors' argumentation might serve as such a criterion (Mabry, 1999; Moss, 1994; Uhlenbeck et al., 2002). Such a criterion might also provide a more adequate basis for improving the quality of assessment than measures of inter-rater reliability do.

A *second reason* for the disproportionate attention to reliability in the evaluation of assessment appears to be the discrepancy between the algorithmic nature of procedures for estimating reliability and the more open nature of procedures for validity inquiry. In contrast to reliability estimation, validity inquiry often relies on human interpretation and is, therefore,

difficult to carry out and report. As a result, validity evidence is often simply not gathered in practice (Crooks, Kane, & Cohen, 1996).

Ensuring validity in assessment

Messick's unified conception of validity outlined above is potentially suitable to guide validity inquiry of authentic teacher assessments, but its breadth and complexity make it hard to work with in practice. This has led some authors to suggest *specialized* validity criteria tailored for authentic assessments (Frederikson & Collins, 1989; Haertel, 1991; Linn, Baker, & Dunbar, 1991). Some of these criteria highlight specific aspects of construct validity, for instance, cognitive complexity of the assessment, content quality, and content coverage. Other criteria, such as meaningfulness and transparency, highlight important additional aspects of validity, as these criteria relate to the extent to which the assessment provides a worthwhile experience for teachers, serving to motivate and direct professional development. Meaningfulness refers to the extent to which the assessment provides a valuable learning experience for the candidates (Linn et al., 1991), and transparency concerns the extent to which the candidates know what they are expected to do and which competences will be assessed, using which standards of performance.

Other authors have suggested that the traditional quality criteria of validity and reliability are simply not appropriate to evaluate the quality of new forms of assessment and have, therefore, proposed *alternative* quality criteria (Driessen, Van der Vleuten, Schuwirth, Van Tartwijk, & Vermunt, 2005; Gipps, 1994; Tigelaar et al., 2005). Examples of such alternative criteria are credibility, dependability, confirmability, and transferability, as used in qualitative research. These criteria refer to the extent to which: the interpretations of assessors match those of stakeholders in the field (*credibility)*; the interpretation process is established, traceable, and documented *(dependability)*; the interpretations and conclusions are supported by the original data *(confirmability)*; and, finally, others are able to determine whether conclusions are applicable to other contexts or situations *(transferability)*.

Although the above-mentioned new conceptions of quality of authentic assessment have proven to be important in identifying validity issues, it is still unclear how assessors should comply with and evaluate specific criteria in the practice of authentic teacher assessment. The literature provides particularly little guidance on important issues related to the judgement processes of assessors, like taking context into account (Dwyer, 1995) and combining data from different sources (Moss, Girard, & Haniford, 2006).

In this dissertation, we hold with the view that the criterion of validity is still important and suitable to guide the development and validation of authentic assessment (Mabry, 1999; Messick, 1989). Nevertheless, we argue that research into the exact nature of judgement processes is needed to gain greater insight into the specific ways in which validity can be ensured in practice. Whereas in traditional modes of assessment, scores are a function of the assessment items or tasks, and validity inquiry is thus aimed at the assessment procedure itself, in the authentic assessment of teacher competence, scores are also a function of the judgement processes of assessors. Therefore, validity inquiry should explicitly address the quality of such processes. As Heller, Sheingold, and Myford (1998) note, "score validation must include theories of the processes, strategies and understandings underlying sound and appropriate raters' judgements, and empirical evidence of the extent to which raters engage in the theoretical processes"(p.7).

In this connection, Heller et al. explicitly applied the two general validity criteria of 'no construct-irrelevant variance' and 'no construct under-representation' (Messick, 1989,1994) to the judgement processes of assessors. Traditionally, in cases of *construct-irrelevant variance*, the assessment is too broad and thus contains variance associated with other — irrelevant— constructs. In cases of *construct under-representation*, the assessment is too narrow and thus fails to capture critical aspects of the target construct. Heller et al. illustrated that, in the context of assessment of competence, *construct-irrelevant variance* may arise when assessors apply irrelevant or idiosyncratic criteria and thus consider evidence which is simply irrelevant to the construct in question. *Construct under-representation* is characterized by the omission of particular assessment criteria and/or evidence, or the idiosyncratic weighing of certain criteria and/or evidence, which may result in insufficient attention to particular aspects of performance.

In line with Heller et al., Moss et al. (1998) introduced two interrelated principles underlying a valid assessment process, based on explorative empirical research on the nature of assessment processes. The first principle is that assessors should search for coherence in evidence from different sources, and explicitly check to see that all relevant evidence has been taken into account. The second principle is that assessors should be encouraged to actively seek and consider counterevidence, thus evidence which challenges their developing interpretations, and explicitly consider alternative interpretations. Compliance with these two principles is assumed to minimize the aforementioned threats to validity, namely, the introduction of construct-irrelevant variance or construct under-representation, or both. This is easier said

than done for assessors. To illustrate the specific problems the assessment process poses, we elaborate in the next section on research on the nature of assessors' judgement processes.

The nature of the judgement processes of assessors

Although the selection of the right assessors and the extensive training of these assessors are important preconditions for a valid assessment process, this alone still does not guarantee high quality assessment processes. In relation to the assessment of competence based on several types of data, the assessment process is often thought of in terms of two essential assessor tasks: (a) consideration of evidence from separate sources and (b) combination of evidence from different sources to attain an overall judgement (Delandshere & Petrosky, 1994, 1998; Delandshere & Arens, 2003; Mehrens, 1990; Moss et al., 1998; Wolf, 1995; Uhlenbeck et al., 2002). These assessor tasks require considerable judgement on the assessor's part and may, therefore, be difficult to carry out in a sound way. Below, we bring up a number of issues pertaining to these assessor tasks.

The *consideration of evidence from separate sources* typically involves applying content criteria (i.e., criteria which conceptualise the relevant competences) to a specific teaching context. For example, the variability of contexts in which teacher competence is assessed impels assessors to determine how to take account of context when judging whether an observed piece of evidence fits in a defined content criterion. Specification of the criteria cannot resolve this issue: as opposed to assessment of a simple set of atomized content criteria, the assessment of complex behaviour like teaching requires a holistic concept of competence (or 'assessment framework'), which explicitly underlies the content criteria and is to be internalized by assessors (Wolf, 1995). Two important issues pertain to the consideration of evidence. Firstly, assessors unavoidably have personal beliefs, values, or perspectives about teaching quality that are not covered in the assessment framework. Some of these perspectives may conflict with the assessment framework, and some may reflect practical knowledge that enhances the quality of the assessment (Moss et al., 1998). Secondly, in the process of consideration of evidence from separate sources, assessors seem to have great difficulty not limiting themselves to recognizing superficial features of competence. That is, assessors seem to show a tendency to attend to the most visible or superficial characteristics of the performance without further interpretation. Such a tendency concerns the quality of the evidence considered in the assessment, as assessors might, for instance, predominantly refer to abstract key words from the assessment framework instead of referring to the salient particulars of the performance of a

specific candidate (Delandshere & Arens, 2003; Delandshere & Petrosky, 1998; Moss et al., 1998).

In the process of *combination of evidence from different sources to attain an overall judgement*, it is to be left to the assessors to determine how the various data sources are combined, as the complexity of teaching makes a strict marking scheme inappropriate (Wolf, 1995). This task constitutes a number of problems for assessors: in the context of teacher portfolio assessment, Moss et al. (1998) have shown that assessors have great difficulty making "a comprehensive weighing of available evidence" instead of a "selective recollection" (p.155). Assessors cannot avoid interpreting parts of the teacher portfolio in terms of a more cohesive understanding of the whole; that is, assessors cannot avoid forming a pattern out of the data from a teacher portfolio in order to evaluate it. And once a particular pattern has been identified, subsequent information may also be interpreted in terms of this pattern. "Even if the subsequent data disconfirms the initial pattern, it is read to some extent in response to the pattern it disconfirms" (Schutz & Moss, pp. 8-9). This strong tendency to seek confirmation of one's initial interpretations can be seen to constitute a bias at times, which can most certainly affect the validity of the judgement process. In connection with this, in reaching an overall judgement, assessors compensate, make allowances, interpret, and explain away even when instructed not to. The more experienced the assessors, the more they operate in a familiar field, and the more they have internalised a concept of competence, which may or may not be in accordance with the assessment framework; and the more they are inclined to do so. Assessors are often unaware of the degree to which they operate in this way (Wolf, 1995).

Up till now, the nature of assessment processes has been addressed particularly by researchers who advocate an 'interpretivist' approach to combining sources of evidence in developing interpretations. Interpretivist approaches stress the importance of holistic integration of evidence, and emphasize the notion that it is simply impossible to strive for a truly objective judgement, as assessment always involves human interpretation. To ensure that judgements are based on sound interpretations, however, interpretivist assessment generally involves the following: (a) discussion among a community of assessors of values and standards; (b) discussion of individual assessments that encourages challenges and revisions to initial interpretations; and (c) transparency of the assessment process: assessors should represent their analysis so that others may review it (e.g., Johnston, 2004; Moss, 1994; Moss et al., 2006; Tigelaar, Dolmans, Wolfhagen, & Van der Vleuten, 2005). In interpretivist assessment practices, the aforementioned processes of consideration of evidence from separate sources and

combination of evidence to attain an overall judgement would ideally be performed in collaboration between different assessors. Moss et al. (1998) suggest that assessors first work through the available data individually (i.e., individually note the relevant evidence and prepare an initial, individual judgement) and then work together in pairs to formulate a definitive collaborative judgement (i.e., an evaluative summary). During this process, the assessors should be encouraged to note not only supportive evidence and examples, but also conflicting evidence and counterexamples; to discuss alternative interpretations; and to revise their interpretations until all relevant evidence has been taken into account. These activities — which clearly bear on the validity of assessment — can particularly be stimulated by discussion between assessors.

1.2 Problem definition and research questions

As mentioned above, validity has been widely acknowledged as the primary criterion for any form of assessment, including authentic assessment. But although the current conception of validity is potentially suitable to guide validity inquiry of authentic teacher assessments, its breadth and complexity make it hard to work with in practice. Even though specialized and additional validity criteria have been proposed, which are now broadly accepted, it remains unclear how assessors can comply with these criteria in teacher assessment practices. Current validity theory provides little guidance on important issues in the judgement processes of assessors, like taking contextual differences into account (Dwyer, 1995) and combining data from sometimes highly differing sources (Moss et al., 2006). Given that the overall validity of the assessment depends heavily on the judgement processes of the assessors, research into the exact nature of such judgement processes is needed to gain greater insight into specific validity issues in teacher assessment and the specific ways in which assessors can comply with validity criteria (i.e., the principles underlying a valid assessment process).

This study was aimed at describing the nature of the judgement processes of assessors, and in such a manner as to illuminate successful strategies and threats underlying a valid assessment process. The general question that prompted the current research was the following: How can the validity of teacher assessment be ensured in practice? More specific questions were addressed which clearly focused on the nature of the judgement processes of assessors. As argued above, construct representation should be a central theme in validity inquiry of not only traditional assessment, but also of authentic assessment, where the con-

structs or competences to be assessed are complex and are not always well conceptualized. For this reason, in answering these research questions, the concept of construct representation was given particular attention (see section 1.4).

1. To what extent do construct-irrelevant variance and construct under-representation occur in the judgement processes of assessors in the consideration of evidence from separate sources and combination of evidence to attain an overall judgement?
2. What is the nature of the assessment process when teachers are assessed collaboratively?
3. What strategies to ensure the validity of the assessment process are performed by assessors in connection with the essential judgement processes of (a) consideration of evidence and (b) combination of evidence to attain an overall judgement?

The problem definition addressed in the present research is theoretically relevant and also has practical relevance, because of the direct implications for the specific ways in which the soundness of interpretation and the validity of assessment can be evaluated and improved in practice. The results have implications for the development of assessment procedures and the training of assessors. This dissertation might thus be of interest to practitioners who use assessment in educational settings or to developers of assessment practices.

1.3 Context of the study: The assessment of interpersonal competence

This dissertation reports on a series of studies which were focused on assessors' judgement processes. The assessors in the present research used a specific procedure to assess the *interpersonal competence* of a student teacher in the context of teacher education. In order to be meaningful for the candidates — prospective teachers in the upper level of secondary education in this particular case — assessment should particularly address typical concerns of these prospective teachers (Dwyer, 1995). Because interpersonal teacher competence, or the quality of interpersonal relationships with students, is a major concern for many beginning teachers (Evans & Tribble, 1986; Evertson & Weinstein, 2006; Veenman, 1984), and experienced teach-ers mention satisfying interpersonal relationships with students as a prerequisite for further professional development (Beijaard, Verloop, Wubbels, & Feiman-Nemser, 2000), this particular competence was selected to be assessed in the present research.

The literature base on interpersonal relationships between teachers and students is extensive (Wubbels, Brekelmans, Den Brok, & Van Tartwijk, 2006), and in contrast with many other teacher competences, the 'interpersonal competence' of teachers is a well-conceptualized and well-operationalized construct. Based on generic theories on interpersonal relationships and communication, and based on empirical research in education, a model is available to describe the quality of teacher-student relationships in terms of teacher behaviour. This model is based on Timothy Leary's research on the interpersonal diagnosis of personality (1957) and its application to teaching (Wubbels, Créton, & Hooymayers, 1985). The Leary model has been investigated extensively in clinical psychology and psychotherapeutic settings, and has proven effective in describing human interaction (Lonner, 1980). In the *Model for Interpersonal Teacher Behaviour (MITB),* the two dimensions are Influence (Dominance-Submission) and Proximity (Opposition-Cooperation). These dimensions can be represented in an orthogonal co-ordinate system. The two dimensions, represented as two axes, underlie eight types of teacher beha-viour: leadership, helping/friendly behaviour, understanding, giving students freedom and responsibility, uncertain, dissatisfied, admonishing, and strict behaviour (see Figure 1.1). The sectors are labelled DC, CD, etc. according to their position in the co-ordinate system (much like the directions in a compass). For example, the sectors 'leadership' and 'helpful/friendly' are both characterized by Dominance and Cooperation. In the DC sector, the Dominance aspect prevails over the Cooperation aspect. A teacher displaying DC beha-viour might be seen by students as enthusiastic, motivating, and the like. The adjacent CD sector, however, includes behaviours of a more cooperative and less dominant type; the teacher might be seen as helpful, friendly, and considerate. Figure 1.1 provides an overview of typical teacher behaviours that relate to each of the eight sectors of the MITB.

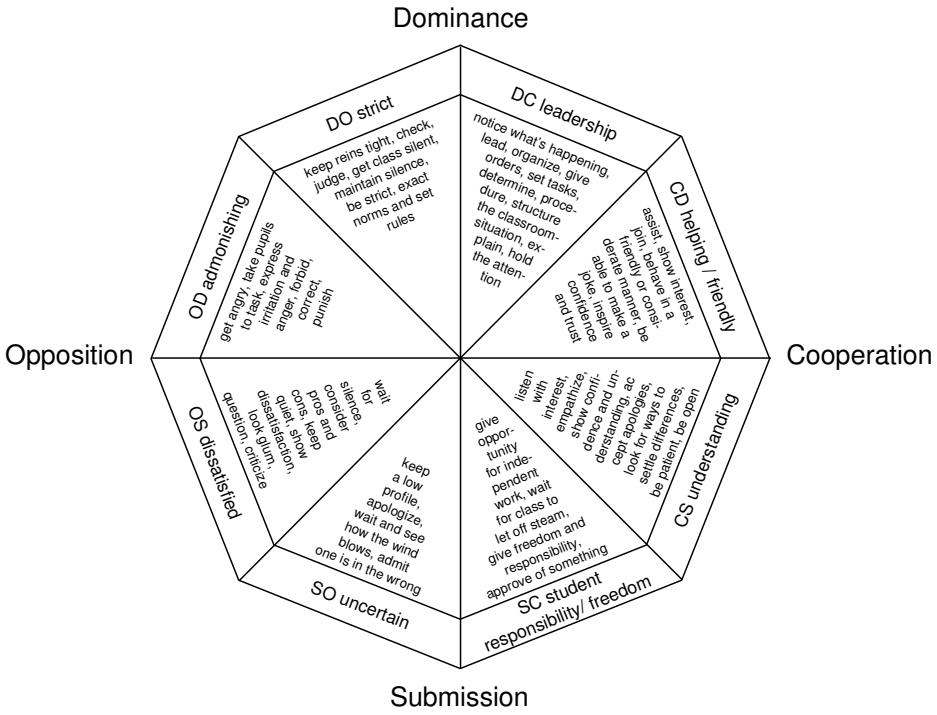


Figure 1.1 The Model for Interpersonal Teacher Behaviour

The interpersonal relationships between a teacher and his or her students can be investigated using the *Questionnaire on Teacher Interaction (QTI)*. The Dutch items were formulated based on large numbers of interviews with both teachers and students (Wubbels & Levy, 1993). The original Dutch version consists of 77 items to be rated on a five-point Likert scale ranging from 'Never/Not at all' to 'Always/Very'. The items are divided into eight scales corresponding with the eight sectors of the MITB. In Table 1.1, a typical item is provided for each of the eight sectors of the QTI. The QTI is designed to question both students and teacher on the interpersonal relationship in a specific classroom group. Students rate their teacher on the QTI, and teachers rate their perceptions about their own behaviour (i.e., self-perceptions). Teachers may also record their responses from the perspective of how they would like to be (i.e., ideal perceptions). The QTI results are typically presented in a report which includes the mean scores for each of the eight scales; the mean scores per item; and graphic representations

of the eight scale scores ('interpersonal profiles'). The profile can be represented by one central point (see Figure 1.2 for an example).

Table 1.1 Typical items for the eight scales of the QTI

	Sector	Typical item
DC	Leadership	S/He is a good leader
CD	Helping / Friendly	S/He is someone we can depend on
CS	Understanding	If we have something to say s/he will listen
SC	Student responsibility / freedom	S/He gives us a lot of free time in class
SO	Uncertain	S/He seems uncertain
OS	Dissatisfied	S/He is suspicious
OD	Admonishing	S/He gets angry quickly
DO	Strict	S/He is strict

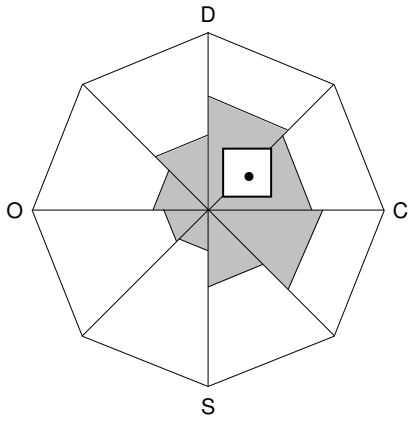


Figure 1.2 Example of a teacher interpersonal profile

Eight different types of interpersonal profile can be distinguished (Brekelmans, 1989; Brekelmans, Levy, & Rodriguez, 1993): Directive, Authoritative, Tolerant/Authoritative, Tolerant, Uncertain/Tolerant, Uncertain/Aggressive, Drudging, and Repressive. Most student and beginning teachers have been found to be of the Tolerant or Uncertain/Tolerant type (Wubbels et al., 2006). In Figure 1.3., the characteristic profile of each of the eight types is depicted on the basis of the two dimension scores of the profile using a central point in the coordinate system.

The assessment procedure developed in the context of the present research included use of the Questionnaire on Teacher Interaction (QTI) and two additional sources of data: a videotaped lesson selected by the student teacher and a self-evaluation assignment. In this self-evaluation assignment, the student teacher was asked to analyze both the results of the QTI and the videotaped lesson to provide (a) an overview of strengths and weaknesses with as much reference to the QTI results and videotaped lesson as possible and (b) a personal professional development plan. The literature base described above provided an adequate basis for the development of an assessment framework which distinguished the key components of interpersonal competence and helped the assessors to analyze the three sources of data in a holistic manner.

A typical characteristic of the present assessment procedure was *discussion with another assessor* after the assessors had formulated an initial, individual judgement based on the three data sources (cf. Moss et al., 1998). Following the procedure, the assessors firstly noted evidence from the three sources of data individually. They then integrated this evidence into an individual judgement in the form of an evaluative summary with supporting evidence for the key components of interpersonal competence. Developing such a summary required the assessors to integrate evidence from the different sources of data. The assessors then used these summaries when they entered into a discussion with another assessor, in which they attempted to reach consensus on the teacher's overall level of performance.

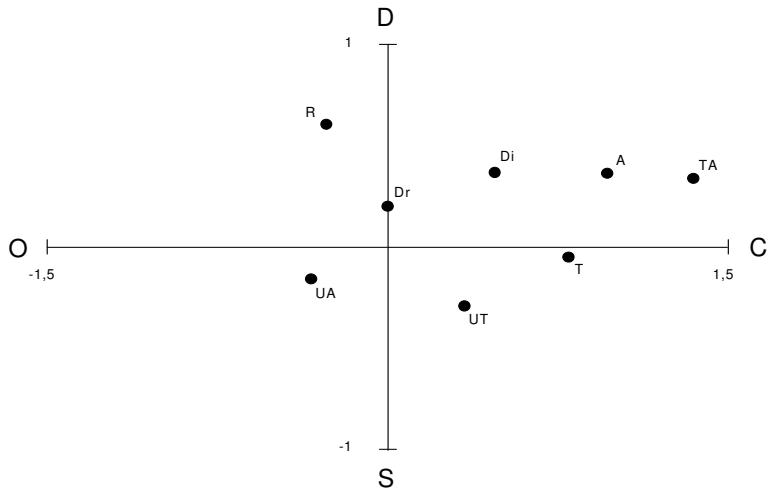


Figure 1.3 Central points of the eight types of pattern of interpersonal relationships. A=Authoritative, Di=Directive, Dr=Drudging ,T=Tolerant, R=Repressive, TA= Tolerant/Authoritative, UA=Uncertain/Aggressive, UT=Uncertain/Tolerant

1.4 Overview of the study

In this dissertation, three studies are presented in which the nature of the judgement processes of assessors was examined in different manners. Note that the results of all three studies are based on the same sample of assessors judging the same student teacher.

In Chapter 2, we describe a small-scale study in which the aforementioned procedure to assess the interpersonal competence of teachers was developed and validated. In this chapter, we address the first research question: to what extent do construct-irrelevant variance and construct under-representation occur in the judgement processes of consideration of sepa-rate sources and combination of evidence to attain an overall judgement? The following data were collected for four separate assessors judging the same student teacher: (a) the assessors' notes for the different sources of data, (b) the overall judgement of the student teacher's interpersonal competence provided by the assessors in the form of an evaluative summary and a numerical judgement along a 10-point scale, and (c) the results of a short interview in which the assessors were asked to explain their overall judgement. To analyze the occurrence of construct-irrelevant variance or construct under-representation, or both, the notes and evaluative summaries of the four individual assessors were compared to a so-called expert judgement.

The study described in Chapter 3 focuses on the aforementioned principles of assessment mentioned by Moss et al. (1998): (a) seeking coherence among the available sources of evidence and checking whether all relevant evidence has been taken into account, and b) challenging developing interpretations with counterevidence and explicitly considering alternative interpretations. Engagement in these two 'key assessment processes' is assumed to minimize the introduction of construct-irrelevant variance or construct under-representation, or both. As it is believed that discussion between assessors may enhance the chances of engagement in these processes, we examined the processes which assessors engage in when collaboratively (i.e., pair-wise) assessing the interpersonal competence of a student teacher. In this study, we addressed the second research question: What is the nature of the assessment process when teachers are assessed collaboratively? The aim of the study was to provide an overview of the specific activities undertaken by the pairs of assessors and to examine the extent to which these activities reflect the key assessment processes. The data consisted of transcripts of the decision-making conversations of 12 pairs of assessors.

In Chapter 4, we address the third research question: Which strategies to ensure the validity of the assessment process are performed by assessors in connection with the essential judgement processes of (a) consideration of evidence and (b) combination of evidence to attain an overall judgement? We examined the occurrence of construct-irrelevant variance or construct under-representation from the perspective of the assessors. It is of great importance for assessors to be aware of their own judgement processes and, specifically, of possible threats to validity in these processes, so that they can take conscious actions during the assessment process to enhance its validity. Experiences of the same assessors who participated in the former study (N=22) were investigated using semi-structured interviews. A qualitative analysis of individual assessors' perceptions with regard to the occurrence of construct-irrelevant variance and construct under-representation in relation to the consideration of evidence and combination of evidence to attain an overall judgement resulted in an overview of potential successful strategies and potential threats underlying a valid assessment process.

Finally, in Chapter 5, we present a general conclusion and discussion based on the findings described in the previous chapters, as well as suggestions for future research and implications for ensuring the validity of the assessment process and the preparation of assessors.

References

Beijaard, D., Verloop, N., Wubbels, Th., & Feiman-Nemser, S. (2000). The professional development of teachers. In R..J. Simons, J. Van der Linden, and T. Duffy (Eds.) *New Learning* (pp. 261-274). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Brekelmans, M. (1989). Interpersonal teacher behaviour in the classroom.[In Dutch]. Utrecht: W.C.C.

Brekelmans, M., Levy, J., & Rodriguez, R. (1993). A typology of teacher communication style. In T. Wubbels & J. Levy (Eds.), Do you know what you look like? (pp.46-55). London: The Falmer Press.

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.

Clark, C.M., & Peterson, P.L. (1986). Teachers' thought processes. In M.C. Wittrock (Ed.), *Handbook of research on teacher education* (pp. 255-296). New York: MacMillan.

Cochran-Smith, M. (2003). The unforgiving complexity of teaching: Avoiding simplicity in the age of accountability. *Journal of Teacher Education, 54*, 3-5.

Cronbach, L. (1988). Five perspectives on validity argument. In H. Weinder and H. Braun (Eds.) *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Crooks, T.J., Kane, M.T., & Cohen, A.S. (1996). Threats to the valid use of assessments. *Assessment in education: principles, policy & practice, 3*, 265-285.

Darling-Hammond, L., & McLaughlin, M.W. (1999). Investing in teaching as a learning profession: Policy problems and prospects. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice*. San Francisco: Jossey-Bass.

Delandshere, G., & Arens, S.A. (2003). Examining the quality of the evidence in preservice teacher portfolios. *Journal of Teacher Education*, *54*, 57-73.

Delandshere, G., & Petrosky, A. R. (1998). Assessment of complex performances: Limitations of key measurement assumptions. *Educational Researcher*, *27*, 14-24.

Dierick, S., Dochy, F., & Van de Watering, G. (2001). Assessment in het hoger onderwijs, implicaties van nieuwe toetsvormen voor de edumetrie. [Assessment in Higher Education, implications of new forms of assessment for edumetrics]. *Tijdschrift voor hoger onderwijs, 19*, 2-18.

Driessen, E.W., Van der Vleuten, C.P.M., Schuwirth, L.W.T., Van Tartwijk, J., & Vermunt, J. (2005). The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: A case study. *Medical Education, 39*, 214-220.

Doyle, W. (1986). Classroom organization and management. In M.C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 392-431). New York: MacMillan.

Dwyer, C.A. (1995). Criteria for performance-based teacher assessments: Validity, standards and issues. In A.J. Shinkfield, & D. Stufflebeam (Eds.), *Teacher evaluation: Guide to effective practice* (pp.62-80). Boston: Kluwer Academic Publishers.

Evans, E.D., & Tribble, M. (1986). Perceived teaching problems, self-efficacy, and commitment of teaching among preservice teachers. *Journal of Educational Research, 80*, 81-85.

Evertson, C.M., & Weinstein, C.S. (2006). Classroom management as a field of inquiry. In C.M. Evertson & C.S. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 3-16). Mahawn: Lawrence Erlbaum Associates.

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

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational measurement*. London, Washington D.C.: The Falmer Press.

Haertel, E. H. (1991). New forms of teacher assessment. In G. Grant (Ed.), *Review of Research in education* (Vol.17, pp. 3-29). Washington DC: American Educational Research Association.

Heller, J. I., Sheingold, K., & Myford, C.M. (1998). Reasoning about evidence in portfolios: Cognitive foundations for valid and reliable assessment. *Educational Assessment*, *5*, 5-40.

Johnston, B. (2004). Summative assessment of portfolios: an examination of different approaches to agreement over outcomes. *Studies in Higher Education*, *29*, 395-412.

Leary, T. (1957). An interpersonal diagnosis of personality. New York: Ronald Press Company.

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

Lonner, W. J. (1980). The search for psychological universals. In H.C. Triandis & W.W. Lambert (Eds.). Handbook of cross cultural psychology. (vol.1) pp. 143-204. Boston: Allyn and Bacon.

Mabry, L. (1999). *Portfolio's plus, a critical guide to alternative assessment*. Thousand Oaks: Corwin press.

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 CA: Sage Publications.

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

Messick, S. (1996). Validity of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1-18). Washington DC: Department of education, office of Educational research and Improvement.

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

Moss, P.A., Girard, B.J., & Haniford, L.C. (2006). Validity in Educational Assessment. *Review of Research in Education*, *30*, 109-162.

Moss, P. A., Schutz, A. M., & Collins, K. M. (1998). An integrative approach to portfolio evaluation for teacher licensure. *Journal of Personnel Evaluation in Education*, *12*, 139-161.

Pitts, J., Coles, C., & Thomas, P.(1999). Educational portfolios in the assessment of general practice trainers: reliability of assessors. *Medical Education, 33*, 515-520.

Rivkin, S.G., Hanushek, E.A., & Kain, J.F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*, 417-458.

Schutz, A., & Moss, P.A. (2004). Reasonable decisions in portfolio assessment: Evaluating complex evidence of teaching. *Education Policy Analysis Archives*, 12(33). Retrieved 7/19/2004 from http://epaa.asu.edu/v12n33/.

Tigelaar, D.E.H., Dolmans, D.H.J.M., Wolfhagen, I.H.A.P., & Van der Vleuten, C.P.M. (2005). Quality issues in judging portfolios: Implications for organizing teaching portfolio assessment procedures. *Studies in Higher Education, 30*, 595-610.

Uhlenbeck, A.M., Verloop, N., & Beijaard, D. (2002) Requirements for an assessment procedure for beginning teachers: Implications from recent theories on teaching and assessment. *Teachers College Record, 104*, 242-272.

Wolf, A. (1995). *Competence-based assessment*. Buckingham: Open University Press.

Wubbels, Th., Brekelmans, M., Den Brok, P., & Van Tartwijk, J. (2006). An interpersonal perspective on classroom management in secondary classrooms in the Netherlands. In. C. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: research, practice and contemporary issues* (pp. 1161-1191). New York: Lawrence Erlbaum Associates.

Wubbels, Th., Créton, H. A., & Hooymayers, H.P. (1985) Discipline problems of beginning teachers, interactional teacher behavior mapped out. Abstracted in Resources in Education, 20, 12, p. 153, ERIC document 260040.

Wubbels, T., & Levy, J. (1993). *Do you know what you look like?* London: Falmer Press.

2

Assessing the interpersonal competence of student teachers[1]

Abstract

In the present study, a procedure to assess the interpersonal competence of teachers was developed and validated. The procedure includes the use of the Questionnaire on Teacher Interaction (QTI), an observation instrument and a self-evaluation instrument. The concepts of construct-irrelevant variance and construct under-representation were used to explore the validity of the assessment procedure with regard to the noting of evidence and combining of evidence to attain an overall judgement by four separate assessors for the same student teacher. The validity of the assessment procedure was found to be satisfactory and the use of multiple assessment instruments to clearly have added value.

[1] This chapter has been published in adapted form as:
Nijveldt, M., Beijaard, D., Brekelmans, M., Verloop, N., & Wubbels, Th. (2005). Assessing the interpersonal competence of beginning teachers: The quality of the judgement process. *International Journal of Educational Research, 43,* 89-102.

2.1 Introduction

The quality of interpersonal relationships with students is a concern for many beginning teachers (e.g., Evans & Tribble, 1986; Evertson & Weinstein, 2006; Veenman, 1984). Particularly in secondary education, student and beginning teachers express a strong need for professional assistance in this domain. In the present study, a procedure to assess the interpersonal competence of student secondary school teachers was thus developed and validated. Such a procedure can provide a helpful impetus for the professional development of teachers but, in order to constitute a valuable learning experience, the procedure must clearly reflect the complexity of the teaching process (e.g., Darling-Hammond & Snyder, 2000; Dwyer, 1995; Uhlenbeck, Verloop, & Beijaard, 2002; Wubbels & Brekelmans, 2006). The assessment procedure should thus meet the following conditions. The specific work and classroom setting of the student teacher should be taken into account; a range of possible solutions and teaching styles should be allowed; both teacher behaviour and cognitions should be assessed; multiple sources of evidence should be consulted; and the evidence should be analyzed holistically. While several studies have shown that these conditions can be met within procedures for teacher assessment (Darling-Hammond & Snyder, 2000; Dwyer, 1995, 1998; Moss, Schutz, & Collins, 1998; Uhlenbeck et al., 2002), ensuring the reliability and validity of such complex assessment procedures nevertheless remains a problem.

In the present study, an interpersonal perspective (see Wubbels, Brekelmans, Den Brok, & Van Tartwijk, 2006) was adopted to develop an authentic assessment procedure which complies with the conditions mentioned above. The assessment procedure includes the Questionnaire on Teacher Interaction (QTI) and two additional instruments, namely a behavioural observation instrument and a self-evaluation instrument. The quality of such an authentic assessment procedure clearly depends on the quality of the assessors' decision-making and a small-scale study of the assessment procedure was therefore undertaken in order to examine the validity of the relevant judgement processes.

2.2 The assessment of interpersonal competence

Teacher competence is defined as the ability of a teacher to deal adequately with the demands of the teaching profession using an integrated set of knowledge, skills and attitudes as manifested in both the performance of the teacher and reflection on his or her performance. This

definition emphasizes the significance of the teacher's ability to reflect upon his or her own teaching practices as teachers — in our opinion — are only competent when they are able to reflect upon their own teaching practices and hence their professional development in addition to being able to teach well. This definition of teacher competence was thus taken, in addition to the aim of stimulating the professional development of student teachers within the domain of interpersonal competence, as the starting point for the selection of the assessment instruments to be used.

The Questionnaire on Teacher Interaction (QTI) (Wubbels et al., 2006) was selected for use as it provides a reliable and valid picture of both the student and teacher perspectives on the *stabilized* interpersonal relationships between the teacher and his or her students. Comparison of the student and teacher perspectives also provides information on the extent to which the teacher is aware of student perceptions of his or her behaviour. The QTI addresses the actual classroom context but does not consider specific situated examples of teacher behaviour while such examples are of critical importance for feedback purposes (i.e., formative assessment). In light of the above, an observation instrument was also included for assessment purposes. More specifically, the Model for Interpersonal Teacher Behaviour (MITB) (Wubbels et al., 2006, cf. Figure 1.1.) was adopted to develop an assessment framework to analyze situated teacher behaviour and identify the key components of interpersonal competence (i.e., specific competences).

The eight sector labels in the MITB (i.e., leadership, helping/friendly, understanding, student freedom and responsibility, uncertain, dissatisfied, admonishing and strict) apply to the stabilized interpersonal style of the teacher but not to concrete examples of situated behaviour or the competences which these reflect. And for this reason and in keeping with our definition of teacher competence, the following competences were distinguished (with the relevant sector from the MITB noted in parentheses).

1. Providing clear structure and guidance for students (DC leadership)
2. Setting norms and standards for students (DO strict)
3. Correction of undesirable student behaviour (OD admonishing)
4. Attention to students (CD helpful / friendly and CS understanding)
5. Provision of responsibility and freedom for students (SC student responsibility/ freedom)

In Figure 2.1, the five competences are depicted according to the submission-dominance (i.e., Influence) and opposition-cooperation (i.e., Proximity) dimensions of the MITB. As can be seen, the sectors OD (admonishing), DO (strict), DC (leadership), CD (helpful/friendly), CS (understanding) and SC (student responsibility /freedom) clearly translate into a competence. Although the sectors SO (uncertain) and OS (dissatisfied) *appear* to have been omitted, the behaviours associated with these sectors are included as counterexam-ples of the competences of (1) providing clear structure and guidance for students and (5) provision of freedom and responsibility for students.

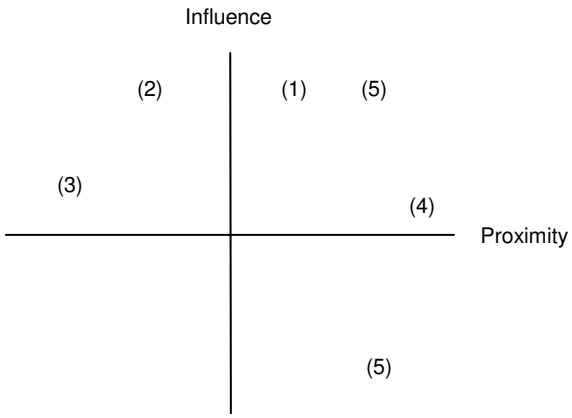


Figure 2.1 The five competences positioned according to the opposition-cooperation (i.e., Proximity) and submission-dominance (i.e., Influence) dimensions of the MITB. (1) Providing clear structure and guidance for students; (2) Setting norms and standards for students; (3) Correction of undesirable student behaviour; (4) Attention to students; (5) Provision of freedom and responsibility for students.

Given that interpersonal competence involves both performance and reflection on one's performance, both behaviour and reflection should be assessed to obtain a comprehensive picture of a teacher's interpersonal competence. For this reason, a self-evaluation instrument was also included to prompt teachers to reflect upon their performance and interpersonal competence. Reflection was operationalized as: (a) awareness of one's interpersonal behaviour and (b) the ability to monitor one's professional development.

2.3 Validity criteria for teacher assessment

In traditional test theory, *validity* is defined as the extent to which a test measures what it was designed to measure. Such a conception of validity is most appropriate within the context of traditional multiple-choice testing where the validity of the judgement depends mainly on the *construct representation* of the test itself as reflected by the content assessed and the assessment criteria employed. For authentic assessment, the content and criteria remain important while the quality of the assessment as a whole is considered with respect to the "ability of assessors to use these criteria to reach technically and professionally defensible conclusions" (Dwyer, 1995, p.77). For almost two decades now and in response to the widespread use of so-called "performance assessment," the validity of this form of assessment has been open to discussion. Some authors recommend a broader conception of validity (e.g., Messick, 1989) while others call for the use of more specialized criteria for purposes of authentic assessment (Frederikson & Collins, 1989; Linn, Baker, & Dunbar, 1991; Haertel, 1991; Gipps, 1994).

In the present study, we specifically considered two major threats to the validity of assessment as pointed out by Messick (1995): construct-irrelevant variance and construct under-representation. In cases of *construct-irrelevant variance*, the assessment is too broad and thus contains variance associated with other — irrelevant — constructs. In cases of *construct under-representation*, the assessment is too narrow and thus fails to capture critical aspects of the target construct. These possibilities apply to not only the assessment criteria themselves but also to the decision-making of assessors with respect to assessment criteria. As studies by Heller, Sheingold, and Myford (1998) and Schutz and Moss (2004) have shown, *construct-irrelevant variance* can arise when assessors apply extraneous, irrelevant or idiosyncratic criteria which are not a part of the rating model on which they have been trained and assessors consider evidence which is simply irrelevant to the construct in question. *Construct under-representation* is characterized by the omission of particular assessment criteria and/or evidence or the idiosyncratic weighting of certain criteria and/or evidence which may result in insufficient attention to particular aspects of performance. In the present study, the occurrence of construct-irrelevant variance and construct under-representation were therefore explicitly analyzed with respect to two essential judgement processes: (1) consideration of evidence from separate sources and (2) the combination of evidence from different sources to attain an overall judgement.

2.4 Method

2.4.1 Participants

Four individuals who were currently teacher educators (N=2) or supervising the training of secondary teachers (N=2) were selected to participate in the present study. These assessors were selected because of experience with use of the QTI for supervision purposes *and* their experience with supervision in general. Two of the assessors also had specific experience with the assessment of teacher competence. The four assessors followed a two-evening preparation programme in which both the content and process aspects of the assessment procedure were addressed. The content aspects included a description of the interpretive framework and its background. The process aspects included the need to take different sources of evidence into consideration, combine evidence from different sources to attain an overall judgement and clearly communicate one's conclusions.

2.4.2 The student teacher case

Using the QTI, the first author of this article selected a female student teacher with 9 months of experience teaching part-time and a Tolerant interpersonal profile in the opinion of her students (cf. Figure 1.2.). This teacher was judged to resemble the average student teacher because most student teachers have been found to be of the Tolerant or Uncertain/Tolerant type.

A lesson taught by the student teacher was next videotaped for observation purposes (i.e., presentation to the four assessors) and self-reflection purposes (i.e., presentation to the teacher herself). Finally, the student teacher was asked to complete a self-evaluation assignment (see sector 2.4.3.3). In such a manner, the four assessors assessed the same student teacher.

2.4.3 Assessment procedure

2.4.3.1 Questionnaire on Teacher Interaction (QTI)

The first instrument administered to assess the interpersonal competence of the student teacher was the QTI, which provides information on (a) student perceptions of the teacher's interpersonal style and (b) the teacher's own self-perceptions.

The assessors in the present study were asked to compare the student perceptions with the profile of the average student teacher, as provided in the QTI report. The assessors were also asked to compare the student perceptions of the student teacher with the student

teacher's own self-perceptions in order to attain an indicator of the student teacher's awareness of her own interpersonal competence.

2.4.3.2 Observation instrument

The second instrument used to assess the interpersonal competence of the student teacher involved the observation of actual situated behaviour. The five competences described above provided the framework for the assessors to analyze the videotaped lesson. Both examples and counterexamples of competent behaviour were provided for illustrative purposes for each of the competences. It was assumed that the provision of counterexamples in addition to confirmatory examples of adequate behaviour would help the assessors attain a comprehensive picture of the teacher's competence and thereby help them determine the level of competence. It was also assumed that referral to concrete counterexamples might help student teachers recognize inadequate behaviour on their part and thereby promote their professional development (Berliner, 2002). It should be noted that the focus in the counterexamples was on those interpersonal problems or pitfalls which characterize student teachers in particular. Illustrative examples and counterexamples are presented for each of the five competences in Table 2.1.

For each of the competences, one or two *guiding questions* were formulated to help focus the assessor's observational analyses. The responses to these questions should provide an extensive description of both concrete examples and counterexamples of competent behaviour for the five competences. For example, the following guiding questions were formulated for the competence of correction of undesirable student behaviour: How does the teacher deal with undesirable situations? and How does the teacher give a sequel to his or her interventions?

Table 2.1 Examples and counterexamples for five competences

1. Providing clear structure and guidance for students

Examples. The teacher:

- is able to make students active learners;
- keeps plenary moments functional;
- realizes smooth transitions between parts of the lesson.

Counterexamples. The teacher:

- appears insecure;
- allows students to frequently interrupt during plenary moments;
- shows a lack of overview over the class;
- involves only part of the class.

2. Setting norms and standards for students

Examples. The teacher:

- makes clear what behaviour is expected of the students;
- acts as a role model;
- formulates rules positively ("please do" as opposed to "do not").

Counterexamples. The teacher:

- sets rules which students perceive as unreasonable;
- communicates distrust in students;
- accompanies the explication of norms and standards with aggressive behaviour.

3. Correction of undesirable student behaviour

Examples. The teacher:

- shows a broad repertoire of low-intrusive corrections;
- aims corrections at student's behaviour and not student's personality;
- checks whether students respond to his or her corrections.

Counterexamples. The teacher:

- looks away while correcting student's behaviour;
- shows "more of the same" corrections;
- does not grant students a new chance after correction.

4. Attention to students

Examples. The teacher:

- shows personal interest in students;
- is patient with students;
- is approachable for students.

Counterexamples. The teacher:

- shows clear preferences in his or her treatment of individual students;
- tries to act as "student-with-the students";
- addresses students in a disrespectful manner.

5. Provision of freedom and responsibility for students

Examples. The teacher:

- gives the students an appropriate level of responsibility;
- gives the students an appropriate level of freedom;
- is able to return to his or her leadership role when he or she chooses so.

Counterexamples. The teacher:

- when providing responsibility and freedom, shows behaviour which students perceive as insecure or weak;
- sticks too rigorously to his or her own ideas;
- cuts any and every initiative on the part of students short.

2.4.3.3 Self-evaluation instrument

The third assessment instrument was designed to allow the assessors to assess the quality of the teacher's self-evaluation. The teacher was asked to analyze both the results of the QTI and the videotaped lesson. Next, the teacher was asked to provide: (a) an overview of her strengths and weaknesses with as much referral to the QTI results and the videotaped lesson as possible and (b) a personal professional development plan.

Once again, a set of guiding questions was formulated to keep the evaluations provided by the assessors in line with our operationalization of the concept of reflection. The guiding questions for the self-awareness aspects of the assessment procedure were as follows: Does the teacher show an awareness of her strengths and her weaknesses? and Is the teacher able to formulate an adequate analysis of her behaviour?

2.4.4 Data collection and analysis

The first author examined the QTI results, observation results and self-evaluation results and then produced a so-called expert report on the student teacher. This report is presented in Table 2.2. The judgement process was next analyzed for the four assessors with the aid of the following sources of data and instruments:

1. any *notes* made by the assessors with regard to the QTI results, the videotaped lesson and the student teacher's self-evaluation;
2. the overall judgement of the student teacher's interpersonal competence provided by the assessors in the form of an *evaluative summary* and a *numerical judgement* along a 10-point scale;
3. and the results of an *open interview* in which the assessors were asked to explain their overall judgement with specific attention to (a) the notation of evidence from the QTI, the videotaped lesson and the student teacher's self-evaluation and (b) the combination of evidence to attain an overall judgement.

Table 2.2 Expert report on the student teacher

The QTI shows a Tolerant profile with the SO sector (uncertain) being relatively more filled but not as much as for an Uncertain/Tolerant profile. The videotaped lesson reveals a similar picture. With respect to the Influence domain, although some difficulties are encountered during the lesson, the teacher appears to be confident and shows basic competence. The teacher's behaviour therefore seems to indicate improvement. In the self-evaluation, the teacher also shows an awareness of some important weaknesses in this domain. At the moment, however, she is not able to sufficiently analyze her own behaviour or formulate concrete plans for improvement. We nevertheless consider this adequate for a student teacher. With respect to the Proximity domain, the teacher shows a basic ability to pay attention to students. Nevertheless, improvement is possible with regard to paying more personal attention to students and dividing one's attention more equally across the students. In her self-evaluation, the teacher expresses satisfaction with her behaviour in this domain and does not mention any weaknesses. The relatively greater amount of attention being paid to the Influence domain of teacher interaction and the dominant aspect of behaviour in particular is common among student and beginning teachers as their problems mainly arise in this domain. Both behaviour and reflection within this domain were thus judged to be sufficient.

For each of the assessors, a description of the judgement process was created for the following two essential aspects: (1) consideration of evidence from separate sources and (2) the combination of evidence from different sources into an overall judgement. The description of *consideration of evidence from separate sources* was based on the transcribed notes of the assessors for the different assessment instruments. The description of *combination of evidence from different sources to obtain an overall judgement* was based on statements from the assessor's evaluative summary and the interview transcript.

To gain insight into the consideration of evidence from separate sources, the *notes* of the assessors were categorized in terms of the following: (a) the five competences derived from the MITB; (b) the MITB dimensions of influence and proximity; and (c) a reflection category . To gain insight into the combination of evidence from different sources to attain an overall judgement, the *statements* of the assessors were categorized in terms of the influence, proximity and reflection domains of interpersonal competence. Finally, the occurrence of *construct-irrelevant variance* and *construct under-representation* during the processes of considering evidence from separate sources and combining evidence from different sources was analyzed using the coding scheme outlined in Table 2.3.

Table 2.3 Coding scheme

Judgement process	**Validity criteria**	**Indicators**
Notation of evidence from separate sources	No construct-irrelevant variance	Evidence refers to only the relevant competences: no extraneous, irrelevant criteria are added; evidence is also connected to the appropriate competence.
	No construct under-representation	All of the competences and components of these are covered, and all of the relevant evidence is considered.
Combination of evidence from different sources to attain an overall judgement	No construct-irrelevant variance	Arguments refer to only the relevant competences: no extraneous, irrelevant criteria are added.
	No construct under-representation	All of the competences and components of these are covered, and all of the relevant evidence is considered.

2.5 Results

2.5.1 Overview of judgements

In Table 2.4, a compact overview of the judgements provided by the assessors for the influence, proximity and reflection domains of interpersonal competence is presented. As can be seen, all of the assessors with the exception of assessor B assessed the competence of the student teacher to be "sufficient." Whereas assessors C and D concluded that the teacher's competence in the domain of proximity was stronger than in the domain of influence, assessor A concluded just the opposite. In the following, we will further illuminate the conclusions of the assessors by summarizing their statements with regard to the influence, proximity and reflection domains of the student teacher's interpersonal competence.

Table 2.4 Summary of overall judgements provided by assessors

	Assessment domain			**Numerical judgement**
Assessor	Influence	Proximity	Reflection	
A	Sufficient	Sufficient	Sufficient	*7*
B	Insufficient	Insufficient	Insufficient	*4*
C	Sufficient	Sufficient	Sufficient	*7*
D	Sufficient	Sufficient	Sufficient	*6*

2.5.1.1 Influence

As can be seen from Table 2.4, assessors A, C, and D judged the student teacher's competence within the domain of influence to be sufficient. The assessors acknowledge the QTI results in connection with the videotaped lesson (i.e., in connection with the situated behaviour of the student teacher). *Assessor A* noted the following example of competence: "The student teacher has no difficulties in beginning to speak, and students listen rather quietly to her plenary introduction." The following counterexamples were also noted: "An inadequate repertoire of corrections with a higher intensity, the student teacher's recurrent invisibility and a lack of overview over the classroom affect her control over the classroom." *Assessor C* stated that the student teacher "obviously has difficulties providing leadership: the structure is not clear for the students, it is not clear what is allowed and not allowed, and disruptive student behaviour is corrected only occasionally." However, the assessor also noted a number of indicators for future professional development in this domain: "The teacher does not appear insecure, students listen to her relatively long plenary introduction and she speaks clearly and calmly." These examples correspond to those of Assessor A. Unlike Assessors A and C, who mentioned a set of concrete examples of competent behaviour, *Assessor D* only mentioned counterexamples of competence: "Students lose their attention primarily due to pedagogical problems. Moreover, the teacher provides too little guidance and loses the rest of the class when she interacts with individual students. The teacher does not set rules and does not correct student behaviour." Assessor D further notes that "nevertheless, it never becomes really chaotic." *Assessor B* was the only assessor to assess the student teacher's competence as insufficient in the domain of influence and based this conclusion on the results of the QTI which showed the student teacher to have less leadership (DC) than the *average teacher*. In the overall judgement provided by assessor B, moreover, the following counterexamples of competence are noted: "The student teacher only provides guidance with respect to content, not interpersonally. She hardly sets any standards and ignores disorder." In keeping with assessors A and C, however, assessor B further notes that the student teacher appears to be secure during the videotaped lesson.

2.5.1.2 Proximity

While assessors A, C and D all rated the competence of the student teacher within the domain of proximity to be sufficient, *assessor A* did not fully acknowledge the positive QTI results for the sectors of CD (helping /friendly) and CS (understanding) in the student teacher's situated

behaviour. Assessor A stated that "although [the student teacher] clearly gets student recognition, acts respectful and states in her self-evaluation to be very satisfied with her relationship with the students, to me she appears primarily distant and detached. She fails to see students' signals, ignores difficult students and does not have any chats with students." Along these lines, *assessor C* stated that she does "not see much interaction with students. Moreover, interaction always seems to be initiated by the students instead of by the teacher. Nonetheless, the student teacher is patient, responds to questions, rewards and walks around the classroom during coursework." *Assessor D* noted that "students consider this teacher 'a good sport'. The atmosphere is open, friendly and trusting." As a counterexample of competence, assessor D noted that the student teacher "frequently pays more attention for her own work than to students and that attention is paid to only those students who ask for it." *Assessor B* assessed the student teacher's competence within the domain of proximity to be insufficient and observed that "although [the student teacher] is patient, praises students frequently, and is helpful, I perceive a considerable distance between teacher and students. I cannot sense a good rapport. For example, when students enter the classroom, the teacher does not call the students by their names and does not make any jokes or chat with them." It should be noted that all of the preceding examples and counterexamples of competence within the domain of proximity were similarly mentioned by the other assessors.

2.5.1.3 Reflection

Just as for the domains of influence and proximity, assessors A, C, and D judged the student teacher's competence within the domain of reflection to be sufficient but assessor B did not. The four assessors generally agreed that although the student teacher showed awareness of her strengths and weaknesses, she was not able to analyze her behaviour adequately or formulate concrete plans for development. The comments of *assessor D* reflected precisely these general statements, and this assessor did not provide any further comments. *Assessor A* further noted with regard to the student teacher's self-evaluation that "no weaknesses in the domain of the relationship with the students whatsoever are mentioned" and clearly stated that she considered this "rather problematic." She concluded that the student teacher's self-awareness in the domain of influence was better than in the domain of proximity. *Assessor C* expressed some doubts about the teacher's ability to change "…because the student teacher herself expresses doubts in her self-evaluation. The teacher could be reluctant to change her behaviour." Nevertheless, assessor C thinks that the student teacher will be able to develop "as she notes

problems with her own behaviour." *Assessor B*, in contrast, places considerable weight on the fact that the student teacher could not adequately analyze her own behaviour and therefore could not formulate concrete solutions for problems. According to this assessor, this is simply unacceptable for even a student teacher. Just as assessor C, assessor B found it difficult to assess the student teacher's potential for future professional development. Assessor B observed that "[the student teacher] often looks for excuses for not realizing her ideal.

2.5.2 The validity of the judgements

In Table 2.5, the reader can find a summary of the extent to which the four assessors complied with the validity criteria of avoiding construct-irrelevant variance and construct under-representation during the assessment processes of noting evidence from different sources and combining evidence to attain an overall judgement.

Table 2.5 Validity of overall judgements provided by four assessors

	Assessor A	**Assessor B**	**Assessor C**	**Assessor D**
(1) Noting evidence				
Construct-irrelevant variance				
(a) evidence refers to content criteria				Pedagogical notes included
(b) evidence is related to appropriate criterion				
(c) observations and interpretation are adequate				
Construct under-representation				
(d) all critical components are considered	Examples of competence 4 under-represented	Examples of competence 1 under-represented	Counterexamples of competence 1 under-represented	Counterexamples of competences 1 and 4 under-represented
(2) Combining evidence into an overall judgement				
Construct-irrelevant variance				
(e) no extraneous, irrelevant criteria are added		Overly severe for Influence, Proximity and Reflection		Pedagogical criteria included
Construct under-representation				
(f) omission or underexposure of criteria does not occur	Overly severe for Proximity	Overly severe for Influence, Proximity and Reflection		

With regard to the process of *noting evidence,* it can be seen that *construct-irrelevant variance* rarely occurred. Only Assessor D appeared to provide a considerable number of notes which were irrelevant for the competence of providing guidance/clear structure (1). *Construct under-representation* was found to occur for the competences of providing guidance /clear structure (1)

and attention to students (4). The notes of three of the assessors, assessor B, C, and D, contained relatively little evidence for the competence of 1, while construct under-representation for the competence 4 occurred when assessed by both assessors A and D. Although such construct under-representation constitutes a source of invalidity, none of the assessors appeared to omit critical aspects of the relevant competences and the variance in their judgements thus remained acceptable. With regard to the process of *combining evidence to attain an overall judgement,* greater differences between the assessors were detected. Assessor A, who slightly under-represented competence 4 in her notes, expressed an overly severe judgement when compared to the expert judgement. Judge A did not acknowledge the positive QTI scores for the CD (helping/friendly) and CS (understanding) sectors of the questionnaire in her evaluation of the teacher's situated behaviour and appeared to explicitly ignore these results in the determination of her overall judgement. The same held for assessor B who also assessed the teacher too negatively in the BS sector of the MITB as well and, relative to the expert judgement, produced overall judgements which were overly severe in the domains of influence, proximity and reflection. For assessor C, who initially underrepresented counterexamples for competence 1 in his notes, the problem largely disappeared when he later included counter-examples which did not appear in his original notes. Finally, the construct-irrelevant notes of assessor D reappeared in her overall judgements. This assessor did not show construct under-representation in her overall judgements because competences 1 and 4 — which were under-represented in her original notes — were sufficiently represented in her overall judgement although the statements referring to the relevant competences were rather abstract.

2.6 Conclusions and discussion

The purpose of the present study was to develop and validate an assessment procedure to evaluate the interpersonal competence of student teachers. In the present paper, the validity of the judgement process was considered in particular. The results for four judges who evaluated the same student teacher using the assessment procedure showed the validity of their judgements to be satisfactory. The amount of construct-irrelevant variance and construct under-representation found for the assessment processes of (a) noting evidence and (b) combining evidence from different sources was small.

The results of the present study have clear implications for just how the reliability and validity of the assessment procedure used to evaluate teacher interpersonal competence can be

strengthened. First, the results showed the differences between the four assessors to stem primarily from the process of combining evidence from different sources to attain an overall judgement. Questions can thus be raised about just how evidence from the QTI (i.e., student perceptions of their relations with a teacher) should be combined with evidence from the observation of situated teacher behaviour. Student perceptions of their interpersonal relations with the teachers tend to stabilize relatively early in the school year, which means that significant changes in the situated behaviour of the teacher may not affect QTI results immediately. In other words, the assessors applying the present assessment procedure should certainly respect the QTI results but also recognize that the QTI results may reflect earlier interpersonal behaviour. Given that the interpersonal behaviour of the teacher may have changed in either a positive or a negative direction during the intervening period of time, the observational data may thus be of critical importance. Explicit discussion of this issue within the preparation program for assessors of interpersonal competence may similarly improve the validity of judgements. A second implication of the present results for just how the reliability and validity of the assessment procedure can be improved stems from the finding that assessor B generally applied an overly harsh standard in her evaluation of the student teacher. That is, a harsh overall standard accounted for the low overall scores assigned by this assessor and not major construct-irrelevant variance or construct under-representation. The inter-rater reliability of the assessment scores might thus be improved with the provision of benchmarked examples of student teachers with insufficient, sufficient and excellent interpersonal skills, respectively (see Gipps, 1994).

The present results also shed light on the added value of using the observation and self-evaluation instruments in addition to the QTI to assess the interpersonal competence of teachers. As already stated, use of the QTI alone does not provide a sufficiently comprehensive picture of the interpersonal competence of teachers. It was therefore argued that both concrete examples of situated behaviour and information on the ability of teachers to reflect upon their own interpersonal competence were also needed. Such additional information was indeed found to enhance the validity of the assessment procedure by providing a more authentic picture of the teacher's interpersonal competence and self-awareness. The results of the present study showed use of situated behaviour and self-evaluation as part of the assessment procedure to provide information which otherwise might not be attained. That is, small changes in the behavioural repertoire of the student teacher may not be revealed by two successive administrations of the QTI while behavioural observation and self-evaluation have

been found to provide detailed and sometimes critical information on the professional development of the teacher. The supplemental information could also serve an important feedback function by providing student teachers with concrete suggestions and directions for their further professional development.

In closing, the results of the present small-scale study showed the validity of the combined assessment procedure to be reasonable but also highlighted some domains for refinement of the assessment procedure and the preparation programme. Examination of the interpersonal competence of a variety of student teachers using the present assessment procedure would nevertheless provide further insight into the utility and validity of the procedure. Particularly challenging student teachers may elicit greater disagreement among assessors, for example, and raise unforeseen issues which have yet to be covered within the assessment framework (see Moss et al., 1998) or assessor preparation programme.

References

Berliner, D. C. (2002). Learning about and learning from expert teachers. *International Journal of Educational Research, 35*, 463-482.

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

Dwyer, C.A. (1995). Criteria for performance-based teacher assessments: Validity, standards and issues. In A.J. Shinkfield & D. Stufflebeam. *Teacher evaluation: guide to effective practice* (pp. 62-80). Boston: Kluwer Academic Publishers.

Dwyer, C. A. (1998). Psychometrics of Praxis III: Classroom Performance Assessments. *Journal of Personnel Evaluation in Education, 12*, 163-187.

Evans, E.D., & Tribble, M. (1986). Perceived teaching problems, self-efficacy, and commitment of teaching among preservice teachers. *Journal of Educational Research, 80*, 81-85.

Evertson, C.M., & Weinstein, C.S. (2006). Classroom management as a field of inquiry. In C.M. Evertson & C.S. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 3-16). Mahawn: Lawrence Erlbaum Associates.

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

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational measurement*. London, Washington D.C.: The Falmer Press.

Haertel, E. H. (1990). Performance Tests, Simulations, and Other Methods. In L. M. Darling-Hammond, J. (Ed.), *The New Handbook of Teacher Evaluation: Assessing Elementary and Secondary School Teachers*. London: Sage.

Heller, J. I., Sheingold, K., & Myford, C.M. (1998). Reasoning about evidence in portfolios: Cognitive foundations for valid and reliable assessment. *Educational Assessment, 5*, 5-40.

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

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

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

Moss, P. A., Schutz, A., & Collins, K. M. (1998). An integrative approach to portfolio evaluation for teacher licensure. *Journal of Personnel Evaluation in Education, 12*, 139-161.

Schutz, A., Moss, P.A., (2004). Reasonable decisions in portfolio assessment: Evaluating complex evidence of teaching. *Education Policy Analysis Archives, 12*. Retrieved 7/19/2004 from http://epaa.asu.edu/v12n33/.

Uhlenbeck, A. M., Verloop, N., & Beijaard, D. (2002). Requirements for an assessment procedure for beginning teachers: Implications from recent theories on teaching and assessment. *Teachers College Record, 104*, 242-272.

Veenman, S.A.M. (1984). Perceived problems of beginning teachers. *Review of Educational Research, 54,* 143-178.

Wubbels, Th., & Brekelmans, M. (2006). Two decades of research on teacher-student relationships in class (introduction). *International Journal of Educational Research, 43*, 6-24.

Wubbels, Th., Brekelmans, M., Den Brok, P., & Van Tartwijk, J. (2006). An interpersonal perspective on classroom management in secondary classrooms in the Netherlands. In. C. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: research, practice and contemporary issues* (pp. 1161-1191). New York: Lawrence Erlbaum Associates.

3

The nature of the collaborative judgement process[2]

Abstract

Given that the assessment of student teachers is generally based on non-standardized, qualitative information derived from multiple sources, the validity of the assessment process largely depends on the judgement capacities of the assessors. It has recently been suggested that the quality of the assessment process can be improved via collaboration in pairs, which is a common practice in higher education. However, there is little empirical research on the nature of the collaborative assessment process or the ways in which such collaboration can enhance the validity of assessment. In the present study, 12 assessor pairs were asked to collaboratively judge the same student teacher. The assessment process was subsequently characterized in terms of the specific communicative activities engaged in. Four different types of collaborative assessment processes could be distinguished and, for each of these, a number of strengths and weaknesses could be identified. The implications of this information for the validity of collaborative assessment processes and the preparation of assessors are discussed in closing.

[2] This chapter has been submitted in adapted form as:
Nijveldt, M., Brekelmans, M., Wubbels, Th., Verloop, N., & Beijaard, D. The nature of the collaborative judgement process in the assessment of student teachers.

3.1 Introduction

Authentic assessment of teaching is widely acknowledged to help promote the professional development of teachers. Such assessment can clarify expectations with regard to the teaching task and provide feedback on the functioning of teachers. Such assessment can also encourage reflection and meaningful dialogue or debate with regard to the essence of good teaching (Darling-Hammond & Snyder, 2000; Delandshere & Arens, 2003; Dwyer & Stufflebeam, 1996). Professional development is currently accepted as not only an explicit goal of formative assessment but also summative assessment. In order to successfully promote teacher learning, however, the relevant assessment procedures must clearly meet a number of conditions. First, the assessment procedures should reflect the complexity of the teaching task in order to be meaningful and constitute a valuable learning experience for teachers (e.g., Darling-Hammond & Snyder, 2000; Dwyer, 1995; Uhlenbeck, Verloop, & Beijaard, 2002). Second, detailed feedback should be provided and the assessors should clearly motivate their judgements and conclusions (Moss, 1994; Tigelaar, Dolmans, Wolfhagen, & Van der Vleuten, 2005).

Given that authentic teacher assessment is generally based on non-standardized, qualitative material derived from sometimes widely differing sources (e.g., a videotaped lesson and written reflections on performance), often originating from a variety of contexts, the assessment process is highly complex. The richness and complexity of the relevant data call for a holistic approach to scoring or, in the words of Delandshere and Petrosky (1994, p.13), "pieces of performance can only be analyzed, interpreted and evaluated in the context of the whole performance, because their significance is determined by that context." The specific teaching context and personal goals of the teacher should be taken into consideration and, when this is done, the assessment procedure is generally taken to have a high degree of validity (e.g., Hager, Gonczi, & Athanasou, 1994). Yet, even though teacher assessment is commonly embedded in a clear conceptual framework (e.g., Dwyer, 1995; Mislevy, Steinberg, Breyer, Almond, & Johnson, 2002), assessors may still attend to evidence which is largely irrelevant to the construct of interest and thereby introduce "construct-irrelevant variance" into the judgement process or simply fail to address all of the assessment criteria and the relevant evidence, which produces "construct under-representation." For valid assessment, however, all of the relevant data must be considered and, in this connection, all conclusions should be supported by clear argumentation which also draws upon the original data (Kane, 1992; Moss, 1994).

In order to minimize the aforementioned threats to validity and improve the teacher assessment process, authors have increasingly suggested that teacher evaluation should be undertaken collaboratively in pairs or even larger teams (e.g., Johnston, 2002; Moss, Schutz, & Collins, 1998; Tigelaar, 2005) That is, "the validity of the conclusion is warranted, in part, in the consensus among [assessors] who are empowered to challenge one another's developing interpretations in light of the cases at hand" (Moss et al., p.142). Collaborative assessment is common practice for the evaluation of complex tasks in higher education and considered very promising in theory, but little empirical research has addressed the exact nature of the collaborative judgement process or the ways in which such collaboration can actually enhance the validity of the assessment process relative to individual assessment.

In the present chapter, we will examine the processes which assessors engage in when collaboratively (i.e., pairwise) assessing a student teacher using a specific assessment procedure. First, an overview of the specific activities undertaken by the pairs of assessors will be provided. Second, the extent to which the collaborative judgement processes reflect two of the key assessment processes as mentioned by Moss et al. (1998) will be examined: (a) collaboratively seeking coherence among the available sources of evidence and checking whether all relevant evidence has been taken into account, and (b) challenging developing interpretations with counterevidence and/or the perspectives of other assessors. Engagement in these two evaluation processes is assumed to minimize the aforementioned threats to validity, namely the introduction of construct-irrelevant variance and/or construct under-representation.

3.2 Collaborative judgement processes

In the assessment literature, numerous studies have addressed the underlying cognitive judgement processes. In this section, we will provide a general overview of those judgement processes considered essential for collaborative assessment, which includes some of the processes essential for the individual assessment process. The relevant studies are of an empirical character (e.g., Heller, Sheingold, & Myford, 1998; Moss, Schutz, & Collins, 1998; Quinlan, 2002; Schutz & Moss, 2004) or a more theoretical, conceptual character and frequently have a focus on the identification of the criteria for high quality assessment (e.g., Delandshere & Petrosky, 1994, 1998; Johnston, 2002; Mabry, 1999; Tigelaar et al., 2005). Research by Moss, Schutz, and Collins (1998) and Schutz and Moss (2004), on assessors' judgement processes in the context of teacher portfolio assessment, has specifically shown that the assessors cannot

avoid interpreting parts of the teacher portfolio in terms of a more cohesive understanding of the whole. Stated differently, assessors cannot avoid forming a pattern out of the data from a teacher portfolio in order to evaluate it. And once a particular pattern has been identified, subsequent information may also be interpreted in terms of this pattern. "Even if the subsequent data disconfirms the initial pattern, it is read to some extent in response to the pattern it disconfirms" (Schutz & Moss, pp. 8-9). This strong tendency to seek confirmation of one's initial interpretations can be seen to constitute a bias at times, which can then most certainly affect the validity of the judgement process and led Moss et al. to introduce two interrelated principles to guide the assessment process. The first principle is that assessors should search for coherence among the evidence coming from different sources and explicitly check to see that all of the relevant evidence has been taken into account. The second principle is that assessors should be encouraged to actively seek and consider counterevidence and thus evidence which challenges their developing interpretations and explicitly entertain alternative interpretations. Adherence to these two principles can particularly be stimulated by collaboration (i.e., discussion between assessors), and Moss et al. therefore suggest that assessors first work through the available data individually (i.e., individually note the relevant evidence) and then work together in pairs to prepare a collaborative judgement (i.e., an evaluative summary) (also see Delandshere & Petrosky, 1994). During this process, the assessors should be encouraged to note not only supportive evidence and examples but also conflicting evidence and counterexamples, to discuss alternative interpretations and to revise their interpretations until all of the relevant evidence has been taken into account. In practice, challenging developing interpretations has been found to constitute a major problem for assessors who primarily seek strictly confirmatory evidence for tentative hypotheses (Moss, Schutz, & Collins, 1998) or simply do not try to confirm or disconfirm their hypotheses at all (Quinlan, 2002).

The aforementioned principles of 'seeking coherence among the available sources of evidence and checking to see that all relevant evidence has been taken into account' and 'challenging developing interpretations with counterevidence and/or the perspectives of other assessors' clearly bear on the validity of the assessment process and can therefore be considered *key processes* for collaborative assessment. In addition to these key processes, however, the following assessment processes will also be considered within the context of the present study: the weighing of the importance of available evidence (i.e., determination of relative importance) and the use of a standard (i.e., evaluation of a candidate in terms of some standard or norm) (e.g., Dwyer, 1995; Heller et al., 1998).

The collaborative judgement processes of the pairs of assessors considered in the present study will be further characterized in terms of the specific communicative activities engaged in. In the literature on collaborative knowledge construction, communicative activities are often analyzed in terms of the types of proposals or contributions put forth and the types of responses offered in return. In such a manner, four categories of activities can be distinguished: (a) contribution, (b) acceptance of contribution, (c) discussion of contribution and (d) ignoring of contribution (cf. Barron, 2003). Translated to the setting of collaborative teacher assessment, contributions can include the presentation of evidence regarding teacher competence, the interpretation of evidence regarding teacher competence or the evaluation of teacher competence with respect to a particular standard. Acceptance of the contribution may then include agreement and/or the provision of additional evidence to support the contribution. Discussion of the contribution can include questioning or rejection of an interpretation, presentation of counterevidence and/or discussion of the relative importance of the evidence presented. Within the context of the present study, ignoring the initial contri-bution will be limited to ignoring *discussion* of contributions — for example — counterevidence which has been presented.

In sum, the following general research question was addressed in the present study: What is the nature of the assessment process when teachers are assessed collaboratively? The following, more specific, research questions were addressed: (a) Which communicative activities do assessors engage in while collaboratively assessing teacher competence? (b) To what extent are the key processes of 'seeking coherence among the available sources of evidence and checking to see that all relevant evidence has been taken into account' and 'challenging developing interpretations with counterevidence and/or the perspectives of other assessors' reflected in the observed collaborative judgement processes? and (c) To what extent does the representation of the key processes vary across the different pairs of assessors?

3.3 Method

3.3.1 Assessment procedure

In the present study, we analyzed the judgement processes of assessors who evaluated the "interpersonal competence" of a student teacher, using a specific assessment procedure. Interpersonal teacher competence or the quality of the interpersonal relationships with students is a major concern for many student and beginning teachers (Evans & Tribble, 1986; Evertson & Weinstein, 2006; Veenman, 1984). To help the assessors attain a comprehensive picture of the teacher's interpersonal competence, a concrete assessment framework in which six key components of interpersonal teacher competence are identified was provided: (1) providing clear structure and guidance for students, (2) setting norms and standards for stu-dents, (3) correction of undesirable student behaviour, (4) attention to students, (5) provision of freedom and responsibility for students and (6) reflection upon one's interpersonal performance. The six components of interpersonal teacher competence are illustrated with both examples and counterexamples of interpersonal competence (see Appendix).

The assessment procedure involved three sources of data: (1) the Questionnaire on Teacher Interaction (QTI), which provides a reliable and valid picture of both the student and teacher perspectives on the interpersonal relationships between the teacher and his or her students (see Wubbels, Brekelmans, Den Brok, & Van Tartwijk), (2) a videotaped lesson selected by the teacher or, in the present study, student teacher and (3) a self-evaluation assignment in which the student teacher is asked to analyze both the results of the QTI and the videotaped lesson to provide (a) an overview of strengths and weaknesses with as much reference to the QTI results and videotaped lesson as possible and (b) a personal professional development plan.

The assessors were first instructed to individually note evidence regarding interpersonal teacher competence from the three data sources. They were then asked to integrate this information into an individual judgement in the form of an evaluative summary and a numerical rating along a 10-point scale. Finally, the assessors were asked to enter into a collaborative judgement process and formulate, in pairs, a definitive summary of the teacher's competence and provide a definitive numerical judgement.

3.3.2 Assessors

Twenty-four individuals who were currently teacher educators (N=4) or teachers supervising the training of prospective secondary school teachers (N=20) were selected to participate in the present study. All of the individuals had experience with the evaluation of student teachers and/or the provision of feedback on the performance of student teachers. Five of the individuals also had specific experience with the assessment of teacher competence.

Prior to their participation in the present study, all of the aforementioned individuals followed a two-evening training programme. During the first evening, the procedure to assess the interpersonal competence of the student teachers was introduced and the assessment framework was discussed. During the second evening, the assessors conducted an entire assessment using an authentic case (i.e., data provided by a student teacher). Following completion of the assessment task, the experiences of the assessors were exchanged in a plenary discussion with specific attention to the difficulties encountered.

3.3.3 Data collection

Following participation in the training programme, the 24 assessors were invited to our institute in groups of 8 and asked to evaluate a new case provided by a student teacher in pairs with a time limit of four hours. All of the 12 assessor pairs evaluated the same student teacher. In keeping with the assessment procedure, the assessors individually noted evidence from the three different sources of data and prepared an individual evaluative summary, including a numerical rating. The assessors then entered into a collaborative judgement process to formulate a definitive summary of the teacher's interpersonal competence and translate this summary into a numerical rating. The collaborative judgement process was audiotaped.

3.3.4 Analyses

The discussions undertaken by the different pairs of assessors were fully transcribed and analyzed in three steps. In the first step, a classification system was generated to characterize the communicative activities which occurred during all of the discussions. A provisional list of categories was formulated and those discussions showing clear variation in the use of communicative activities at first sight (N=5) were then coded by the first author using the aforementioned provisional list. Codes were added as needed during this process, which resulted in the formulation of a comprehensive classification system.

In the second step, two of the five discussions coded by the first author were coded by a second researcher using the comprehensive classification system. The differences between the two coders were discussed until a consensus could be reached. Based on these coding discussions, the definitions for some of the coding categories were refined and examples to clarify what was meant were added to the classification system. All of the transcripts were next coded by the first author using Atlas-ti, which is a software program for qualitative analyses. This coding was verified by the second researcher. Agreement occurred on nearly 90% of the assigned codes; the remaining codes were discussed until a consensus could be reached.

In the third step of the analyses, the frequencies of the different communicative activities were taken to be an indicator of the nature of the collaborative judgement process. The absolute frequencies for the 12 assessor pairs were compared and the pairs were further categorized according to the occurrence of those activities which best reflect the two key processes for collaborative assessment. In such a manner, four types of collaboration were distinguished. And for each type of collaboration, we examined which specific communicative activities occurred to a more or less frequent extent when compared to the entire group mean. If the mean frequency of a communicative activity was found to be more than 0.8 standard deviation higher or lower than the overall mean (cf. Cohen, 1988), the activity was considered typical of that specific type of collaboration. In addition, the overall nature of the corresponding judgement processes were described more qualitatively, which included identification of specific strengths and weakness with respect to the validity of the assessment process.

3.4 Results

3.4.1 Communicative activities

The first and second steps in the analyses led to the formulation of the classification system presented in Table 3.1. This classification system represents the communicative activities which the various assessors engaged in while collaboratively assessing teacher competence. As can be seen, the communicative activities are grouped into four general categories of collaborative interaction: (A) contribution (16 categories), (B) acceptance of contribution (3 categories), (C) discussion of contribution (3 categories, 8 subcategories) and (D) ignoring of discussion. For each category of communicative activity, moreover, the total frequency of occurrence, mean per collaborative pair, range and standard deviation is presented.

Table 3.1 Means, ranges and standard deviations for communicative activities of assessors

Activities / codes	Total	M	Range	SD
A. Contribution				
1. Present evidence	309	25.8	13 - 51	10.5
2. Present an interpretation of evidence	133	11.1	3 - 19	5.1
3. Present a judgement	143	11.9	8 - 18	3.1
4. Judge the candidate against a standard	40	3.3	0 - 10	3.7
5. Present a numerical judgement along a 10-point scale	48	4.0	2 - 6	1.5
6. Note coherence between available sources of evidence	21	1.8	2 - 4	1.5
7. Note incoherence between available sources of evidence	3	0.3	0 - 2	0.6
8. Note agreement between assessors	23	1.7	0 - 6	1.9
9. Note disagreement between assessors	1	0.0	0 - 1	0.3
10. Explain definition of interpersonal competence or aspects thereof	7	0.6	0 - 3	1.0
11. Check whether all important evidence has been taken into account	8	0.8	0 - 3	1.0
12. Suggest an approach to the judgement task	32	2.6	0 - 6	2.3
13. Request justification or clarification	20	1.7	0 - 5	1.9
14. Rephrase initial contribution	7	0.6	0 - 1	0.5
15. Request addition or confrontation	34	2.8	0 - 6	1.9
16. Judgement with regard to sufficiency of available evidence	4	0.3	0 - 4	1.1
B. Acceptance of contribution				
1. Agreement with description of reasons	78	6.5	0 - 13	4.9
2. **Provide additional evidence (ADDITION)**	124	10.3	0 - 16	5.3
3. Agreement with no justification	198	16.5	8 - 29	5.6
C. Discussion of contribution				
1. Reject contribution (evidence, interpretation, judgement)				
a. Rejection based on construct irrelevance	7	0.6	0 - 2	0.8
b. Rejection based on the quality of the evidence	3	0.3	0 - 1	0.5
c. Rejection based on the standard to be met	4	0.3	0 - 2	0.7
d. Rejection with no justification	16	1.3	0 - 4	1.5
2. Discuss importance or weight of contributions				
a. Based on the quality of the evidence	38	3.2	0 - 11	2.8
b. Based on the standard to be met	31	2.5	0 - 7	2.4
c. Based on the specific context	25	2.1	0 - 8	2.3
d. With no justification	33	2.8	0 - 9	2.9
3. **Challenge an interpretation via presentation of counterevidence or alternative interpretation (CHALLENGE)**	72	6.0	0 - 16	5.0
D. Ignoring of discussion	10	0.8	0 - 3	0.9

Inspection of the results in Table 3.1 shows the key assessment process of "seeking coherence among the available sources of evidence and checking whether all relevant evidence has been taken into account" to be reflected particularly in the communicative activity of "providing additional evidence" (B2). The key assessment process of "challenging developing interpretations with counterevidence and/or the perspectives of other assessors" is reflected in the communicative activity of "challenging an interpretation via the presentation of counterevidence or an alternative interpretation" (C3). The total occurrence of these communicative activities was high (i.e., 124 and 72, respectively) while the total occurrence of a number of other communicative activities was found to be relatively low. Within the general category of "contribution," for example, the following activities did not regularly occur: "note incoherence between available sources of evidence" (A7), "note disagreement between assessors" (A9), "check whether all important evidence has been taken into account" (A11) and "judge sufficiency of the available evidence" (A16). Within the general category of "discussion of contribution," rejections on the basis of "construct-irrelevance" (C1a), "the quality of the evidence" (C1b) and "the standard to be met" (C1c) were very infrequent. Although the assessors frequently discussed the quality of the evidence, the standard to be met and the specific context (C2), in other words, such discussions rarely led to explicit rejection of the relevant contributions.

3.4.2 Four types of collaborative assessment

In this section, the discussions conducted by the 12 pairs of assessors will be considered in greater detail. For this purpose, the 12 discussions were characterized in terms of the occurrence of the two key assessment processes, which are best reflected by the communicative activities of "providing additional evidence" (i.e., *addition*) and "challenging an interpretation via the presentation of counterevidence or an alternative interpretation" (i.e., *challenging*).

The collaborative judgement processes for the 12 pairs of assessors were found to differ considerably with respect to *both* the "provision of additional evidence" and the "challenging of an interpretation via the presentation of counterevidence or an alternative interpretation." Based on the observed patterns of addition and challenging, four types of collaborative assessment could be distinguished: collaboration involving both addition and challenging (Type I), collaboration involving challenging (Type II), collaboration involving addition (Type III) or collaboration involving neither addition nor challenging (Type IV).

Those communicative activities which occurred more than average or less than average for each of the different types of collaborative assessment processes will be considered in the following. Theoretically, assessment processes involving both addition and challenging may be most likely to minimize any threats to the validity of the assessment process and, for this reason, the particular strengths and weaknesses of the different types of collaborative assessment processes will also be considered in this regard.

3.4.2.1 Collaborative Assessment Type I: Addition and challenging

The collaborative judgement processes for 3 of the 12 pairs in the present study could be characterized as Type I. The occurrence of the key assessment components of addition and challenging was greater than average for pairs A, B and C. As can be seen from Table 3.2, moreover, a variety of communicative activities characterized the assessment process for Type I pairs with a considerable number of the activities occurring more than average but none of the activities occurring less than average. It can be further noted that pairs A, B and C accounted for all 3 of the rejections based on the quality of the evidence and 5 of the 7 rejections based on construct irrelevance. Such rejections clearly reflect the key assessment process of challenging developing interpretations.

Table 3.2 Communicative activities found to be above or below average for Type I Collaborative Assessment (addition and challenging)

Activities		**Mean**	
		Type I	**Overall**
Activities which occur **above** average:			
Contribution	Judge the candidate against a standard	8.7	3.3
	Explain definition of interpersonal competence	1.7	0.7
	Suggest an approach to the judgement task	5.0	2.6
	Request justification or clarification	3.0	1.7
	Judge sufficiency of the available evidence	1.3	0.3
Acceptance	Provide additional evidence	13.0	10.3
Discussion	Reject contribution on the basis of construct irrelevance	1.7	0.6
	Reject contribution on the basis of the quality of the evidence	1.0	0.3
	Challenge with counterevidence or alternative interpretation	10.7	6.0
Activities which occur **below** average:			
None			

The frequency of communicative activities specifically involving the discussion of a contri-bution was also above average for the Type I pairs. For pair A, 4 interactions involved discussion of contribution (i.e., C1, C2 and/or C3 activities) which subsequently resulted in clear conclusions for 3 of the 4 interactions. For pair B, 2 of the 3 interactions based on discussion involved a single discussion activity which was subsequently ignored by the assessors. The third interaction did not result in a consensus but nevertheless prompted specification of the evidence on which the assessment (i.e., interpretations and judgements) should be based. A fragment from this specific interaction is presented in Box 3.1. Finally for pair C, 3 of the 4 discussion of contribution interactions resulted in a clear conclusion; the fourth interaction involved a single discussion activity which was subsequently ignored by both parties.

Box 3.1 Illustration of Type I Collaborative Assessment: Addition and challenge (pair B)

Assessor 1	This candidate should especially work on her insecurity. When she becomes more secure, her performance will be much more powerful in a great many domains. *(A2)*
Assessor 2	Did you find her insecure? I myself didn't think so, actually. *(C1d)*
Assessor 1	Well, insecure in the sense of being rather invisible throughout the lesson. *(A1)*
Assessor 2	Yes, I agree… *(B3)* …but I really don't think she's tremendously insecure. She herself mentions that she would like to be a bit more secure, but I personally had the impression that this was just an attitude or stance. *(C3)*
Assessor 1	This is a form of insecurity. *(A10)*
Assessor 2	Well, yes, but this insecurity is not visible in the lesson. It's something which merely takes place in her own head. I mean, to me, she is clearly able to control the classroom. During the beginning of the lesson I was kind of worried: she enters the classroom and doesn't do anything for a considerable amount of time. I found this rather remarkable, not very effective. But then, without any difficulties, she starts to speak and the students listen to her. During plenary moments she was able to control the students with low-intrusive corrections: a single gesture or comment such as "ladies" suffices; the students stopped chatting immediately. But this situation changes as the plenary introduction becomes a bit too long. At this point, she doesn't know what to do. In other words, in the beginning, she controls the classroom very well but loses control at a certain point. This can be explained by her corrections: she's very good at low-intrusive corrections, but when such small interventions do not suffice, she doesn't know what to do. She starts raising her voice, without effect. *(C3)*
Assessor 1	Well, I guess that explains my perception of insecurity. Um, not intervening when necessary. I didn't hear her say anything like "come on, pay attention everybody" or "please do not interrupt me while I'm speaking." In this respect, she does not provide sufficient structure for the different parts of the lesson. *(B2)*
Assessor 2	Yes, that's true. *(B3)*

The strength of collaborative assessment encompassing both addition and challenging is that this can lead to judgements derived from both confirmatory and disconfirmatory evidence. The argumentation of the Type I pairs was also found to be relatively extensive and

transparent. However, the fact that the discussions of contributions did not always lead to clear conclusions constitutes a major shortcoming. In some cases, for example, the challenges of interpretations or judgements were simply ignored and all of the available evidence did not, therefore, contribute to the assessment process.

3.4.2.2 Collaborative Assessment Type II: Challenging

Only one of the 12 pairs, pair D, showed a judgement process which involved mainly challenging. Those communicative activities which occurred above or below average for this collaborative pair are summarized in Table 3.3.

Table 3.3 Communicative activities found to be above or below average for Type II Collaborative Assessment (challenging)

Activities		**Mean**	
		Type II	**Overall**
Activities which occur **above** average:			
Contribution	Explain definition of interpersonal competence	2.0	0.7
Discussion	Discuss importance of contributions based on the standard to be met	5.0	2.5
	Challenge via counterevidence or alternative interpretation	15.0	6.0
Activities which occur **below** average:			
Acceptance	Agreement with description of reasons	2.0	6.5
	Provide additional evidence	3.0	10.3

Three discussions could be distinguished although one of these was very extensive and encompassed almost the pair's entire interaction. This discussion resulted in an extensive conclusion in which the assessors elaborated on both those judgements on which they could clearly reach a consensus and those judgements on which they could *not* reach a consensus or sufficiently support. A fragment from the latter discussion is presented in Box 3.2.

Box 3.2 Illustration of Type II Collaborative Assessment: Challenging (pair D)

Assessor 1	The QTI… I just can't relate these student perspectives [on the teacher-student relationship] to what I see in the lesson. In her self-evaluation the teacher expresses that she is still very insecure and in the lesson this was visible in all possible ways. But it seems her students simply do not notice this. *(A7)*
Assessor 2	What *I* concluded from the self-evaluation is that at present, she is perfectly able to deal with her insecurity. She still has to work on her leadership skills, but she is fully aware of this and is therefore able to work this out. *(C3)* But then, you say you are just not convinced…. *(A13)*
Assessor 1	No. Firstly, in the self-evaluation, insecurity is expressed more than once and secondly, what concerns me most, she constantly comes up with excuses to not have to work on certain things in the end. [reads a passage from the self-evaluation as a concrete example] *(C3)*
Assessor 2	Well, I *did* ask myself if she merely *says* that she wants to change certain things or *really* wants to change her style. *(C2a)*
Assessor 1	… and at a certain point she also states that she wants to improve keeping the attention of the whole class during the plenary moments by not constantly paying attention to individual students. So there she proposes a clear plan. But then she says something like "this is not that easy to do in practice and I'm not completely sure if I would *like* to adopt this alternative approach." So I noted "This is going to fail!" *(C3)*
Assessor 2	Yeah, okay. But I really don't consider this a hopeless case: she is a student teacher and has sufficient time to develop professionally. *(C1c)*
Assessor 1	[reads another passage from the self-evaluation to illustrate excuses] *(C3)*
Assessor 2	I agree that this is not very strong *(B3)*, but I do not think it's that strange to express your uncertainties in a self-evaluation assignment. I really think she's aware of her strengths and weaknesses with respect to the area of providing guidance and she's also become more aware of how to overcome her weaknesses. I think she can accomplish major improvements. *(C3)*

A clear strength of the pair D assessment process is that their judgements with regard to the components of providing clear structure and guidance for students, setting norms and standards for students, and correction of undesirable student behaviour are carefully supported with evidence: Both confirmatory and disconfirmatory observations are discussed along with construct relevance, the quality of the available data and the relative importance of the available data. Given the focus of the assessors on reaching a consensus with regard to primarily the aforementioned components, however, the component of attention to students went underrepresented in their discussion. More specifically, the assessors showed consensus with regard to this component of the teacher's interpersonal competence during an early stage of their discussion and therefore failed to check the plausibility of their judgements, which remained quite abstract with no reference to concrete evidence as a result.

3.4.2.3 Collaborative Assessment Type III: Addition

The judgement processes for 5 of the 12 collaborative pairs were found to be mainly based on addition (i.e., the provision of additional evidence for a particular interpretation or judgement). As can be seen from Table 3.4, another communicative activity which was found to occur on a greater than average basis for pairs E through I was "check if all the important evidence has been taken into account." In fact, this activity for pairs E through I accounts for 7 of the total of 8 occurrences of the activity (see Table 3.1). And checking that all evidence has been taken into account constitutes one of the key processes of valid assessment.

Table 3.4 Communicative activities found to be above or below average for Type III Collaborative Assessment (addition)

Activities		**Mean**	
		Type III	**Overall**
Activities which occur **above** average:			
Contribution	Check whether all important evidence has been taken into account	1.4	0.8
Acceptance	Provide additional evidence	13.8	10.3
Activities which occur **below** average:			
Discussion	Challenge with counterevidence or alternative interpretation	3.2	6.0

The judgement processes of the five Type III pairs also had in common that the discussions of contributions occurred on a *below* average basis. Closer examination of these discussions showed them to generally involve just a single discussion of contribution. The discussions also concerned only minor details in most cases. Four of the 10 relevant discussions resulted in an explicit consensus; two remained undecided; and four were subsequently ignored.

An assessment approach in which the assessors are primarily aimed at the addition of information typically results in a transparent trail of evidence leading to a particular judgement or interpretation. While the activity of "check if all the important evidence has been taken into account" can be seen to occur more than average for the five Type III pairs, the assessors seek primarily confirmatory evidence and virtually no counterevidence is provided. The discussions of the Type III pairs are also of a very concrete character. In Box 3.3, a fragment from a Type III collaborative pair is presented.

Box 3.3 Illustration of Type III Collaborative Assessment: Addition (pair F)

Assessor 1	I think the teacher's analysis of her own behaviour and the classroom situation is not very strong. She never finishes her analyses properly. *(A1)*
Assessor 2	No, her analysis is not that convincing. Also, she constantly contradicts herself. *(B2)*
Assessor 1	Can we provide an example? Um, (…) do you think line 83 is an example of such a contradiction? *(B2)*
Assessor 2	Let me see (…) I noted line 85 to 87 as an example… Yeah, okay, yes, that's the same example you mentioned. Um, another example can be found around line 125, when the teacher says that she thinks it's okay for students to have discussions but also that she expects students to pay attention to her lesson at the same time. *(B2)*
Assessor 1	What's with student teachers wanting to allow students to discuss during lectures?
Assessor 2	Discussion and paying attention to the lesson pose quite a paradox in my opinion.
Assessor 1	Yes, they do! And she really seems to think that students discuss *the subject matter*…
Assessor 2	So, our point is that the teacher does not finish her analyses and that she contradicts herself regularly. In addition, I think she is not able to formulate concrete plans for improvement. *(A3)*
Assessor 1	Yes, no concrete plans whatsoever. And this is perfectly in line with the fact that she does not know what to improve. I mean, when your analysis is inadequate… *(A6)*

3.4.2.4 Collaborative Assessment Type IV: Neither addition nor challenging

For 3 of the 12 collaborative pairs, the assessment process was characterized by neither addition nor challenging. As can be seen from Table 3.5, the communicative activity of "noting agreement between assessors" occurred on a greater than average basis for pairs J, K and L. The fact that the assessors noted their agreement rather early in the interaction might explain the fact that the provision of counterevidence or alternative interpretations (i.e., challenging) was found to occur far below average.

Table 3.5 Communicative activities found to be above or below average for Type IV Collaborative Assessment: (neither addition nor challenging)

Activities		**Mean**	
		Type IV	**Overall**
Activities which occur **above** average:			
Contribution	Note agreement between assessors	3.3	1.5
Activities which occur **below** average:			
Contribution	Suggest an approach to the judgement task	0.3	2.6
Acceptance	Provide additional evidence	4.3	10.3
Discussion	Challenge with counterevidence or alternative interpretation	1.7	6.0

Similar to the Type III pairs, discussions of contributions were found to occur below average for the Type IV pairs. The nature of the Type IV discussions was also very similar to the nature of the Type III discussions which generally involved just a single discussion of contribution. Five of the 11 relevant discussions resulted in a consensus; three remained undecided; and three were ignored. A fragment from a typical Type IV interaction is presented in Box 3.4.

No specific validity strengths can be noted for Type IV collaborative assessment. A lack of argumentation or failure to provide supportive evidence was found to occur for all of the collaborative pairs but stood out in particular for the Type IV pairs, however. For the three Type IV pairs, the assessment process was either too abstract with little or no reference to concrete evidence or too concrete with little or no abstraction.

Box 3.4 Illustration of Type IV Collaborative Assessment: Neither addition nor challenge (pair K)

Assessor 1	What I also found essential, the QTI results show that her self-perspective [on the teacher-student relationship] differs considerably from the student perspectives. She actually has a very poor self-awareness. *(A2)*
Assessor 2	A poor self-awareness? No, I don't really agree with you. *(C1d)*
Assessor 1	She thinks she's not strict enough and comes across very insecure, but from the student perceptions can be concluded that they really do not notice this. So, she could really use coaching on this specific aspect. *(A1, A2)*
Assessor 2	Yes, and also on some other aspects. *(B3)*
Assessor 1	Okay, but we were talking about her awareness of her own behaviour.
Assessor 2	I also feel she could use coaching with respect to the accurate use of her voice, for example. And with respect to how do you call it, using body language, um, and presentation style? She has an interesting subject for the students, namely AIDS in Africa, um, condom use, and still the students are not really involved. You and I would consider it a piece of cake to get the students interested in this subject. So, I see a few clear directions for coaching. *(A2 ,D)*
Assessor 1	Okay…*(B3)*

3.5 Conclusions and discussion

In the present study, an overview of the communicative activities undertaken when assessors engage in the collaborative assessment of teacher competence was presented. Theoretically, we argued that a process aimed at both the addition of evidence and the challenging of interpretations should produce the most valid judgements. The present results show such a process to be very difficult to realize in actual practice and thus confirm the results of Moss, Schutz, and Collins (1998) and Quinlan (2002). Assessors nevertheless undertake activities which clearly reflect the essential assessment processes of seeking coherence among available sources of

evidence, checking whether all of the relevant evidence has been taken into account and challenging developing interpretations with counterevidence and/or perspectives of the other. Only the discussions by 3 of the 12 collaborative pairs were found to involve both addition and challenging or a Type I pattern of collaborative assessment. The discussion of a different pair was characterized by a predominantly challenging or Type II pattern of collaborative assessment. The discussions by another five pairs were mainly characterized by addition or a Type III pattern of collaborative assessment. And the discussions by the three remaining pairs were characterized by neither addition nor challenging and thus a Type IV pattern of collaborative collaboration.

The twelve collaborative pairs clearly differed with respect to the communicative activities which occurred and, more specifically, the occurrence of contributions, the acceptance of contributions, the discussion of contributions and the ignoring of a discussion. For each type of collaboration, we determined which communicative activities occurred on a greater than average basis and found those pairs representing the Type I pattern of collaborative assessment to display the greatest variability. In addition to the activities of "providing additional evidence" and "challenging via the provision of counterevidence or alternative interpretations," these pairs undertook many other communicative activities.

When the exact nature of the discussions of the contributions made by the collaborative pairs is examined in particular, considerable variability across the pairs can again be seen to occur. A rather remarkable finding is that the relevant discussions differed with respect to not only the variability of the discussion activities but also the extent to which the discussion activities resulted in a clear conclusion (i.e., a consensus). Whereas the Type I and Type II interactions involved a series of discussion activities and thus entailed a real discussion, the Type III and Type IV interactions frequently involved a single discussion activity which was often concerned with only minor details. The Type I and II discussions were also more likely to produce explicit conclusions than the Type III and IV discussions although some of the discussions for all of the pairs remained largely undecided or more or less ignored right from the start. These differences in the occurrence and exact nature of the discussion activities undertaken across collaborative pairs may be explained in part by the level of generality characterizing the discussions. The judgement produced by the collaborative pair should have a higher level of generality than the relevant particulars (e.g., Delandshere & Petrosky, 1994; Mislevy et al., 2002; Moss et al., 1998). Our data suggest that overly holistic judgements with only minor referral to specific data (as in Type IV assessment processes) and overall judge-

ments which do not go beyond the level of specifics (as in Type III assessment processes) do not trigger serious discussion, the mention of counterevidence or the consideration of alternative interpretations.

Although Type I judgement processes involving both the addition of evidence and the challenging of interpretations can be assumed to produce the most valid judgements, the quality of the collaborative assessment process for the three pairs of Type I assessors can nevertheless be improved. The discussion activities did not always produce clear conclusions and some of the discussion activities were even ignored at times, which means that the counterevidence presented and alternative interpretations put forth may not always be reflected in the final judgements of these pairs. In other words, even Type I collaborative pairs showed a disposition to consider only confirmatory evidence as opposed to counterevidence or alternative interpretations.

It should be noted that all of the assessors in this study were using the procedure to assess the interpersonal competence of student teachers for the first time. It is often asserted that the expertise of assessors stabilizes as they gain experience with a specific assessment procedure (e.g., Uhlenbeck et al., 2002). Also, the supervisory backgrounds of the assessors in the present study may have predisposed them to focus on the weaknesses and not the strengths of the student teachers. It may be that a predisposition to provide tips and tricks coloured the judgements of these assessors and that more balanced judgements may emerge as the assessors grow more accustomed to their role as assessor as opposed to tutor. All evidence, including both confirmatory *and* disconfirmatory information, will presumably then be taken into account.

With regard to the potential advantages of collaborative assessment over individual assessment, it is difficult to draw conclusions on the basis of the present results. It was assumed that the two key assessment processes of "seeking coherence among the available sources of evidence and checking whether all relevant evidence has been taken into account," on the one hand, and "challenging developing interpretations with counterevidence and entertaining alternative interpretations" on the other hand, would be stimulated by collaborative assessment. However, only 3 of the 12 assessor pairs displayed a judgement process which jointly reflected these key processes. While this suggests that collaborative assessment does not specifically elicit the desired judgement processes, we still believe that taking all of the relevant evidence — both confirmatory and disconfirmatory — into account may be easier in collaboration than individually on one's own. A suggestion for future research is therefore to

explore the specific conditions needed to enable assessor pairs to engage in a more balanced assessment process. Similarly, those collaborative judgement processes involving both addition and challenge should be analyzed in greater detail to gain greater insight into just how the collaborative process can increase the validity of the assessment process. Think-aloud sessions or interviews with expert assessors can also provide insight into the necessary conditions. In future research, moreover, not only the enabling conditions for valid collaborative assessment but also the disabling conditions (i.e., disadvantages of collaboration) should be taken into consideration.

3.6 Implications for the training of assessors

The present overview of the communicative activities undertaken during the collaborative assessment of teacher competence can certainly be introduced into teacher assessment preparation programmes. The four types of collaborative assessment together with their relative strengths and weaknesses should certainly be considered within such a programme. In doing this, the discussion of such validity-related issues as construct irrelevant variance and construct under-representation can be made more tangible.

Although the present results might also be called upon to formulate guidelines for collaborative assessment, the results of research by Moss et al. (1998) suggest that the provision of highly specific guidelines should be avoided. A tendency to draw too heavily upon such guidelines may simply preclude in-depth consideration of all possible evidence. Rather than explicit guidelines and monitoring of the assessment process, thus, assessors may be asked to discuss their judgements following each assessment or a series of assessments. How do they perceive the quality of the judgement process? Did they consider both confirmatory and disconfirmatory evidence? How could the assessment process be improved? The four types of collaborative assessment highlighted in the present study can provide a helpful basis for the analysis of assessment activities and the formulation of specific suggestions and directions to improve the quality of assessment processes.

References

Barron, B. (2003). When smart groups fail. *The Journal of the Learning Sciences*, 12(3), 307-359.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences, 2nd edition.* New Jersey: Lawrence Erlbaum Associates.

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

Delandshere, G., & Arens, S.A. (2003). Examining the quality of the evidence in preservice teacher portfolios. *Journal of Teacher Education*, *54*, 57-73.

Delandshere, G., & Petrosky, A. R. (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.

Delandshere, G., & Petrosky, A. R. (1998). Assessment of complex performances: Limitations of key measurement assumptions. *Educational Researcher*, *27*, 14-24.

Dwyer, C.A. (1995). Criteria for performance-based teacher assessments: Validity, standards and issues. In A.J. Shinkfield, & D. Stufflebeam (Eds.), *Teacher evaluation: Guide to effective practice* (pp.62-80). Boston: Kluwer Academic Publishers.

Dwyer, C. A., & Stufflebeam, D. (1996). Teacher evaluation. In D. C. Berliner & R.C. Calfee (Eds.), *Handbook of Educational Psychology* (pp. 765-768). New York: Macmillan.

Evans, E.D., & Tribble, M. (1986). Perceived teaching problems, self-efficacy, and commitment of teaching among preservice teachers. *Journal of Educational Research, 80*, 81-85.

Evertson, C.M., & Weinstein, C.S. (2006). Classroom management as a field of inquiry. In C.M. Evertson & C.S. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 3-16). Mahawn: Lawrence Erlbaum Associates.

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

Heller, J. I., Sheingold, K., & Myford, C.M. (1998). Reasoning about evidence in portfolios: Cognitive foundations for valid and reliable assessment. *Educational Assessment*, 5(1), 5-40.

Johnston, B. (2004). Summative assessment of portfolios: An examination of different approaches to agreement over outcomes. *Studies in Higher Education*, *29*, 395-412.

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

Mabry, L. (1999). *Portfolio's plus, a critical guide to alternative assessment*. Thousand Oaks: Corwin press.

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

Moss, P. A., Schutz, A. M., & Collins, K. M. (1998). An integrative approach to portfolio evaluation for teacher licensure. *Journal of Personnel Evaluation in Education*, *12*, 139-161.

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., & Johnson, L.(2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15*, 363-389.

Nijveldt, M., Beijaard, D., Brekelmans, M., Verloop, N., & Wubbels, Th. (2005). Assessing the interpersonal competence of beginning teachers: The quality of the judgement process. *International Journal of Educational Research*, *43*, 89-102.

Quinlan, K. M. (2002). Inside the peer review process: How academics review a colleague's teaching portfolio. *Teaching and Teacher Education*, *18*, 1035-1049.

Schutz, A., & Moss, P.A. (2004). Reasonable decisions in portfolio assessment: Evaluating complex evidence of teaching. *Education Policy Analysis Archives*, 12(33). Retrieved 7/19/2004 from http://epaa.asu.edu/v12n33/.

Tigelaar, D. E. H., Dolmans, D.H.J.M, Wolfhagen, I.H.A.P., & Van der Vleuten, C.P.M. (2005). Quality issues in judging portfolios: Implications for organizing teaching portfolio assessment procedures. *Studies in Higher Education, 30*, 595-610.

Uhlenbeck, A.M., Verloop, N., & Beijaard, D. (2002) Requirements for an assessment procedure for beginning teachers: Implications from recent theories on teaching and assessment. *Teachers College Record, 104*, 242-272.

Veenman, S.A.M. (1984). Perceived problems of beginning teachers. *Review of Educational Research, 54,* 143-178.

Wubbels, Th., Brekelmans, M., Den Brok, P., & Van Tartwijk, J. (2006). An interpersonal perspective on classroom management in secondary classrooms in the Netherlands. In. C. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: research, practice and contemporary issues* (pp.1161-1191). New York: Lawrence Erlbaum Associates.

4

Assessors' perceptions of their judgement processes[3]

Abstract

As current procedures for teacher assessment are often based on non-standardized information derived from multiple sources, the overall validity of the assessment depends heavily on the judgement processes of the assessors. Because it is of great importance for assessors to be aware of their own judgement processes and of the possible threats to validity in these processes, investigating assessors' perceptions is of vital significance. In the present study, the perceptions of 22 assessors who judged a student teacher pair-wise using a specific assessment procedure were explored using semi-structured interviews. A qualitative analysis of the individual assessors' perceptions with regard to the essential judgement processes of consideration of evidence and combination of evidence to attain an overall judgement resulted in an overview of successful strategies and threats underlying a valid assessment process. General implications for ensuring the validity of the assessment process and the preparation of assessors are discussed.

[3] This chapter has been accepted in adapted form for publication in Studies in Educational Evaluation:
Nijveldt, M., Beijaard, D., Brekelmans, M., Wubbels, Th., & Verloop, N. Assessors' perceptions of their judgement processes: Successful strategies and threats underlying valid teacher assessment.

4.1 Introduction

Authentic assessment of teacher competence, including assessment based on teaching portfolios, has become increasingly common practice in teacher education. Such assessment is used for both formative purposes — to stimulate reflection and professional development of student teachers — and summative purposes — to inform decisions about certification of student teachers. In order to do justice to the complex nature of teaching, authentic assessment is generally based on non-standardized, qualitative material derived from sometimes widely differing sources (e.g., a videotaped lesson and written reflections on performance), often originating from a variety of contexts. Because of its authentic nature, this form of assessment is generally perceived as highly valid. However, whereas a considerable number of researchers have addressed the reliability of authentic assessment, few have specifically addressed its validity. Given that assessments are based mainly on qualitative, non-standardized evidence, validity rests largely on the professional judgement of assessors (Dwyer, 1995). That is, even though the selected tasks and instruments and the assessment criteria employed are important conditions for a valid judgement, the quality of the assessment as a whole depends primarily on the ability of assessors to use these instruments and criteria to reach a valid overall judgement. While current teacher assessment is generally based on an assessment framework which explicitly underlies the content criteria (Wolf, 1995) and helps assessors determine what to look for in a performance and what to exclude from the overall judgement, even trained assessors may introduce potential sources of invalidity into the assessment process. To date, little is known about the exact nature of the assessment process and the exact threats to validity within this process. Therefore, it remains unclear just how assessors realize a valid judgement of teacher competence in practice. In the present study, we aimed to gain greater insight into the 'strategies' and 'threats' underlying valid assessment on the basis of the perceptions of the assessors themselves. We analyzed the perceptions of assessors who judged a student teacher pair-wise using a specific assessment procedure. Based on the individual assessors' perceptions of the relevant judgement processes and perceived difficulties, an overview is provided of successful strategies and threats pertaining to the validity of the assessment process. In closing, based on this information, general implications for ensuring the validity of assessment processes and the preparation of assessors are discussed.

4.2 Validity of the assessment process

As assessment of teacher competence is generally based on several sources and types of data, the assessment process is often thought of in terms of two essential assessor tasks or processes: (a) consideration of evidence from separate sources and (b) combination of evidence from different sources to attain an overall judgement (Delandshere & Petrosky, 1998; Delandshere & Arens, 2003; Mehrens, 1990; Moss, Schutz, & Collins, 1998; Uhlenbeck, Verloop, & Beijaard, 2002; Wolf, 1995). In practice, the assessment process normally has an iterative character (Heller, Sheingold, & Myford, 1998): assessors do not perform these processes as two subsequent steps, but typically alternate between the two processes during the assessment process. As these tasks require considerable judgement on the assessor's part, they are difficult to carry out in a sound way. Assessors may introduce two major threats to the validity of these processes: construct-irrelevant variance and construct under-representation (Messick, 1989, 1996). *Construct-irrelevant variance* may arise when assessors apply irrelevant or idiosyncratic criteria and when assessors consider evidence that is irrelevant to the construct in question (i.e., the competence to be assessed). In terms of the two essential judgement processes, construct-irrelevant variance may typically arise in the process of consideration of evidence from separate sources. *Construct under-representation* may arise in the processes of both consideration and combination of evidence. In the process of consideration of evidence, it is characterized by the omission of particular assessment criteria and/or relevant evidence, and in the process of combination of evidence from separate sources it is characterized by idiosyncratic weighing of certain criteria and/or evidence, which may result in insufficient attention to particular aspects of performance (Heller et al., 1998). In order to prevent idio-syncratic weighing, it is important that assessors not only consider evidence which confirms their initial interpretations, but also actively seek 'counterevidence' which may challenge these interpretations. Research has shown that assessors generally show a strong tendency to search for approval of their initial interpretations, and even evidence which disconfirms this initial judgement is read to some extent as supporting it (Moss et al., 1998; Schutz & Moss, 2004). In connection with this, it is important that assessors warrant their judgements by plausible argumentation. That is, when assessors carefully check whether their interpretations and conclusions are sufficiently supported by the original data (Kane, 1992; Moss, 1994), possible threats to the validity of the assessment process are minimized. In Table 4.1, the above-mentioned 'indicators of validity' are summarized in connection with the essential judgement processes of 'consideration of

evidence from separate sources' and 'combination of evidence from different sources to attain an overall judgement' and the validity criteria of no construct-irrelevant variance and no construct under-representation. It is still unclear just how assessors comply with these indicators in practice. Particularly the process of combination of evidence from different sources and contexts into an overall judgement remains obscure, and in connection with this, current validity theory does not offer clear directions (Dwyer, 1995; Moss, Girard, & Haniford, 2006). The aim of the present study was to provide an overview of successful strategies and threats underlying an assessment process which complies with the aforementioned indicators. Such an overview contributes to the knowledge about ensuring the validity of assessment processes. In sum, the question that prompted our study was the following: 'Which strategies to ensure the validity of the assessment process are performed by assessors in connection with the essential judgement processes of (a) consideration of evidence and (b) combination of evidence to attain an overall judgement?'

Table 4.1 Indicators of validity for the essential judgement processes.

Judgement process	**Validity criteria**	**Indicators**
Consideration of evidence from separate sources	No construct-irrelevant variance	(a) Evidence refers only to the relevant competences: no extraneous, irrelevant criteria are added;
	No construct under-representation	(b) All of the relevant competences and components of these are covered; all relevant evidence is considered.
Combination of evidence from different sources to attain an overall judgement	No construct under-representation	(a) All of the evidence noted for the separate sources is also clearly reflected in the overall judgement; (b) Both confirmative evidence and counterevidence are considered in the overall judgement; (c) No idiosyncratic weighing of evidence; (d) All conclusions are supported by clear argumentation which also draws upon the original data.

4.3 Method

4.3.1 Assessment procedure

We analyzed the perceptions of assessors who evaluated the interpersonal competence of a student teacher, using a specific assessment procedure. Interpersonal teacher competence, involving the quality of interpersonal relationships with students, is a major concern for many student teachers and beginning teachers (e.g., Evans & Tribble, 1986; Evertson & Weinstein, 2006; Veenman, 1984). In this section, we elaborate on the characteristics of the assessment procedure used, which are considered preconditions for a valid assessment process.

First, to help assessors obtain a comprehensive picture of the candidate's interpersonal competence, an *assessment framework* was provided distinguishing six key components of interpersonal competence: (1) providing clear structure and guidance for students, (2) setting norms and standards for students, (3) correcting undesirable student behaviour, (4) attention to students, (5) provision of freedom and responsibility for students, and (6) reflection upon one's interpersonal performance. The specific assessment framework was based on the scientific literature on interpersonal relationships between teachers and students; more specifically, on the Model for Interpersonal Teacher Behaviour (MITB)(see Wubbels, Brekelmans, Den Brok, & Van Tartwijk, 2006). See Chapter 2 for greater detail on the alignment between the MITB and the assessment framework. For each of the six components of the assessment framework, both examples and counterexamples of interpersonal competence were provided for illustrative purposes. It was assumed that the provision of counterexamples in addition to confirmatory examples would help the assessors attain a com-prehensive picture of the candidates' interpersonal competence. In addition, for each of the six components, one or two *guiding questions* were formulated to help focus assessors' analysis of student teacher's interpersonal competence. For example, the following questions were formulated for the component of correcting undesirable student behaviour: How does the teacher respond to undesirable student behaviour? How does the teacher check the effects of his or her interventions? The guiding questions for the reflection component of interpersonal competence were as follows: Does the teacher show an awareness of personal strengths and weaknesses? Is the teacher able to formulate an adequate analysis of his or her behaviour?

Second, in order to obtain a comprehensive picture of the student teacher's interpersonal competence, *multiple sources of data* were to be consulted. The procedure used was

based on three data-sources: (1) the Questionnaire on Teacher Interaction (QTI), which provides a reliable and valid picture of both the student and teacher perspectives on the interpersonal relationships between the teacher and his or her students (see Wubbels et al., 2006), (2) a video-taped lesson selected by the student teacher, and (3) a self-evaluation assignment by the student teacher, in which he/she was asked to analyze both the results of the QTI and the videotaped lesson and to provide (a) an overview of strengths and weaknesses with as much referral to the QTI results and the videotaped lesson as possible and (b) a personal professional development plan.

Third, for each assessment instrument, a specific *format* was provided. That is, to ensure that all six key components of interpersonal competence were systematically considered in both of the essential judgement processes, a form was provided for each of the three instruments, containing a column to note evidence for each of the six components.

Fourth, a typical characteristic of the present assessment procedure is *discussion with another assessor* after assessors have formulated an initial, individual judgement. Following the procedure, the assessors firstly noted evidence from three data-sources, individually. They then integrated this evidence into an individual judgement in the form of an evaluative summary. Finally, they entered into discussion with another assessor, in which a collaborative, definitive summary was to be formulated. Such discussion is considered to encourage assessors to comply with the above-mentioned indicators of validity (Moss et al., 1998; Tigelaar, Dolmans, Wolfhagen, & Van der Vleuten, 2005).

Fifth, a *training programme for assessors* was also a part of the procedure. Extensive preparation of assessors is widely acknowledged to be an important condition for a valid assessment process. Assessor training generally includes extensive discussion of the assessment framework, helping assessors to recognize evidence and distinguish between different performance levels (Dwyer, 1995).

4.3.2 Assessors and assessor preparation

Twenty-two individuals who were at the time teacher educators (N=4) or experienced teachers supervising the training of prospective secondary school teachers (N=18) were selected to participate in the present study. All of the individuals had experience in the evaluation of student teachers and /or in providing feedback on the performance of student teachers. Five of the individuals also had specific experience in the assessment of teacher competence. Prior to their participation in the present study, all of the aforementioned individuals followed a two-

evening training programme. During the first evening, the above-mentioned procedure to assess the interpersonal competence of student teachers was introduced, and the assessment framework was discussed. During the second evening, the assessors conducted an entire assessment using an authentic case (i.e., data provided by a student teacher). Following completion of the assessment task, the experiences of the assessors were exchanged in a plenary discussion in which specific attention was paid to the evidence noted, the interpretation of this evidence, and the difficulties encountered. Note that the current preparation programme did not provide assessors with specific guidelines for the process of combining evidence from the three sources to attain an overall judgement. As the literature does not provide straightforward guidelines for the combination of situated, non-standardized evidence, we were particularly interested in the strategies assessors themselves would consider. Therefore, we presented the assessors with only the validity indicators pertaining to the combination of evidence, as presented in Table 4.1. That is, we stressed the importance of (a) clearly reflecting all of the evidence noted for the separate sources in the overall judgement, (b) actively searching for counterevidence which might challenge initial interpretations, (c) a balanced weighing of the available evidence, and (d) checking whether all conclusions are supported by clear argumentation which also draws upon the original data.

4.3.3 Data collection and analysis

Following participation in the training programme, the 22 assessors were invited to our institute in 3 groups and asked to assess the interpersonal competence of another student teacher. All assessors evaluated the same student teacher. In keeping with the assessment procedure, the assessors individually noted evidence from the three different sources of data and prepared an individual, initial, evaluative summary. The assessors then entered into discussion with another assessor to formulate a definitive summary of the teacher's interpersonal competence and translate this summary into a numerical rating. The data relevant to the present study consisted of retrospective, semi-structured interviews with individual assessors, conducted shortly after the assessors finished their assessment process. A total of 22 interviews were held, with an average duration of approximately 40 minutes (ranging from 20 minutes to 60 minutes). Each interview was audio-taped.

The interview scheme consisted of two parts: (I) questions focusing on the two essential judgement processes of consideration of evidence from separate sources, and combination of evidence from different sources to attain an overall judgement, and (II) more specific

questions focussing on the indicators of validity distinguished. Examples of interview questions are presented in Table 4.2.

Table 4.2 Examples of interview questions

Interview part	**Examples**
I Essential judgement processes	Did you feel capable of noting relevant evidence for the six key components of interpersonal competence? Why? Did you perceive any difficulties in this process? What approach did you take with respect to the combination of evidence from different sources to attain an overall judgement, (a) individually and (b) in collaboration with the other assessor? Did you perceive any difficulties in this process?
II Indicators of validity	During the assessment process, were you engaged in the following activities? If so, in what way? o Seeking coherence between the three data sources o Considering the specific weight of the relevant data o Checking whether all relevant evidence is clearly reflected in the overall judgement Did you reconsider any of your interpretations of the data or any judgements based on alternative interpretations or judgements of the other assessor? If so, in what way, or which specific interpretations or judgements?

The analysis of the interview data was conducted as follows. For each of the 22 assessors, a report of perceptions of the assessment process was created for the following two essential judgement processes: (1) consideration of evidence from separate sources and (2) combination of evidence from different sources to attain an overall judgement. For each judgement process, perceptions of the assessors were categorized in terms of the validity criteria of (a) no construct-irrelevant variance and (b) no construct under-representation and their indicators (as outlined in Table 4.1). Moreover, a distinction was made between strategies which assessors perceived as likely to result in a valid judgement and specific problems which in the opinion of the assessors might hamper the realization of a valid judgement. An overview of 'successful strategies' and 'problems' was then generated for each of the validity indicators, as outlined in Table 4.1, by grouping similar strategies and problems of individual assessors.

4.4 Successful strategies for the consideration of evidence from separate sources

An overview is provided in Table 4.3 of the successful strategies employed and problems encountered in connection with the process of consideration of evidence from separate sour-

ces. As can be seen, the assessors mentioned a great variety of strategies for the consideration of evidence. In comparison to the number of successful strategies, relatively few problems (i.e., threats to validity) were mentioned. In this section, the results presented in the table are exemplified with relevant quotes from the interviews.

Table 4.3 Overview of perceived successful strategies and problems concerning the consideration of evidence from separate sources

Indicators of validity	**Perceived successful strategies**	**Perceived problems**
(a) Evidence refers to only the relevant competence *(N=18)*	1. 'Correct' use of the assessment framework *(N=18)* 2. Correction for irrelevant evidence by discussing evidence with the other assessor *(N=2)*	1. A predisposition to also consider evidence for *didactical* competence, which is irrelevant to the assessment of *interpersonal* competence *(N=2)*
(b) All relevant evidence is considered *(N=19)*	1. Checking collaboratively all evidence noted, resulting in the addition of evidence *(N=6)* 2. Checking collaboratively all evidence noted and concluding that both assessors noted the same evidence *(N=6)* 3. Correction for a predisposition to note a disproportionate amount of evidence for one or more aspects of competence *(N=4)* 4. Correction for a predisposition to note relatively more negative evidence than positive evidence *(N=3)* 5. Correction for a predisposition to overlook relatively tangible aspects of competence *(N=2)* 6. Also considering relevant evidence for interpersonal competence which was not explicitly mentioned in the examples and counterexamples of interpersonal competence provided in the assessment framework *(N=2)* 7. Using the assessment framework to check whether aspects of interpersonal competence were overlooked *(N=2)*.	1. A predisposition to note a disproportionate amount of evidence for one or more aspects of interpersonal competence *(N=4)* 2. A predisposition to note relatively more negative evidence than positive evidence *(N=3)* 3. A predisposition to overlook relatively tangible aspects of interpersonal competence *(N=2)*

4.4.1 Evidence refers to only the relevant competence

All twenty-two assessors stated that the essential process of consideration of evidence from separate sources was well supported by the assessment framework underlying the assessment procedure, and the discussion of this framework in the training programme. In line with this,

eighteen of the assessors believed that as they were able to apply the assessment framework correctly, only evidence for the relevant competence — in this case *interpersonal* competence — was considered.

> "I felt well able to recognize evidence. I positively distinguished all the relevant information and *only* relevant information. The classification of the different components certainly proved helpful in this regard as it clearly focused and helped structure my observations" (A4).

Two other assessors believed that no irrelevant variance was considered, partly because 'questionable' data were extensively discussed with the other assessor until a consensus was reached on the relevance or irrelevance of the particular evidence. Another two assessors, however, questioned their ability to consider only the relevant evidence without adding any irrelevant criteria: they both perceived difficulties with distinguishing interpersonal competence from specifically *didactical* competence. Both assessors explained this difficulty as resulting from their background as didactical experts.

> "I do feel that my observations are quite coloured by a didactical lens. Being a didactical expert I just couldn't get rid of this lens" (A15).

4.4.2 All relevant evidence is considered

In describing the essential process of noting evidence from separate sources, several assessors claimed to have been explicitly engaged in the process of considering whether all relevant evidence had been considered *(N=19)*. These assessors took different approaches to this process. *First*, 6 assessors thought all relevant evidence had been considered as a result, in part, of their having discussed all the evidence noted with the other assessor, resulting in the addition of evidence which had been overlooked by the individual assessors, and, in this way, in the consideration of *all* relevant evidence.

> "We compared our evidence noted for the QTI, video-taped lesson, and the self-evaluation assignment — in this order — which led to the addition of evidence. We literally collected all facts available. [The other assessor] specifically considered different evidence from the self-evaluation than I did; vice versa, she said I noted evidence she had not considered." (A4).

Second, another 6 assessors, who also discussed all the evidence noted with the other assessor, concluded from this discussion that "exactly the same evidence" had been noted, which led them to believe that *all* relevant evidence had been considered. *Third*, 4 assessors found they had noted a disproportionate amount of evidence for particular aspects of competence. To correct for this, 2 of these 4 assessors checked whether more evidence was to be found for the aspects of competence which were underrepresented:

> "I detected that I had noted much more evidence for the first of the [six] components of interpersonal competence than for the others. This is certainly a result of my specific way of observing student teachers. This is certainly a result of a somewhat restricted view" (A10).

Fourth, three assessors found they had noted relatively more negative evidence than positive evidence. To correct for this, they checked whether more positive evidence was to be found.

> "I first made an overview of strengths and weaknesses [of the student teacher] accompanied with numerous specific examples. For all six components of interpersonal competence I explicitly tried to consider both weaknesses and strengths. For some components I had considerable difficulties in also noting essential strengths, but I really tried to also consider these" (A9).

Fifth, two assessors noted a tendency to overlook relatively intangible aspects of competence, particularly those pertaining to the aspect of setting norms and standards. To correct for this, they checked whether in addition to tangible evidence, more intangible evidence had also sufficiently been considered:

> "It struck me that the greater part of my notes pertained to the component of 'providing clear structure and guidance for students.' As a result, I added evidence like 'being patient with students' and other evidence I had not noted first, because it does not involve a specific action. Some components of interpersonal competence particularly comprise such 'invisible evidence', for instance, 'setting norms and standards for students'. You cannot actually observe [the student teacher] doing it, but it can be derived from the fact that [the student teacher] corrects students when they — apparently — break a rule" (A21).

Sixth, another two assessors stated that they had considered evidence for interpersonal competence which was not explicitly included in the examples and counterexamples of inter-

personal competence provided in the assessment framework, resulting in a comprehensive consideration of all relevant evidence for this particular student teacher.

Finally, another two assessors believed they had considered all relevant evidence because — after having noted evidence for the three data-sources — they used the assessment framework to check whether any aspects of interpersonal competence had been overlooked.

4.5 Successful strategies for the combination of evidence to attain an overall judgement

An overview is provided in Table 4.4 of the successful strategies employed and problems encountered in connection with the process of combination of evidence from separate sources to attain an overall judgement. As was the case for the results pertaining to the process of consideration of evidence, a great variety of strategies was mentioned for the combination of evidence and a relatively greater number of strategies than problems was mentioned. Below, we illustrate the results presented with relevant quotes from the interviews.

4.5.1 All evidence considered is clearly reflected in the overall judgement

In relation to the essential process of 'combining the evidence to attain an overall judgement', only 1 of the 22 assessors explicitly referred to the process of 'checking whether all relevant evidence is considered.' He explained that he explicitly checked, together with the other assessor, whether "all relevant evidence considered by him and his co-assessor was sufficiently reflected in the eventual overall judgement" (A3).

Table 4.4 Overview of perceived successful strategies and problems concerning the combination of evidence to attain an overall judgement

Indicators of validity	Perceived successful strategies	Perceived problems
(a) All relevant evidence is considered in the overall judgement *(N=1)*	1. Explicitly checking, together with the other assessor, whether all relevant evidence is reflected in the collaborative judgement *(N=1)*	
(b) Both confirmative evidence and counterevidence are considered *(N=7)*	1. Noting incoherence of some of the evidence noted and searching for a comprehensive judgement, including counterevidence *(N=3)* 2. Discussion of alternative interpretations or perspectives of the other assessor, resulting from clear differences of opinion *(N=2)* 3. Consideration of counterevidence from the other assessor, resulting in slight alteration of a specific judgement *(N=2)*	1. A predisposition to focus particularly on data which "make sense" *(N=1)*
(c) No idiosyncratic weighing of evidence *(N=16)*	1. Equally reflecting the three data sources in the overall judgement *(N=12)* 2. Explicitly explaining *why* specific aspects of interpersonal competence are more important to a teacher's overall performance than others *(N=3)* 3. Explicitly checking whether the relevant negative and positive evidence was equally reflected in the overall judgement (N=2)	1. A specific data source was given less weight, without clear argumentation *(N=4)* 2. A specific aspect of interpersonal competence was given more weight, without clear argumentation, i.e., arbitrarily *(N=2)* 3. A predisposition to reflect more negative than positive evidence in the overall judgement (N=3).
(d) All conclusions are supported by clear argumentation which also draws upon the original data. *(N=9)*	1. Use of the formats offered clearly stimulated assessors to explicitly support conclusions with concrete data *(N=2)* 2. Discussion with the other assessor stimulated the forming of real opinions, as opposed to simply summarizing the relevant evidence *(N=1)* 3. In the discussion with the other assessor, all facts were brought together *(N=1)*	1. A lack of data to support a specific interpretation or judgement *(N=4)* 2. A general feeling that interpretations of the self-evaluation assignment were too subjective, insufficiently based on concrete evidence *(N=4)*

4.5.2 Both confirmative evidence and counterevidence are considered

Seven of the 22 assessors stated that they explicitly aimed at checking whether not only confirmative evidence (i.e., evidence which confirms initial judgements), but also counterevidence

had been considered (i.e., evidence which challenges initial judgements). Three of these 7 assessors aimed at taking counterevidence into account, both in their individual and collaborative judgement processes. These assessors detected incoherence in some of the evidence they noted, and sought to make a comprehensive judgement, including counter-evidence.

> "I noticed a considerable number of discrepancies, or inconsistency between the different sources of data, so I wrote these down for a start. (…) Based only on the results of the QTI and the self-evaluation, without the video-taped lesson, my overall judgement would have been less positive than it is now. Retrospectively, I see that I took the discrepancies between these sources as a starting point for the formulation of my overall judgement. Basically, these discrepancies formed the basis for my evaluative summary" (A10).

The other 4 of these 7 assessors mainly considered counterevidence in their collaborative processes, through discussion of alternative interpretations or perspectives of the other assessor. Whereas for two of these assessors this counterevidence was exchanged as a result of a clear difference of opinion (see the example below), the other assessors stated that the other assessor mentioned evidence which they had not considered themselves, resulting in a minor adjustment of a specific judgement.

> "After I had explained my overall judgement [to the other assessor], he made clear why he did not agree with some aspects of my argumentation. He did not agree with two or three vital aspects of my overall judgement, and we extensively discussed these. During our discussion I sometimes approved of his argumentation and he sometimes approved of mine, as we presented information the other had not considered, or interpreted differently. Thus, we really tried to convince each other" (A1).

Another assessor said that although he did not note any clear counterevidence or contradictions in the evidence, he did, however, note the tendency to "focus specifically on information that makes sense, as you agree more with some data than with other data. As a result, the overall judgement is without a doubt not very objective" (A18). This tendency reflects the aforementioned threat to validity of primarily searching for evidence which confirms an initial judgement as opposed to also actively seek for evidence which may challenge this initial judgement.

4.5.3 No idiosyncratic weighing of evidence

Sixteen of the 22 assessors mentioned strategies or threats with regard to the process of weighing the relative importance of evidence. These assessors mentioned the weighing of the relative importance of *data sources*; the weighing of the relative importance of *aspects of interpersonal competence,* and the weighing of *specific evidence*. With respect to consideration of the relative importance of the *data sources,* 12 assessors stated that, as a result of clear coherence between the different sources of data, the three data sources were equally reflected in the final judgement, indicating a balanced weighing of the available evidence. Four other assessors stated that one or more data sources were given greater or less emphasis in their final judgement, and explicitly questioned this approach.

> "The QTI caused us considerable trouble in this particular case, because it did not provide the results we expected from the video-taped lesson. We discussed whether we should simply put aside these [QTI] results or not. I found it very difficult to account for ignorance of these results. [The incongruence] particularly pertained to the nature of the relationships with the students. The students were positive, indicating a good rapport between teacher and students, but I did not sense this in the video-taped lesson. Was it this particular lesson? Eventually, we put less emphasis on the QTI results, without sound argumentation. Which of course is not particularly strong" (A5).

With respect to the weighing of the relative importance of *aspects* of interpersonal competence, three assessors stated — in their evaluative summaries — that they explicitly explained *why* specific aspects of interpersonal competence were more important to the teacher's overall performance than others. Two other assessors, who weighed the relative importance of different aspects of interpersonal competence without clear argumentation, explicitly questioned this approach (i.e., recognized the threat of idiosyncratic weighing). The first of these assessors said he indicated in his evaluative summary which specific strengths and weaknesses of the student teacher weighed particularly heavy in his overall judgement: "this of course is rather subjective as another person could simply think differently about this" (A20). The second assessor stated that she was concerned whether she — as a result of her supervisory background — might have put too much emphasis on the component of reflection upon one's interpersonal competence or the *potential* competence of the teacher as opposed to *current* competence.

Two assessors mentioned the weighing of *specific evidence*. Both of these assessors explicitly checked whether the relevant negative and positive evidence was equally reflected in their overall judgements, as they recognized a tendency to focus more on the weaknesses than the strengths of the student teacher.

4.5.4 All conclusions are supported by clear argumentation which draws upon the original data

While describing the essential process of combining data into an overall judgement, 9 of the 22 assessors mentioned the support of all conclusions by clear argumentation which also draws upon the original data. Only two of these nine assessors explicitly stated that they managed to realize clear argumentation, for several reasons. First, both of these assessors stated that the assessment framework, and particularly the accompanying 'formats', clearly stimulated them to carefully support conclusions with concrete data: "as the formats contained all the evidence I noted, I could literally use my notes as evidence" (A2). Second, one of these assessors claimed in addition that the discussion with the other assessor stimulated "the forming of a real opinion, as opposed to simply summarizing the relevant evidence" (A4). Third, the same assessor stated also that in the discussion with the other assessor, "all facts were brought together, resulting in a strong argumentation, very well supported by the facts" (A4).

The other seven assessors noted several difficulties relating to the process of supporting all conclusions by clear argumentation which also draws upon the original data. Firstly, the assessors found that not all of their remarks (i.e., conclusions) could be sufficiently supported with evidence (N=4). Whereas three of these assessors felt information was lacking primarily in the self-evaluation assignment, when "the candidate did not provide a comprehensive story or explanation", another assessor missed information in the video-taped lesson, because at some critical moments, the video-tape did not provide an overview of the whole classroom. All four of these assessors expressed a need for clarification from the candidate. Secondly, the assessors experienced difficulties in distinguishing actual data or evidence and interpreting these data or evidence (N=4), particularly in connection with the self-evaluation assignment. All these assessors found it difficult to note concrete evidence underlying their judgements and feared that their interpretations of the material were too subjective.

> "When interpreting the results of the self-evaluation assignment, to a certain degree I fell into the pitfall of 'interpreting away'. Although most information could be interpreted clearly and

> unequivocally, at times I did not feel convinced about my interpretations. In such cases I felt the need to discuss these interpretations with the student teacher" (A9).

4.6 Conclusion and discussion

In the present study, assessors' perceptions of essential judgement processes were investigated so as to provide an overview of successful strategies and threats underlying a valid assessment process. The present overview, based on the perceptions of 22 assessors, shows the assessors to have employed a considerable diversity of *strategies* (see Tables 4.3 and 4.4). However, these different strategies were generally mentioned by only small numbers of assessors. An interesting finding is that more strategies were mentioned for the process of 'consideration of evidence from separate sources' than for the process of 'combination of evidence to attain an overall judgement', which might suggest that the consideration of evidence is more tangible for assessors than the combination of evidence into an overall judgement. This aligns with the findings in the literature, where the process of consideration or 'recognition' of relevant evidence is more crystallized than that of combination of evidence (Dwyer, 1995; Moss et al., 2006), as current validity theory does not offer clear guidelines for the combination of evidence from different sources and contexts.

The strategies pertaining to the process of *consideration of evidence from separate sources* show that a great many assessors referred to the use of the assessment framework and accompanying formats as an indicator of a valid assessment process, as they helped assessors to include only the relevant evidence in the assessment and all the relevant evidence. Particularly in connection with the inclusion of *all* relevant evidence, the assessors in the present study employed a great variety of additional strategies, including correction for predispositions which may constitute bias, and discussion with the other assessor.

In the process of consideration of evidence from separate sources, assessors used a great variety of strategies which were aimed at the notation of all relevant evidence; in the process of *combination of evidence to attain an overall judgement,* only one of the 22 assessors stated having explicitly aimed at systematically checking whether all relevant evidence noted for the separate sources was equally reflected in the overall judgement. From these results it may be concluded that assessors show more awareness of construct under-representation with regard to consideration of evidence than with regard to the combination of evidence to attain an overall judgement.

In connection with the inclusion of all relevant evidence in the final judgement, it is important that assessors — when combining evidence in an overall judgement — not only reflect evidence which clearly fits their initial judgement, but also include and explain counter-evidence. We assumed that collaboration in pairs would encourage assessors to comply with this specific validity indicator. However, when the strategies directly related to discussion with another assessor were examined, it was found that assessors were more likely to focus on seeking confirmation of their initial judgement than to explicitly challenge each others' judgements. This constitutes an important threat to validity, namely, a disposition to consider only confirmatory evidence for an initial judgement as opposed to also considering counter-evidence or alternative interpretations.

Another important aspect of combination of evidence in an overall judgement is the balanced weighing of the available evidence. The assessors in the present study mentioned strategies which referred to the relative weighing of the three sources of data, the different aspects of interpersonal competence, and specific evidence. The present results show that whereas numerous assessors explicitly aimed to equally reflect all data-sources in their overall judgement, only a small number of assessors explicitly aimed at the balanced weighing of either the different aspects of interpersonal competence or the specific evidence. To avoid idio-syncratic weighing, it is important for assessors to engage in all the aforementioned forms of balanced weighing.

A final aspect of the combination of evidence constitutes that all inferences are supported by clear argumentation which draws upon the original data. A remarkable finding is that few assessors in the present study explicitly referred to this aspect. The results might suggest that the present assessors were more concerned with the mere recognition of the relevant evidence than with the *interpretation* of this evidence.

The assessors in the present study not only mentioned strategies which they considered successful for the consideration and combination of evidence, but also described several problems encountered during the assessment process. These problems clearly reflect *threats* to the validity of the assessment (see Tables 4.3 and 4.4). The numbers of assessors who showed awareness of these threats were, however, found to be small. In this respect it is important to note that the threats mentioned might also apply to the judgement process of the assessors who were not aware of these threats or did not experience specific problems. The results also show that whereas for some threats mentioned by the assessors a strategy to overcome these specific threats was also mentioned, the greater part of the threats mentioned

remained unresolved (i.e., no successful strategies for overcoming these threats were employed by the assessors). Although these results might suggest that awareness of potential threats does not necessarily result in the adoption of successful strategies to overcome these threats, we consider awareness of threats an important precondition for an assessment process in which construct-irrelevant variance and construct under-representation are minimized. When assessors are able to recognize their biases (i.e., threats to the validity of the judgement process) they are more likely to take conscious action during the assessment process to overcome these biases or threats (Szpara & Wylie, 2005).

4.7 Implications for the preparation of assessors

As mentioned above, it is of great importance for assessors to be aware of the nature of their judgement processes, and of strategies and threats underlying a valid judgement process. Assessor training should thus explicitly aim to increase awareness of the relevant judgement processes, in addition to focusing on the actual content or soundness of assessors' judgements. A vital part of assessor preparation programmes would be to have assessors judge a number of authentic cases and to discuss the nature of their judgement processes and the specific ways in which assessors tried to comply with relevant indicators of validity (for instance, the indicators distinguished in Table 4.1). The present overview of successful strategies undertaken during the assessment of teacher competence (see Tables 4.3 and 4.4) can certainly be considered in assessor preparation programmes. Such strategies could be discussed in explicit relation to potential threats to validity. Assessors' awareness of possible threats to a valid assessment process might increase as a result of exercises which focus on recognizing threats and encourage the adoptation of a broad repertoire of strategies to overcome these threats. Possible strategies could be discussed, with specific attention to the potential of these strategies. We suggest introducing the following issues pertaining to the processes of consideration and combination of evidence: In training in 'consideration of evidence from separate sources' (i.e., recognizing the relevant evidence for the competence to be assessed), it is important not only to exten-sively discuss the assessment framework underlying the assessment criteria, but also to attend to personal predispositions which might constitute bias. The present study highlighted a num-ber of such disabling dispositions (see Table 4.3), like a predisposition to note relatively more negative than positive evidence for interpersonal competence, and it is important for assessors to be able to recognize their personal biases. With respect to the 'combination of evidence

from separate sources to attain an overall judgement', we suggest that particular attention be paid to potentially successful strategies for reflection on all available evidence in forming the overall judgement; active searching for evidence which might challenge developing interpretations; balanced weighing of the available evidence; and supporting of all conclusions by clear argumentation which also draws upon the original data.

References

Delandshere, G., & Arens, S.A. (2003). Examining the quality of the evidence in preservice teacher portfolios. *Journal of Teacher Education*, 54(1), 57-73.

Delandshere, G., & Petrosky, A. R. (1998). Assessment of complex performances: Limitations of key measurement assumptions. *Educational Researcher*, 27(2), 14-24.

Dwyer, C.A. (1995). Criteria for performance-based teacher assessments: Validity, standards and issues. In A.J. Shinkfield, & D. Stufflebeam (Eds.), *Teacher evaluation: Guide to effective practice* (pp.62-80). Boston: Kluwer Academic Publishers.

Evans, E.D., & Tribble, M. (1986). Perceived teaching problems, self-efficacy, and commitment of teaching among preservice teachers. *Journal of Educational Research, 80* (2), 81-85.

Heller, J. I., Sheingold, K., & Myford, C.M. (1998). Reasoning about evidence in portfolios: Cognitive foundations for valid and reliable assessment. *Educational Assessment*, 5(1), 5-40.

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

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 CA: Sage Publications.

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

Messick, S. (1996). Validity of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1-18). Washington DC: Department of education, office of Educational research and Improvement.

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

Moss, P. A., Schutz, A. M., & Collins, K. M. (1998). An integrative approach to portfolio evaluation for teacher licensure. *Journal of Personnel Evaluation in Education*, 12(2), 139-161.

Moss, P.A., Girard, B.J., & Haniford, L.C. (2006) Validity in Educational Assessment. *Review of Research in Education* (Vol 30, pp.109-162).

Nijveldt, M., Beijaard, D., Brekelmans, M., Verloop, N., & Wubbels, Th. (2005). Assessing the interpersonal competence of beginning teachers: The quality of the judgement process. *International Journal of Educational Research*, 43, 89-102.

Quinlan, K. M. (2002). Inside the peer review process: How academics review a colleague's teaching portfolio. *Teaching and Teacher Education*, 18, 1035-1049.

Schutz, A., & Moss, P.A. (2004). Reasonable decisions in portfolio assessment: Evaluating complex evidence of teaching. *Education Policy Analysis Archives*, 12(33).

Szpara, M.Y., & Wylie, E.C.(2005). National Board for Professional Teaching Standards Assessor Training: Impact of Bias Reduction Exercises. *Teachers College Record,* 107 (4), 803-841.

Tigelaar, D. E. H., Dolmans, D.H.J.M, Wolfhagen, I.H.A.P., & van der Vleuten, C.P.M. (2005). Quality issues in judging portfolios: Implications for organizing teaching portfolio assessment procedures. *Studies in Higher Education,* 30 (5), 595-610.

Veenman, S.A.M. (1984). Perceived problems of beginning teachers. *Review of Educational Research, 54,* 143-178.

Wolf, A. (1995). *Competence-based assessment*. Buckingham: Open University Press.

Wubbels, Th., Brekelmans, M., Brok, P. den, & Tartwijk, J. van (2006). An interpersonal perspective on classroom management in secondary classrooms in the Netherlands. In. C. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: research, practice and contemporary issues* (pp.1161-1191). New York: Lawrence Erlbaum Associates.

5

General discussion

The aim of the research presented in this dissertation was to gain insight into how the validity of assessors' judgement processes can be ensured in the practice of authentic teacher assessment. To this end, three studies were conducted which focussed specifically on the qualities of the judgement processes of assessors. It was argued that even though the assessment procedures used and the criteria employed are important conditions for a valid judgement, the quality of the assessment as a whole depends heavily on the ability of assessors to use these procedures and criteria to reach a valid judgement. Up till now, however, few empirical studies have addressed the exact nature of assessors' judgement processes. In the work reported here, we investigated the judgement processes of assessors who judged the interpersonal competence of a student teacher using a specific assessment procedure. In this final chapter, we summarize and discuss the results of the three studies presented (sections 5.1 and 5.2). In addition, the limitations of these studies are discussed (section 5.3) and suggestions for future research offered (section 5.4). The chapter ends with several practical implications for ensuring the validity of teacher assessment practices (section 5.5).

5.1 Main conclusions

Construct-representation in assessors' judgement processes

In Chapter 2 we report on a small-scale study (Study I) in which a procedure to assess the interpersonal competence of teachers was developed and validated. The validity of the judgements of four assessors who evaluated the interpersonal competence of the same student teacher was explored using the concepts of construct-irrelevant variance and construct under-

representation. The research question of this first study was the following: To what extent do construct-irrelevant variance and construct under-representation occur in the judgment processes of consideration of evidence from separate sources and combination of evidence to attain an overall judgement? To investigate the amount of construct-irrelevant variance and construct under-representation in the consideration and combination of evidence, assessors' *notes* (i.e., evidence noted for the separate sources) and overall judgements in the form of *statements* in an evaluative summary were compared to a so-called expert judgement. The results showed the amount of construct-irrelevant variance and construct under-representation in these notes and statements to be small. That is, in both their notes and their statements, assessors practically did not refer to evidence which was irrelevant to interpersonal competence. In addition, they included all relevant aspects of interpersonal competence as operationalized in the specific assessment framework underlying the procedure. Small differences between the four assessors were found to stem primarily from the process of combining evidence from separate sources to attain an overall judgement. That is, the four assessors noted more or less the same evidence for the separate sources, but relatively greater variance was found in their statements.

On the basis of these results we conclude that the validity of the assessment of interpersonal competence — based on the present assessment procedure — could be improved specifically with regard to the judgement process of the combination of evidence from separate sources into an overall judgement. The consideration or recognition of the relevant data did not pose specific problems, as the assessors in the present study noted more or less the same data for the different components of interpersonal competence and for the different sources of data.

The nature of the collaborative judgement process

Although the results with regard to construct-representation in the notes and statements of assessors presented in Study I may be regarded as an indicator of the validity of the judgement processes of these assessors, they do not provide practical guidance for the specific ways in which the quality of the relevant judgement procedures can be ensured or improved. In Study II (Chapter 3), the judgement processes of twenty-four assessors were analyzed in greater detail. We aimed to describe the nature of assessors' judgement processes and, particularly, to examine the extent to which these processes reflected two principles of assessment mentioned

by Moss et al. (1998): (a) seeking coherence among the available sources of evidence and checking whether all relevant evidence has been taken into account, and (b) challenging developing interpretations with counterevidence and explicitly considering alternative interpretations. Engagement in these two 'key assessment processes' is assumed to minimize the introduction of construct-irrelevant variance or construct under-representation, or both. As it is assumed that discussion between assessors may enhance the chances of engagement in these processes, we examined the processes which assessors engage in when *collaboratively* (i.e., pairwise) assessing the interpersonal competence of a student teacher. The research question concerned the nature of the assessment process when teachers are assessed collaboratively. The following more specific questions were addressed: (a) What communicative activities do assessors engage in while collaboratively assessing teacher competence? (b) To what extent are the aforementioned key processes reflected in the observed collaborative judgement processes? (c) To what extent does the representation of the key processes vary across the different pairs of assessors?

The results of the study showed that the assessors clearly undertook communicative activities which reflected the key assessment processes of (a) seeking coherence among available sources of evidence and checking whether all relevant evidence has been taken into account and (b) challenging developing interpretations with counterevidence or perspec-tives, or both, of the other assessor. Inspection of the overview of communicative activities showed the two key assessment processes to be reflected particularly in the activities of "providing additional evidence for an argument"(i.e., *addition*) and "challenging an inter-pretation through the presentation of counterevidence or an alternative interpretation" (i.e., *challenging*). It seems, nevertheless, extremely difficult for assessors to engage in *both* key processes in practice. That is, the collaborative judgement processes of few collaborative pairs were found to involve both *addition* and *challenging*, called a Type I pattern of collaborative assessment. The discussions of the other pairs were characterized by either a predominantly *challenging* or Type II pattern of collaborative assessment; *addition* or a Type III pattern of collaborative assessment; or neither *addition* nor *challenging*, called a Type IV pattern of collaborative assessment. Although judgement processes involving both *addition* and *challenging* may be assumed to produce the most valid judgements, the results showed that even the quality of the collaborative assessment processes of the few pairs who were able to realize such a process could be improved. The discussion activities did not always produce clear conclusions, and some of the discussion activities were even ignored at times, meaning that the counterevidence presented and the

alternative interpretations put forth were not always reflected in the final judgements of these pairs. Even collaborative pairs who were engaged in both key assessment processes eventually showed a disposition to consider only confirmatory evidence as opposed to actively considering counterevidence or alternative interpretations.

Based on these results, it can be concluded that a judgement process in which initial interpretations are actively challenged with counterevidence or alternative interpretations, or both, which is assumed to produce the most valid judgements, is difficult to realize in practice. As such a process proved difficult to realize for collaborative pairs, we believe that taking all of the relevant evidence into account may be even harder to do individually. The validity of assessors' judgement processes can be improved when assessors engage in both desired judgement processes.

Assessors' perceptions of successful strategies to ensure validity

Chapter 4 (Study III) addresses the occurrence of construct-irrelevant variance and construct under-representation from the perspective of the assessors. Assessors' perceptions of essential judgement processes were investigated to provide an overview of successful strategies and threats underlying a valid assessment process. The following research question was addressed: Which strategies to ensure the validity of the assessment process are performed by assessors in connection with the essential judgement processes of consideration of evidence from separate sources and combination of evidence to attain an overall judgement? The overview of strategies presented in Tables 4.3 and 4.4, based on the perceptions of 22 assessors, shows that the assessors employed a great diversity of successful strategies. However, these different strategies were generally mentioned only by a small numbers of assessors. The results show that more strategies were mentioned for the process of consideration of evidence from separate sources than for the process of combining this evidence in an overall judgement. Whereas in the process of consideration of evidence from separate sources nearly all assessors used strategies aimed at the consideration of all relevant evidence, in the process of combination of evidence to attain an overall judgement few assessors were explicitly aimed at also systematically reflecting all relevant evidence in the overall judgement. From these results it may be concluded that assessors show more awareness of construct under-representation with regard to consideration of evidence than with regard to the process of combination of evidence to attain an overall judgement.

The assessors in the present study not only described successful strategies underlying a valid assessment process, but also showed awareness of several *threats* to the validity of the assessment process. The numbers of assessors who mentioned these threats were, however, small. For a number of threats mentioned by the assessors, a successful strategy to overcome these specific threats was introduced; however, the greater part of the threats mentioned remained unresolved. (i.e., no successful strategies for overcoming these threats were employed by the assessors).

On the basis of the present results it may be concluded that the validity of the assessment can be improved when assessors are able to employ a broad repertoire of strategies which underlie a valid assessment. When assessors are able to recognize threats to the validity of their judgement processes and, in addition, are aware of a range of potential successful strategies to overcome such threats, they might be more likely to use these strategies during the assessment process, which may result in a valid assessment.

5.2 Discussion

In this section, before we discuss the conclusions presented above, we consider the characteristics of the procedure used to assess the interpersonal competence of student teachers. These specific characteristics explain in part the results found, as they certainly affect the nature of the judgement processes of assessors.

Characteristics of the assessment procedure

A characteristic of the present assessment procedure which was assumed to help assessors obtain a valid picture of student teachers' interpersonal competence was an *assessment framework* in which six key components of interpersonal competence were distinguished (i.e., specific competences). Both examples and counterexamples of competent behaviour were provided for illustrative purposes for each of the six key components. The specific assessment framework used was based on an extensive empirical research on teacher-student relationships (see Wubbels, Brekelmans, Den Brok, & Van Tartwijk, 2006) and, therefore, on a well-defined and operationalized theoretical *construct*. As argued above, an assessment framework should have the right level of specificity. If the assessment criteria are too general or vague, the framework is difficult to apply fairly, and if they are too specific the framework is unlikely to capture the essence of interpersonal competence (Dwyer, 1995) and does not leave room for different

teaching styles or different teaching contexts (Uhlenbeck et al., 2002). The results of Study I show that the amount of construct-irrelevant variance and construct under-representation in the notes and evaluative summaries of the assessors was small, indicating that the present criteria were not too broad or vague. Moreover, based on the results of Study III, it can be concluded that the assessment framework allowed the assessors room to consider specific indicators of interpersonal competence which were not explicitly mentioned in the assessment framework, but clearly pertain to the six key components of interpersonal competence. This indicates that the criteria were not too specific. In general, the results of Study III show that the assessors felt well supported by the assessment framework underlying the assessment procedure, as they believed the framework helped them to consider all relevant evidence and only the relevant evidence.

Another typical characteristic of the assessment procedure used was the consultation of *three sources of data* which differed in nature. The results of Study I shed light on the added value of using an observation and self-evaluation instrument in addition to the existing QTI to assess the interpersonal competence of teachers. Although use of the QTI alone provides a reliable and valid picture of both the student and the teacher perceptions of the pattern in the stabilized interpersonal relationship between the teacher and his or her students, we assumed that it did not provide a sufficiently comprehensive picture of the interpersonal competence of teachers. It was argued that both information on a teacher's everyday behaviour in classrooms and information on the ability of teachers to reflect upon their own interpersonal competence were also needed. Such additional information was considered to enhance the validity of the assessment procedure by providing a more comprehensive picture of the teacher's interpersonal competence. The results of Study I indicate that the combination of data sources provided detailed information on the students' and teacher's perceptions of their interpersonal relationships, actual teacher behaviour, and self-reflection. In this respect it should be noted that some assessors nevertheless stated that specific interpretations or judgements could not sufficiently be supported with the data available (Study III). All of these assessors suggested discussing these specific interpretations with the candidate and giving the candidate the opportunity to react and to supply additional information. The latter suggestion is in line with a recommendation of Tigelaar et al. (2005) to organize tripartite meetings with an assessor, the candidate, and his or her supervisor. In such a manner, candidates are given the chance not only to supply additional information, but also to gain insight into the argumentation underlying the judgement given. Another note with regard to the mixing of selected sources of

data is that many assessors stated that they gave the observational instrument (i.e., analysis of a video-taped lesson) relatively more emphasis in their judgement process than the QTI and the self-evaluation assignment. One reason for this is that some assessors could not account for small incongruencies between their own observations and the students' perceptions reflected in the QTI results (Study III). Other assessors explained the emphasis on the observational instrument with the fact that they were more accustomed to working with classroom observations than to working with the QTI or self-evaluations. As the instruments are designed to provide complementary evidence, and emphasis on *one* of the three data sources might constitute idiosyncratic weighing of the available evidence, a more balanced use of the three instruments would be desirable.

The judgement processes of assessors

In the research presented, we analyzed assessors' judgement processes in three different studies. The results of these studies shed light on the nature of the assessment process and the specific difficulties in this process which may affect the soundness of the overall judgement. In our opinion, such insights provide guidance for judging validity and developing more valid judgements in practice. Although this research was conducted in a specific assessment context, we believe that the results presented clearly point to general issues in authentic assessment and are thus relevant to a wider array of assessment practices.

The results of the presented work show that the essential assessment process of combination of evidence to attain an overall judgement is more challenging for assessors than the process of consideration of evidence from separate sources. The results of Study I show the differences between assessors to stem primarily from the combination of evidence. The results of Study II and Study III indicate that the assessors in the present study put relatively much emphasis on the consideration of evidence. The results of Study II show that the judgements of five of the twelve collaborative pairs did not truly go beyond the level of specific data. That is, in their collaborative judgement processes, these assessors took the approach of summarizing the relevant data and did not clearly interpret these data so as to come up with a holistic judgement of the data. In addition, the results of Study III show that the assessors considered more strategies pertaining to the consideration of evidence than to the combination of evidence in an overall judgement. With regard to this latter process, assessors should aim to reflect all relevant evidence in the final judgement; they should not only reflect evidence which clearly fits initial judgements, but also involve and explain counterevidence and

support the final judgement by clear argumentation representing all relevant evidence. All assessors in the present study, however, showed a tendency to consider only confirmatory evidence for their initial interpretations as opposed to also considering counterevidence or alternative interpretations. The results of Study II show that even when assessor pairs *do* discuss counterevidence or alternative interpretations, or both, this does not always produce clear conclusions, indicating that counterevidence found and alternative interpretations put forth are not reflected in the argumentation of assessors and, therefore, in the final judgement. The results of Study III show that a great many assessors used the discussion with another assessor primarily to seek confirmation of their initial judgement, and did not aim at challenging each other's or their own judgements. In connection with this, the results of Study III implicate that many assessors were aware of slight incongruencies between the different sources of data. For example, a considerable number of assessors found the students' perceptions put forth in the QTI to be considerably more positive than their own judgements based on the video-taped lesson and the teacher's perceptions reflected in the QTI and the self-evaluation assignment. Whereas some of these assessors aimed to explain why students were more positive about the student teacher than they expected at first hand, and altered their initial judgement accordingly (i.e., based on this counterevidence), other assessors gave the student perceptions less emphasis than their own observations, without searching for a coherent explanation. The latter approach constitutes an important threat to the validity of the assessment process, namely, the threat of idiosyncratic weighing.

In our study, we suggested that it is important for assessors to be aware of their own judgement processes and the specific threats to validity in these processes. Such awareness is likely to enable assessors to take conscious actions during the assessment process to overcome such threats to validity and, in this connection, to reach a sound judgement. With respect to the process of combination of evidence to attain a sound overall judgement, awareness of the process particularly involves being able to elicit the line of argumentation underlying the overall judgement. Being able to elicit the line of argumentation makes it possible to evaluate the extent to which assessors comply with the validity indicators summarized in Table 4.1. (i.e., all relevant evidence is clearly reflected in the overall judgement; both confirmative and counterevidence are considered; no idiosyncratic weighing of evidence occurs; and all conclusions are supported by clear argumentation which also draws upon the original data). Note that the indicators of validity employed in this study bore clear resemblance to the criteria of credibility (i.e., interpretations of assessors matched those of stakeholders in the field), dependability (i.e.,

the interpretation process was established, traceable, and documented), and confirmability (the interpretations and conclusions were supported by the original data) described in Chapter 1. Like the indicators of validity distinguished in our study, these criteria clearly pertain to the quality of the interpretation process of the assessors. Measures to establish and evaluate these criteria would include the requirement for assessors to write extensive interpretative summaries in which the argumentation leading to the overall judgement is documented (Moss et al., 1998), and checking of interpretations with involvement of the candidates (Tigelaar et al., 2005).

The specific methods used in our three studies also provide examples of the ways in which the quality of these judgement processes could be evaluated. In essence, the three different analyses of the judgement processes in this dissertation represent three forms of validity inquiry. In Study I (Chapter 2), the soundness of the judgements of the assessors was judged using the general concepts of construct-irrelevant variance and construct under-representation. The occurrence of construct-irrelevant variance and construct under-representation was operationalized with respect to the judgement processes of the consideration of evidence from separate sources and the combination of evidence in an overall judgement. To investigate the amount of construct-irrelevant variance and construct under-representation in the consideration and combination of evidence, the assessors' *notes* (i.e., evidence noted for the separate sources) and overall judgements in the form of *statements* in an evaluative summary were compared to an expert judgement. In this manner, we checked whether both notes and statements referred to interpersonal competence only (no construct-irrelevant variance) and whether both notes and statements covered all components of interpersonal competence (no construct under-representation). In Study II (Chapter 3), we explored the extent to which assessors were able to realize the following principles of assessment: (a) seeking coherence among the available sources of evidence and checking whether all relevant evidence has been taken into account, and (b) challenging developing interpretations with counterevidence and explicitly considering alternative interpretations. Engagement in these two 'key assessment processes' is assumed to minimize the introduction of construct-irrelevant variance or construct under-representation, or both, and can be considered an indicator of sound interpretations. In Study III, the concepts of construct-irrelevant variance and construct under-representation were again taken as a starting point for our analysis. In connection with the essential judgement processes of consideration of evidence and combination of evidence to attain an overall judgement, we investigated the strategies to ensure the validity of the assessment process used by assessors, according to the assessors themselves. These three

specific methods used for the analysis of assessors' judgement processes can be regarded as an example of validity inquiry of authentic assessment.

5.3 Limitations of the study

In this section we indicate some limitations of our study pertaining to the application of the assessment procedure and the research design. Firstly, all the assessors in this study were using the procedure to assess the interpersonal competence of student teachers for the first time. Even though the participating assessors were carefully selected on the basis of their experiences with use of the QTI for supervision purposes and experiences with supervision or evaluation of student teachers in general, it is often asserted that the expertise of assessors stabilizes as they gain experience with a specific assessment procedure (e.g., Uhlenbeck et al., 2002). In several fields of research, 'expert-research' is a common approach to gain in-depth insight into complex cognitive activities. We expect that experienced or expert assessors who are well accustomed to their roles may provide the most interesting results in further research. The assessors in the present research differed with regard to their prior experiences with the assessment of teacher competences, and based on general impressions from our three studies, we expect that relatively experienced assessors would be most likely to elicit 'good practices' or high-quality judgement processes, characterized by in-depth consideration and combination of the available evidence. In addition, compared to inexperienced assessors, more experienced assessors are expected to show greater awareness of the nature of their judgement process, the threats to validity in this process, and possible strategies to overcome these threats. Therefore, it might also be easier for them to describe the salient aspects of their judgement processes. Research with expert assessors might thus specifically shed light on critical characteristics of high-quality assessment processes.

Secondly, the assessors in the present research assessed only one student teacher, who we expected to be an 'average' student teacher with respect to the characteristics to be assessed. We aimed to generate as much variation between assessors' judgement processes as possible, and therefore included a maximum number of assessors. However, examination of the interpersonal competence of a variety of student teachers using the present assessment procedure might provide further insight into the nature of the assessment process. Particularly student teachers who present clear ambiguities between the different sources of data may elicit a greater variance in the nature of assessors' judgement processes, for instance, more

disagreement among assessors and greater consideration of counterevidence or alternative interpretations, or both.

Thirdly, in Study III, the perceptions of the assessors were examined using semi-structured retrospective interviews. The interview questions triggered in some, but not all, of the assessors an extensive description of their judgement process. Note that it is more difficult for some people to express their perceptions in this way than for others, as perceptions might be of a tacit nature (e.g., Calderhead, 1996; Kagan, 1992). Additional methods like stimulated recall or retrospective think-aloud protocols (Van der Schaaf, Stokking, & Verloop, 2005), or focus groups (Krueger & Casey, 2000), which would enable assessors to respond to each other's perceptions of the judgement process, might reveal more tacit perceptions or processes, and might thus provide greater insight into the nature of assessors' judgement processes.

Fourthly, the results of Studies II and III were based on the judgement processes of 24 and 22 assessors, respectively. Although four types of collaborative assessment were distinguished in Study II, the number of collaborative pairs which made up these four categories was small. Also, the absolute frequencies of the different communicative activities which were taken to be an indicator of the nature of the collaborative judgement process were rather small at times. A larger sample of assessors might have provided a more detailed description of the four categories distinguished. With regard to Study III, a larger sample of assessors would possibly have generated additional successful strategies and threats underlying the validity of the assessment process.

Finally, as stated above, the assessment framework which guided the judgement processes of the assessors in the present study was based on a well-defined theoretical construct. Such a well-conceptualized and well-operationalized construct was not available for all relevant teacher competences as was the case for 'interpersonal competence'. An advantage of the selection of this particular competence for assessment in the present research is that this clearly enabled us to focus on the assessment process itself, as opposed to putting much emphasis on developing a definition of the competence to be assessed according to certain procedures. A limitation of this choice for a well-operationalized competence, however, pertains to the generalizability of the results to other assessment procedures. Although our results clearly point to general issues in authentic assessment, other assessments of teacher competence might be based on less well-defined constructs, resulting in additional issues in the judgement processes of assessors.

5.4 Suggestions for future research

Below, we elaborate on two directions for future research: further investigations into assessors' judgement processes and the effects of teacher assessment on the professional development of teachers.

Research into assessors' judgement processes

So far, few empirical studies have addressed the nature of the judgement processes which underlie valid teacher assessment. Insight into the nature of assessors' judgement processes is, however, needed to provide guidance for obtaining and evaluating validity in teacher assessment, which depends heavily on the professional judgement of assessors. In this dissertation, we explored the nature of the judgement processes of assessors in evaluating teacher competence, and the extent to which assessors engage in relevant judgement processes. For future research into the quality of assessors' judgement processes, we suggest further investigation of the specific conditions which enable assessors to engage in the relevant judgement processes.

A potential enabling condition which we explored in the present research was discussion between pairs of assessors after they had formed an initial, individual overall judgement. Of course, collaboration should not be a goal in itself: it is a way to encourage assessors to reflect all relevant evidence in the overall judgement, to carefully explicate the line of argumentation which led to the final judgement, and to explicitly consider counterevidence and alternative interpretations which might challenge initial interpretations. Think-aloud sessions or interviews with expert assessors may provide further insight into the conditions which enable either individual assessors or collaborative pairs to engage in such key processes. In future research, moreover, not only the enabling conditions for valid assessment (or successful strategies underlying valid assessment) but also the disabling conditions (or threats underlying valid assessment) should be taken into consideration.

Another important subject of inquiry would be the actual soundness of assessors' argumentation. To what extent are assessments and conclusions supported by the original data? Closer examination of the argumentation of assessors might provide further insight into how the validity of the assessment process can be ensured in practice, and may help refine the relevant criteria for validity in teacher assessment.

Research into the consequences of assessment for teachers' professional development

In our studies into validity in teacher assessment, the focus was on the soundness of the judgement processes of assessors, and more specifically on construct representation and the extent to which assessors' interpretations are based on the total set of relevant evidence. We also discussed the validity of the assessment procedure itself. A comprehensive inquiry into the validity of teacher assessment, however, would also include inquiry into the consequences of the assessment. In our view, a high-quality assessment procedure and sound interpretations of assessors can certainly be regarded as a prerequisite for consequential validity (i.e., providing clear directions for further professional development of teachers). The results of the present study do not provide explicit implications or criteria for enhancing the consequences of assessment for professional development. Does teacher assessment based on a procedure such as that followed in this study foster reflective skills and meaningful dialogue or debate among teachers with regard to the essence of good teaching? What additional criteria for the presentation of the overall judgement to the candidate or feedback are needed? And how should candidates be engaged in the different phases of the assessment process (Davis & LeMahieu, 2003)? In future research into the effects of teacher assessment on teachers' professional development, the relations between the nature of the assessment process, on the one hand, and the quality of the feedback or the meaningfulness for the candidate, on the other hand, may be investigated.

5.5 Implications for assessment practices

Assessment practices

Assessment procedures and practices should be designed to support desirable judgement processes of assessors. The aforementioned issues with regard to the assessment process may be translated into the following practical suggestions for the organization of assessment procedures and practices.

Consideration of evidence from separate sources The assessment framework which formed the basis of the assessment in the present study was designed to be not too broad and certainly not too specific or check-list like, so that it would leave room for different teaching styles and contexts. Some assessors in the present study felt they considered indicators of interpersonal competence which were not explicitly mentioned in the assessment framework, but yet are

clearly part of interpersonal competence (Study III). This finding stresses the importance of a suggestion by Moss et al. (1998), to frequently discuss the assessment framework based on specific cases at hand. Specific cases might raise issues which stimulate further discussion about the essence of the competence to be assessed.

Combination of evidence from separate sources to attain an overall judgement As stated in our general introduction, in the context of authentic assessment, it is to be left to the assessors to determine how the various sources of data are exactly combined into an overall judgement, as the complexity of teaching makes a strict marking scheme inappropriate. To guarantee the soundness of this process of combination of data, it is important that the process be transparent and that assessors comply with the validity criteria outlined in Table 4.1: namely, (a) all relevant evidence is clearly reflected in the overall judgement; (b) both confirmative and counterevidence are considered; (c) no idiosyncratic weighing of evidence occurs; and (d) all conclusions are supported by clear argumentation which also draws upon the original data. We suggest that, as a part of any assessment procedure, assessors should explicitly check whether their overall judgement constitutes a coherent interpretation of the total set of evidence. Transparency of the trail of evidence leading to the interpretations allows users to evaluate the conclusions for themselves, and can, therefore, be regarded as a criterion which supports the soundness of interpretations (Moss, 1994). Following Moss et al. (1998), we suggest assessment practices should require assessors to write an 'interpretative summary' in which the argumentation leading to the overall judgement is carefully documented. Should assessors feel that specific interpretations cannot be sufficiently supported on the basis of the data available, or notice incongruencies between the different sources of data for which they cannot find a coherent explanation, they should explicitly document these uncertainties. And when such uncertainties arise, assessors should particularly be enabled to check their interpretations with the candidates involved. When necessary, the candidate should be given the opportunity to provide additional information (cf. Moss et al., 1998). Note, however, that documentation of the argumentation leading to the overall judgement might be easier said than done for assessors. As stated above, although the assessors in the present study were required to write such an interpretive summary, the summaries of many assessors could be classified as mere summaries of the evidence considered rather than summaries of the *argumentations* based on this evidence. Extensive training would be needed to help assessors to elicit their interpretation processes, specifically with regard to the combination of evidence in an overall judgement. In section 5.5.2., we elaborate on implications for assessor preparation programmes.

Collaborative assessment Should developers of assessment procedures opt for a procedure which comprises collaborative assessment, the expected benefits of such an approach should outweigh issues of cost and efficiency. On the basis of the present research, it is difficult to draw firm conclusions with regard to the potential advantages of collaborative assessment over individual assessment, as we did not explicitly compare these different approaches to assessment. A middle ground between a fully individual and a collaborative approach would be to give individual assessors the possibility to discuss their interpretations with another assessor when particular judgements raise difficult issues.

Assessor preparation programmes

The literature provides little guidance for the combination of non-standardized evidence from different sources of data and from different contexts. Therefore, the assessors in our study were not provided with highly specific guidelines in this respect, as we were particularly interested in the approaches assessors would consider themselves. Based on the insights generated by our study, we present the following suggestions for the training of assessors.

Consideration of evidence from separate sources As a part of the assessor training, assessors should be provided with a number of authentic teacher cases and be given the opportunity to express a thoroughly argued judgement using the specific framework underlying the assessment procedure. Assessors should then be enabled to extensively discuss the assessment framework, based on their evaluation of the specific cases at hand. An additional preparation for the process of consideration of evidence would be to discuss possible strategies and threats underlying the consideration of evidence, illustrated in Chapter 4. In our study, we argued that awareness of personal predispositions which might constitute bias is assumed to encourage the adoption of successful strategies to overcome such bias.

Combination of evidence in an overall judgement In assessor preparation programmes, extensive attention should be given to the process of combination of evidence from different sources to attain an overall judgement. The findings of all three studies show that this specific process is the most difficult and intangible for assessors. Assessors should be prepared to compose a coherent interpretation of the total set of evidence and to carefully explicate the argumentation which led to the judgement given. The validity indicators outlined in Table 4.1 might provide guidance for the combination of evidence. Again, based on the discussion of a number of authentic teacher cases, assessors may extensively discuss possible strategies and threats underlying the combination of evidence, as illustrated in Chapter 4.

Collaborative assessment When discussion with another assessor is part of the assessment procedure, we suggest that the preferred character of this collaboration be extensively addressed in assessor preparation programmes. The overview of the communicative activities undertaken during the collaborative assessment of teacher competence (Study II) can certainly be introduced into teacher assessment preparation programmes. The four types of collaborative assessment, together with their relative strengths and weaknesses, should be considered in such a programme. This would allow the discussion of such validity-related issues as construct irrelevant variance and construct under-representation to be made more tangible.

References

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

Davis, A., & LeMahieu, P. (2003). Assessment for learning: Reconsidering portfolios and research evidence. In M. Segers & F. Dochy & E. Cascallar (Eds.), *Optimising new modes of assessment: in serach of qualities and standards* (pp. 174-171). Dordrecht: Kluwer Academic Publishers.

Dwyer, C.A. (1995). Criteria for performance-based teacher assessments: Validity, standards and issues. In A.J. Shinkfield, & D. Stufflebeam (Eds.), *Teacher evaluation: Guide to effective practice* (pp.62-80). Boston: Kluwer Academic Publishers.

Kagan, D.M. (1992). Implications of Research on Teacher Belief. *Educational Psychologist, 27,* 65-90.

Krueger, R.A., & Casey, M.A. (2000). *Focus Groups: A practical guide for applied research.* California: Sage Publications.

Tigelaar, D.E.H., Dolmans, D.H.J.M., Wolfhagen, I.H.A.P., Van der Vleuten, C.P.M. (2005). Quality issues in judging portfolios: Implications for organizing teaching portfolio assessment procedures. *Studies in Higher Education,* 30(5), 595-610.

Uhlenbeck, A.M., Verloop, N., & Beijaard, D. (2002) Requirements for an assessment procedure for beginning teachers: Implications from recent theories on teaching and assessment. *Teachers College Record, 104*, 242-272.

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.

Wubbels, Th., Brekelmans, M., Den Brok, P., & Van Tartwijk, J. (2006). An interpersonal perspective on classroom management in secondary classrooms in the Netherlands. In. C. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: research, practice and contemporary issues* (pp.1161-1191). New York: Lawrence Erlbaum Associates.

Appendix

Assessment framework: 'Interpersonal competence'

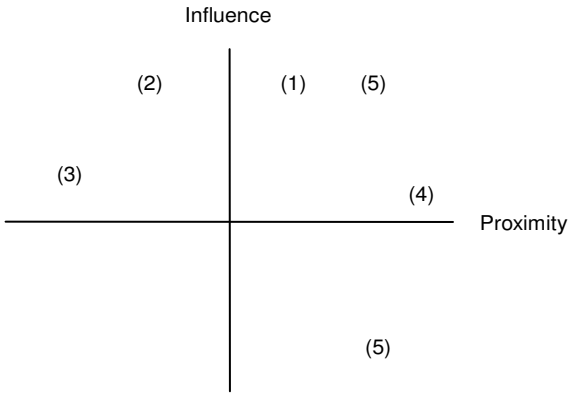
A. Analysis of interpersonal competence in a video-taped lesson

All teacher behaviour in a specific lesson can be evaluated against the extent to which this behaviour can be seen as an indicator of interpersonal competence. To guide such an analysis of interpersonal competence in a video-taped lesson, five key components of interpersonal competence were distinguished. Together, these five components provide a comprehensive view of interpersonal competence. An interpersonally competent teacher is able to apply the appropriate component in the appropriate situation. Generally, in every single lesson, teachers will demonstrate all five components of interpersonal competence.The five components of interpersonal competence are

(1) Providing clear structure and guidance for students
(2) Setting norms and standards for students
(3) Correction of undesirable student behaviour
(4) Attention to students
(5) Provision of freedom and responsibility for students

The five components of interpersonal competence are based on the Model for Interpersonal Teacher Behaviour (MITB). In the figure, the components are depicted according the dimensions of the MITB.

.



The five components differ in the extent to which the teacher decides what happens in the classroom (Influence) and in the amount of emotional distance between the teacher and the students (Proximity). The higher the position on the influence-dimension, the more the teacher influences what happens in the classroom. The more to the right on the proximity-dimension, the smaller the emotional distance between teacher and students. Below, we describe the components in greater detail, and for each of the components both examples and counterexamples of competent behaviour are provided for illustrative purposes.

1. Providing clear structure and guidance for students

The interpersonally competent teacher

A teacher who is interpersonally competent is able to make clear when the lesson starts and to motivate the students to actively participate. During the lesson, the teacher shows enthusiasm for the content of the lesson and for working with the students. It appears – from the reactions of the students – that the teacher is able to provide students with clear instructions. The amount of time the teacher uses for his or her instructions is not greater than strictly necessary. The teacher clearly has an overview of the class and is able to take a central position during plenary moments.

Possible examples of competent behaviour

The teacher

- is able to make clear when the lesson starts;
- is able to make students active learners;
- activates students;
- is able to interest students
- communicates enthusiasm for his or her subject;
- communicates enthusiasm for working with the students;
- provides students with clear instructions;
- uses for his or her instruction no more time than strictly necessary;
- keeps plenary moments functional;
- is able to make clear which activities he or she expects from students;
- realizes smooth transitions between the different parts of the lesson;
- takes a central position during plenary moments;
- shows the students that he or she has an overview of the individual students and the class as a whole;
- shows the students that he or she has expertise in the subject-matter.

Possible counterexamples of competent behaviour

The teacher

- does not manage to activate the students;
- does not communicate enthusiasm for his or her subject
- does not communicate enthusiasm for working with the students;
- does not manage to interest students;
- appears insecure;
- allows students to take control of the class;
- shows a lack of overview of the class;
- involves only part of the class;
- allows students to interrupt frequently during plenary moments;
- frequently interrupts plenary moments his or herself.

Guiding questions

- How does the teacher behave during plenary moments?
- In what way does the teacher provide structure and guidance to the students?

2. Setting norms and standards for students

The interpersonally competent teacher

A teacher who is interpersonally competent is able to make clear what specific behaviour he or she expects from the students, that is, which norms and standards he expects the students to comply with in his or her class. The teacher noticeably lives by his own norms and standards, and explicates these when necessary. When explicating norms and standards, the teacher clearly presents these as a matter of course, without any aggression and without communicating distrust in the students.

Possible examples of competent behaviour

The teacher

- makes clear what behaviour is expected from students;
- communicates whether or not students meet his or her expectations;
- acts as a role model;
- reminds students of rules when necessary;
- explicates his or her rules in such a way that students perceive these as self-evident;
- formulates rules positively ("please do" as opposed to "do not");
- does as he or she says (for example, when the teacher asks for silence, he or she waits until the class is indeed silent before continuing the lesson);
- employs a limited set of rules.

Possible counterexamples of competent behaviour

The teacher

- does not manage to make clear what behaviour is expected from students;
- continually changes the norms and standards set;
- accompanies the explication of norms and standards with aggressive behaviour
- communicates distrust in students;
- does not live by his or her own rules;
- sets rules which are perceived by students as unreasonable;
- is not able to convince the students of the necessity of his or her rules (for example, the teacher acts as if he or she does not stand behind the rules him or herself, or apologises for the rules);
- employs standards which are either too high or too low for students.

Guiding question

- How does the teacher make clear what is allowed in his or her class?

3. Correction of undesirable student behaviour

The interpersonally competent teacher

The interpersonally competent teacher intervenes timely in case of undesirable student behaviour, and is, therefore, able to employ corrections which are as low-intrusive as possible. The teacher has a broad repertoire of corrections at his or her disposal, particularly of low-intrusive corrections. When the teacher starts with low-intrusive corrections, the rest of the class is not – or almost not – disturbed, and the teacher still has the possibility to follow up with a correction of a slightly higher intensity if necessary. In addition, the teacher clearly aims his or her corrections at the behaviour or offence of the student and not at the personality of the student. The teacher takes care that the student does not feel threatened. Corrections are addressed to the right students, and – depending on the specific situation – are addressed to an individual student, a group of students, or the class as a whole. It is clear to students to whom the correction is addressed. The teacher waits until students respond to a correction before continuing the lesson, and in this way intensifies the effect of corrections.

Possible examples of competent behaviour

The teacher

- intervenes timely in case of undesirable student behaviour;
- is able to employ corrections which are as low-intrusive as possible;
- shows a broad repertoire of corrections, particularly of low-intrusive corrections;
- aims corrections at student's behaviour and not student's personality;
- clearly addresses his or her corrections to the right student or group of students;
- checks whether students respond to his or her corrections;
- clearly adjusts the intensity of corrections to the nature of the student's offence;
- is able to switch from low-intrusive corrections to corrections of a higher intensity when necessary.

Possible counterexamples of competent behaviour

The teacher

- does not manage to employ mainly low-intrusive corrections;
- uses "more of the same" corrections;
- tries to be consistent in corrections at all costs;
- does not grant students a new chance after correction;
- tends to avoid confrontation with the students;
- creates an unfriendly, unsafe atmosphere through his or her performance;
- makes comments at the expense of individual students;
- looks away while correcting students' behaviour;
- smiles when correcting students' behaviour;
- speaks very loudly for a large amount of time when correcting students.

Guiding questions

- How does the teacher respond to undesirable student behaviour?
- How does the teacher check the effects of his or her interventions?

4. Attention to students

The interpersonally competent teacher

A teacher who is interpersonally competent shows interest in students, takes into account students' feelings and, in connection with this, makes allowances. The teacher regularly expresses praise and appreciation for students by paying compliments. In addition, the teacher is patient with students, for instance, giving them sufficient time to react to questions or to complete assignments, or being willing to explain more than once. The teacher moreover manages to adjust to the experiences and performance level of the students.

Possible examples of competent behaviour

The teacher

- is able to maintain a balance between emotional proximity and distance between him or herself and the students;
- shows personal interest in students;
- responds to signals of the students;
- when necessary, makes allowances that result in a desirable solution for both teacher and students;
- praises students whenever appropriate, at the right level of intensity;
- is patient with students;
- adjusts his or her lessons to the experiences of students;
- adjusts his or her lessons to the performance level of the students;
- makes students feel accepted and appreciated;
- listens actively to students;
- shows a broad repertoire of compliments;
- greets the students when they enter the classroom;
- is approachable for students.

Possible counterexamples of competent behaviour

The teacher

- does not show interest in the students;
- addresses students in a disrespectful manner;
- shows clear preferences in his or her treatment of individual students;
- aims to act as "student-with-the-students".

Guiding questions

- In what way does the teacher give attention to students?
- How does the teacher respond to students' signals?

5. Provision of freedom and responsibility for students

The interpersonally competent teacher

The interpersonally competent teacher is able to determine how much freedom and responsibility a specific classroom group can handle in a specific situation. When providing freedom or responsibility, the competent teacher is able to return to his or her leadership role whenever he or she chooses to. The teacher is able to make clear to students how much freedom they have in a specific situation. Giving responsibility means, for instance, that the teacher encourages contributions and initiatives of students, or responds to relevant suggestions of students.

Possible examples of competent behaviour

The teacher

- gives the students an appropriate amount of responsibility;
- gives the students an appropriate amount of freedom;
- when providing freedom or responsibility, is able to return to his or her leadership role when he or she chooses to;
- addresses relevant questions of students;
- responds to relevant suggestions of students;
- encourages initiatives of students.

Possible counterexamples of competent behaviour

The teacher

- gives the students either too much or too little responsibility and freedom;
- cuts short any and every initiative on the part of the students;
- when providing freedom or responsibility, shows behaviour which students perceive as insecure or weak;
- sticks too rigorously to his or her own ideas.

Guiding questions

- What kind of responsibility do students have in this teacher's class?
- What is the role of the teacher in this respect?

B. Analysis of reflection on interpersonal competence in a self-evaluation assignment

To guide the analysis of the self-evaluation assignment, we distinguished two aspects of reflection:

- Awareness of one's own interpersonal competence;
- Monitoring one's own professional development in the area of interpersonal competence.

Below, we elaborate on these two aspects of reflection.

6a. Awareness of one's own interpersonal competence

Possible indicators of competence

- The teacher shows awareness of the effects of his or her behaviour on the behaviour of the students: the teacher is able to indicate how students respond to his or her behaviour;
- The teacher's self-awareness (reflected in the teacher's analysis and in perceived strengths and weaknesses) corresponds to the impression of the assessors, based on the video-taped lesson and the QTI report.
- The teacher refers to adequate examples from the video-taped lesson and QTI report to describe his own strengths and weaknesses. The teacher is able to describe why the examples chosen are relevant.

Guiding questions

- Is the teacher able to formulate an adequate analysis of his or her behaviour?
- Does the teacher show an awareness of his or her strengths and his or her weaknesses?

6b. Monitoring one's own professional development

Possible indicators of competence

- The points of improvement mentioned by the teacher him or herself correspond to the points of improvement that the assessors deduced from the video-taped lesson and the QTI report;
- The teacher is able to formulate a concrete plan for further professional development;
- The teacher is aware of a number of alternatives for present – less adequate – interpersonal behaviour.

Guiding questions

- Is the teacher able to formulate which aspects of interpersonal competence can be improved?
- Is the teacher able to describe just how further professional development can be realized?

Nederlandse samenvatting

In dit proefschrift staat validiteit in de beoordeling van docentcompetenties centraal. Als gevolg van de toegenomen aandacht voor de kwaliteit en professionaliteit van docenten wordt beoordeling van docentcompetenties steeds gebruikelijker. Voor vele vormen en niveaus van onderwijs zijn inmiddels competenties gespecificeerd die docenten nodig hebben om goed te kunnen functioneren in deze specifieke onderwijscontexten. Zowel het formuleren van competenties als het beoordelen ervan dragen bij aan de professionalisering van docenten: ze bevorderen discussie over de essentie van goed onderwijzen evenals reflectie van docenten op het eigen functioneren. Om de kwaliteit van competentiebeoordelingen te waarborgen en optimale leereffecten te realiseren, moeten beoordelingsprocedures aan bepaalde voorwaarden voldoen. De procedure moet recht doen aan de complexiteit van het leraarsberoep, aan de specifieke context waarin de docent werkt en aan de persoonlijke stijl en doelen van de docent. Om aan deze voorwaarden te kunnen voldoen, worden beoordelingen van docentcompetenties doorgaans gebaseerd op meerdere databronnen, zoals lesopnames, lesplannen, lesmateriaal en reflecties van de docent op zijn of haar eigen functioneren.

In dergelijke vormen van competentiebeoordeling spelen beoordelaars een cruciale rol. De beoordelingsprocedure en beoordelingscriteria zijn belangrijke condities voor het realiseren van een valide oordeel, maar de kwaliteit van de beoordeling als geheel hangt vooral af van het vermogen van de beoordelaars om deze procedure en criteria te kunnen gebruiken. Bij beoordeling gebaseerd op meerdere databronnen worden doorgaans de volgende essentiële beoordelingstaken onderscheiden: (a) aanwijzen van 'bewijs van competentie' uit de afzonderlijke bronnen en (b) combineren van dit bewijs tot een totaaloordeel over de kandidaat. Deze taken vragen om veel interpretatie door beoordelaars. De validiteit van beoordelingen hangt daardoor voor een belangrijk deel af van de kwaliteit van de beoordelingsprocessen van de beoordelaars.

Twee belangrijke bedreigingen van de validiteit van beoordelingsprocessen zijn 'construct-irrelevante variantie' en 'construct onderrepresentatie' (Messick, 1989, 1996). *Construct-irrelevante variantie* ontstaat wanneer beoordelaars er irrelevante, idiosyncratische criteria op na houden en bewijsmateriaal in overweging nemen dat niet relevant is voor de te beoordelen competentie. *Construct onderrepresentatie* wordt gekenmerkt door het niet in beschouwing nemen

van relevante criteria of relevant bewijs of de idiosyncratische afweging van ver-schillende criteria of bewijs. Dit resulteert in onvoldoende aandacht voor relevante aspecten van de te beoordelen competentie. Empirische studies van Moss et al. (1998) en Schutz & Moss (2004) — uitgevoerd in de context van beoordeling van docentportfolio's — illustreerden dat beoordelaars al heel snel een indruk hebben van kandidaten en ongewild de neiging hebben om vooral te zoeken naar bewijs dat deze indruk bevestigt. Dit is een belangrijke bedreiging van de validiteit van het beoordelingsproces. Op basis van deze bevinding introduceerden zij de volgende principes om sturing te geven aan het beoordelingsproces. Het *eerste* principe houdt in dat beoordelaars coherentie zouden moeten zoeken tussen het beschikbare bewijsmateriaal en expliciet na zouden moeten gaan of al het relevante bewijsmateriaal in overweging genomen is. Een *tweede*, verwant principe houdt in dat beoordelaars aangemoedigd zouden moeten worden om expliciet te zoeken naar bewijs dat de zich ontwikkelende indruk zou kunnen ontkrachten, en om expliciet na te gaan of er alternatieve interpretaties mogelijk zijn van het beschikbare bewijs. De veronderstelling is dat wanneer beoordelaars weten te voldoen aan deze twee principes, bias wordt geminimaliseerd en ook de kans op construct-irrelevante variantie en construct onderrepresentatie kleiner wordt. Samenwerking ofwel discussie tussen beoordelaars, nadat zij een voorlopige individuele indruk hebben gevormd, zou beoordelaars bij uitstek kunnen stimuleren om te voldoen aan de genoemde principes.

Ondanks het feit dat de validiteit van de beoordeling afhangt van de wijze waarop beoordelaars essentiële beoordelingsprocessen invullen, heeft slechts een aantal studies zich expliciet gericht op de aard of kwaliteit van beoordelingsprocessen van beoordelaars. Dergelijke studies dragen echter bij uitstek bij aan de discussie over de validiteit van de beoordeling van docentcompetenties, omdat ze inzicht geven in de principes die ten grondslag liggen aan een valide beoordelingsproces. Moeilijkheden die beoordelaars bijvoorbeeld ervaren tijdens het beoordelingsproces zijn te koppelen aan bedreigingen van de validiteit van de beoordeling.

De centrale vraag in dit onderzoek luidde: *Hoe kan validiteit van beoordelingen van docentcompetentie gewaarborgd worden?* Deze probleemstelling is nader geconcretiseerd in drie onderzoeksvragen waarin het beoordelingsproces van beoordelaars centraal wordt gesteld:

(1) In hoeverre komt construct-irrelevante variantie en construct onderrepresentatie voor in de beoordelingsprocessen van beoordelaars?

(2) Wat is de aard en kwaliteit van beoordelingsprocessen van beoordelaars die paarsgewijs een docent beoordelen?

(3) Welke strategieën om tot een valide oordeel te komen hanteren beoordelaars in relatie tot de essentiële processen van aanwijzen van bewijsmateriaal uit de afzonderlijke bronnen en combineren van dit bewijsmateriaal tot een totaaloordeel?

De onderzoeksvragen zijn beantwoord in drie studies. In deze drie studies analyseerden we de beoordelingsprocessen van beoordelaars die de *interpersoonlijke competentie* van een docent-in-opleiding (dio) beoordeelden aan de hand van een hiervoor ontwikkelde beoordelingsprocedure. Er is gekozen voor beoordeling van interpersoonlijke competentie — het kunnen creëren van een goede werksfeer en een goede relatie met leerlingen — omdat dit voor veel dio's een belangrijk aandachtspunt is. De ontwikkelde beoordelingsprocedure is gebaseerd op drie bronnen: een door de dio geselecteerde video-opname van een les, de Vragenlijst Interpersoonlijk Leraarsgedrag (VIL) die een valide en betrouwbaar beeld geeft van de percepties van leerlingen van de interpersoonlijke relatie met hun docent, en een zelf-evaluatie van de dio waarin hij of zij de eigen interpersoonlijke competentie analyseert met zo veel mogelijk verwijzing naar de lesopname en de VIL-resultaten. Om de beoordelaars te ondersteunen bij het aanwijzen van bewijsmateriaal en het combineren van bewijsmateriaal tot een totaaloordeel, werd een beoordelingskader aangeboden waarin zes essentiële aspecten van interpersoonlijke competentie worden onderscheiden en geïllustreerd met mogelijke indicatoren en contra-indicatoren.

In totaal werden vierentwintig personen geselecteerd om als beoordelaar deel te nemen aan de studie. Onder hen waren lerarenopleiders (N=4) en zogeheten 'begeleiders-op-school' van dio's (N=20). Allen hadden ervaring met het evalueren van dio's en het geven van feedback. Vijf van hen hadden daarnaast ervaring als beoordelaar van competenties van zij-instromers. Ter voorbereiding op de studie volgden alle deelnemers een beoordelaarstraining van twee avonden. De eerste avond werd de ontwikkelde procedure voor het beoordelen van interpersoonlijke competentie geïntroduceerd en werd het beoordelingskader uitvoerig bediscussieerd. De tweede avond beoordeelden de beoordelaars een dio op basis van door een dio ter beschikking gesteld materiaal. Volgens de procedure vormden zij eerst individueel een indruk; daarna ging iedere beoordelaar in discussie met een andere beoordelaar. De avond werd afgesloten met een vergelijking van de eindoordelen en een plenaire discussie over de gevolgde aanpak.

In **hoofdstuk 2** wordt een kleinschalig onderzoek (Studie I) naar de validiteit van de ontwikkelde beoordelingsprocedure beschreven. Om een beeld te krijgen van het voorkomen van construct-irrelevante variantie en construct onderrepresentatie werden de oordelen van vier beoordelaars geanalyseerd. De onderzoeksvraag die centraal stond in deze eerste studie was de volgende: In hoeverre komt construct-irrelevante variantie en construct onderrepresentatie voor in de beoordelingsprocessen van beoordelaars? Om deze vraag te beantwoorden werden zowel de *aantekeningen* van de beoordelaars (het aangetekende bewijsmateriaal bij de drie databronnen) als hun totaaloordelen in de vorm van *uitspraken* in een evaluerende samenvatting vergeleken met een 'expertoordeel'. De resultaten wezen uit dat beoordelaars vrijwel geen construct-irrelevante variantie en construct onderrepresentatie introduceerden: beoordelaars verwezen in hun aantekeningen en in hun totaaloordeel vrijwel niet naar bewijsmateriaal dat geen onderdeel is van interpersoonlijke competentie en betrokken alle relevante aspecten van interpersoonlijke competentie in hun oordeel. Opvallend was dat de verschillen tussen beoordelaars hoofdzakelijk ontstonden tijdens het proces van combineren van data: de vier beoordelaars wezen nagenoeg hetzelfde bewijsmateriaal aan, maar in de uiteindelijke oordelen waren grotere verschillen te zien. Uit de resultaten van deze studie werd geconcludeerd dat de validiteit van de beoordeling met name verbeterd zou kunnen worden met betrekking tot het proces van combineren van data uit verschillende bronnen tot een totaaloordeel. Het herkennen of aanwijzen van bewijsmateriaal bleek geen specifieke problemen op te leveren: de beoordelaars in de onderhavige studie wezen grotendeels hetzelfde bewijs aan voor de verschillende aspecten van interpersoonlijke competentie en voor de verschillende databronnen.

Hoewel de resultaten van Studie I gezien kunnen worden als een indicator voor de validiteit van de beoordelingsprocessen van beoordelaars, bieden ze geen inzicht in de specifieke moeilijkheden van beoordelaars en daarmee geen aanknopingspunten voor de wijze waarop de validiteit van beoordelingsprocessen verder aangescherpt kan worden. In **hoofdstuk 3** (Studie II) worden de beoordelingsprocessen van vierentwintig beoordelaars onderworpen aan een meer diepgaande analyse. De doelstelling van dit onderzoek was om de aard van beoordelingsprocessen van beoordelaars te beschrijven en, meer specifiek, om na te gaan of beoordelaars de twee belangrijke beoordelingsprincipes weten te realiseren zoals uiteengezet door Moss et al. (1998): (a) het zoeken van coherentie tussen de verschillende databronnen en nagaan of al het

relevante bewijs in overweging is genomen, en (b) het uitdagen van de zich ontwikkelende totaalindruk door het actief op zoek gaan naar tegenbewijs of alternatieve interpretaties. Verondersteld wordt dat toepassing van deze principes door beoordelaars de kans op de introductie van construct-irrelevante variantie en construct onderrepresentatie verkleint. Tevens wordt verondersteld dat de toepassing van deze twee principes in het bijzonder gestimuleerd wordt door discussie tussen beoordelaars. Om deze reden werden de beoordelingsprocessen van beoordelaars geanalyseerd die paarsgewijs de interpersoonlijke competentie van een docent-in-opleiding beoordeelden. De centrale onderzoeksvraag was: Wat is de aard en kwaliteit van beoordelingsprocessen van beoordelaars die paarsgewijs een docent beoordelen? En meer specifiek: Wat zijn de exacte collaboratieve activiteiten die beoordelaars uitvoeren bij de beoordeling van docentcompetenties? In welke mate representeren deze activiteiten de twee bovengenoemde beoordelingsprincipes? Welke verschillen tussen de verschillende beoordelingsparen zijn in dit verband te onderscheiden? Uit de resultaten van de studie bleek dat beoordelaars duidelijk activiteiten uitvoeren conform de genoemde beoordelingsprincipes (a) het zoeken naar coherentie tussen de verschillende databronnen en nagaan of al het relevante bewijs in overweging is genomen, en (b) het uitdagen van de zich ontwikkelende totaalindruk door actief op zoek te gaan naar tegenbewijs of alternatieve interpretaties. De bestudering van het overzicht van de activiteiten in paren liet zien dat deze twee beoordelingsprincipes het best worden gerepresenteerd door de activiteiten 'aanvullend bewijsmateriaal aandragen' (*aanvullen*) en 'uitdagen van een interpretatie door het aandragen van tegenbewijs of het presenteren van een alternatieve interpretatie' (*uitdagen*). Het blijkt echter moeilijk voor beoordelaars om elkaar in de discussie zowel aan te vullen als uit te dagen. Slechts een klein aantal paren liet een gezamenlijk beoordelingsproces zien waarin zij elkaar zowel *aanvulden* als *uitdaagden*. De discussie van de overige paren was of voornamelijk gericht op het elkaar *uitdagen*, voornamelijk gericht op *aanvullen*, of was gericht op *aanvullen* noch *uitdagen*. Hoewel verondersteld wordt dat een samenwerking die gericht is op zowel aanvullen als uitdagen leidt tot de meeste valide oordelen (d.w.z., oordelen gebaseerd op *al* het relevante bewijsmateriaal, zowel bewijs als 'tegenbewijs'), bleek dat zelfs de processen van de beoordelaars die erin slaagden om elkaar zowel aan te vullen als uit te dagen verbeterd zouden kunnen worden. Discussies van de beoordelaars leidden niet altijd tot duidelijke conclusies, waardoor het aangedragen tegenbewijs en alternatieve interpretaties vaak niet werden meegenomen in het totaaloordeel. Dit betekent dat zelfs beoordelaars die *wel* tegenbewijs en alternatieve interpretaties aandroegen de neiging hadden om uiteindelijk vooral bevestiging te zoeken voor hun aanvankelijke oordeel. Op basis van de

resultaten van de onderhavige studie werd geconcludeerd dat een beoordelingsproces waarin de zich ontwikkelende totaalindruk actief wordt uitgedaagd met tegenbewijs en/of alternatieve interpretaties van bewijs, moeilijk te realiseren is in de praktijk. Omdat een dergelijk proces al moeilijk te realiseren blijkt voor duo's, veronderstellen we dat dit nog moeilijker te realiseren zal zijn voor individuele beoordelaars.

In **hoofdstuk 4** (Studie III) wordt de validiteit van beoordelingen van docentcompetentie bekeken vanuit het perspectief van beoordelaars. We onderzochten de percepties van 22 beoordelaars met betrekking tot de beoordelingsprocessen van het aanwijzen van bewijsmateriaal van afzonderlijke bronnen en het combineren van dit bewijsmateriaal tot een totaaloordeel op basis van retrospectieve interviews. De centrale onderzoeksvraag luidde: Welke strategieën om tot een valide oordeel te komen worden ingezet door beoordelaars in relatie tot de essentiële processen van aanwijzen van bewijsmateriaal uit de afzonderlijke bronnen en combineren van dit bewijsmateriaal tot een totaaloordeel? De percepties van beoordelaars met betrekking tot het aanwijzen en combineren van bewijsmateriaal werden gecategoriseerd in termen van de validiteitscriteria 'geen construct-irrelevante variantie' en 'geen construct onderrepresentatie'. Indicatoren voor deze criteria werden gedistilleerd uit de literatuur. Voor het aanwijzen van bewijsmateriaal werden de volgende indicatoren onderscheiden: (a) het aangewezen bewijsmateriaal is relevant voor de te beoordelen competentie: de beoordelaar voegt hier geen extra, irrelevante criteria aan toe, (b) alle relevante competenties en aspecten hiervan worden gedekt, al het relevante bewijsmateriaal wordt in overweging genomen. Voor het combineren van bewijsmateriaal tot een totaaloordeel waren de indicatoren als volgt: (a) al het bewijsmateriaal aangewezen voor de afzonderlijke bronnen komt duidelijk tot uitdrukking in het eindoordeel, (b) zowel bevestigend bewijs als tegenbewijs wordt in overweging genomen in het totaaloordeel, (c) er is geen sprake van idiosyncratische weging van bewijsmateriaal, (d) alle conclusies worden duidelijk onderbouwd op basis van de originele data. Percepties van beoordelaars van het aanwijzen en combineren van bewijsmateriaal werden gecategoriseerd aan de hand van deze indicatoren. Op deze manier stelden we een overzicht op van succesvolle strategieën en problemen met betrekking tot het aanwijzen van bewijsmateriaal uit de afzonderlijke bronnen en het combineren van dit bewijsmateriaal tot een totaaloordeel.

De resultaten laten zien dat beoordelaars een grote diversiteit aan *succesvolle strategieën* gebruikten. De afzonderlijke strategieën werden echter door slechts een klein aantal beoordelaars genoemd. Opmerkelijk was dat beoordelaars meer succesvolle strategieën noemden met

betrekking tot het aanwijzen van bewijsmateriaal dan tot het combineren van bewijsmateriaal. Waar beoordelaars tijdens het proces van aanwijzen van bewijsmateriaal sterk gericht waren op het in overweging nemen van *al* het relevante bewijsmateriaal, richtten zij zich bij het combineren van bewijsmateriaal niet expliciet op de vraag of *al* het aangewezen bewijs ook daadwerkelijk tot uitdrukking kwam in het uiteindelijke oordeel. Verder bevestigen de resultaten van deze studie die van de voorgaande studie, waarin we constateerden dat het voor beoordelaars moeilijk is om hun aanvankelijke oordeel uit te dagen met tegenbewijs of alternatieve interpretaties. Bestudering van de door beoordelaars genoemde strategieën laat zien dat beoordelaars de discussie met een andere beoordelaar voornamelijk gebruiken om bevestiging te zoeken van hun aanvankelijke, individuele oordeel: beoordelaars verwezen veel vaker naar de bevestigende functie dan naar de uitdagende functie van de andere beoordelaar. De beoordelaars verwezen niet alleen naar strategieën, maar waren zich ook bewust van problemen die mogelijk nadelig waren voor de kwaliteit van hun oordeel. Het aantal beoordelaars dat zich bewust was van potentiële 'bedreigingen van de validiteit' was echter klein. En hoewel beoordelaars voor een aantal problemen een oplossing vonden (een succesvolle strategie introduceerden), bleef een aantal problemen onopgelost. Uit de resultaten van deze studie werd geconcludeerd dat de validiteit van de beoordeling zou kunnen worden verbeterd als beoordelaars in staat zouden zijn een breed repertoire van strategieën in te zetten dat samenhangt met een valide beoordeling. Als beoordelaars bedreigingen van validiteit kunnen herkennen en zich bewust zijn van potentiële strategieën om deze bedreigingen te voorkomen, zou dit de kans kunnen vergroten dat zij succesvolle strategieën bewust inzetten tijdens de beoordeling.

In **Hoofdstuk 5** worden de belangrijkste conclusies van het proefschrift samengevat en bediscussieerd. De resultaten van het onderzoek geven niet alleen een illustratie van de aard van beoordelingsprocessen van beoordelaars, maar geven ook een indicatie van de problemen die spelen bij deze processen. Van de essentiële beoordelingsprocessen 'aanwijzen van bewijsmateriaal' en 'combineren van bewijsmateriaal tot een totaaloordeel' blijkt het laatste het lastigst en minst doorzichtig voor beoordelaars. De resultaten van Studie I lieten zien dat de verschillen tussen beoordelaars hoofdzakelijk ontstonden in het proces van combineren van bewijsmateriaal. Studie II en III wezen uit dat de beoordelaars in hun beoordelingsproces relatief veel nadruk leggen op het aanwijzen van bewijsmateriaal. De resultaten van Studie II lieten verder zien dat de oordelen — in de vorm van evaluerende samenvattingen — van vijf van de twaalf beoordelaars het niveau van concreet bewijsmateriaal niet ontstegen: in hun

gezamenlijke beoordelingsproces somden deze beoordelaars hoofdzakelijk het relevante bewijsmateriaal op, zonder dit bewijs duidelijk te interpreteren teneinde tot een holistisch oordeel te komen. Daarnaast lieten de resultaten van Studie III zien dat beoordelaars meer strategieën noemden met betrekking tot het aanwijzen van bewijsmateriaal dan met betrekking tot het combineren van dit bewijsmateriaal tot een totaaloordeel. Bij het combineren van bewijsmateriaal zouden beoordelaars ervoor moeten zorgen dat al het relevante bewijsmateriaal in het eindoordeel tot uitdrukking komt, dat niet alleen gezocht wordt naar bevestigend bewijs voor hun aanvankelijke indruk maar dat ook tegenbewijs gezocht en verklaard wordt, en dat alle conclusies duidelijk worden onderbouwd op basis van de originele data. Alle beoordelaars hadden de neiging om vooral bevestigend bewijsmateriaal voor hun zich ontwikkelende indruk te zoeken en niet ook actief op zoek te gaan naar tegenbewijs voor deze indruk, of naar mogelijke alternatieve interpretaties van bewijsmateriaal. Uit de resultaten van Studie II blijkt dat zelfs als beoordelaars daadwerkelijk tegenbewijs en/of alternatieve interpretaties bediscussiëren, dit niet altijd resulteert in duidelijke conclusies. Dit tegenbewijs en deze alternatieve interpretaties komen niet duidelijk terug in de argumentatie van beoordelaars en daarmee ook niet in het uiteindelijke oordeel. Uit de resultaten van Studie III blijkt dat vrijwel alle beoordelaars de discussie met de andere beoordelaar hoofdzakelijk gebruiken om bevestiging te zoeken voor hun aanvankelijke, individuele oordeel en niet gericht zijn op het uitdagen van elkaars argumentatie en oordelen.

In dit proefschrift werd gesuggereerd dat het belangrijk is dat beoordelaars zich bewust zijn van hun eigen beoordelingsprocessen en de specifieke bedreigingen van validiteit in deze processen. Een dergelijk bewustzijn maakt het mogelijk om acties te ondernemen om potentiële bedreigingen van validiteit te voorkomen en tot een valide oordeel te komen. Met betrekking tot het combineren van bewijsmateriaal tot een totaaloordeel is het belangrijk dat beoordelaars in staat zijn om de argumentatie die ten grondslag ligt aan hun totaaloordeel te verduidelijken. In staat zijn om de lijn van argumentatie te specificeren maakt het mogelijk om te evalueren in hoeverre men als beoordelaar voldoet aan de indicatoren van validiteit zoals uiteengezet in Studie III (al het bewijsmateriaal aangewezen voor de afzonderlijke bronnen komt duidelijk tot uitdrukking in het eindoordeel; zowel bevestigend bewijs als tegenbewijs worden in overweging genomen in het totaaloordeel; er is geen sprake van idiosyncratische weging van bewijsmateriaal; alle conclusies worden duidelijk onderbouwd op basis van de originele data).

De beperkingen van het onderzoek hangen samen met de specifieke context waarbinnen de drie studies werden uitgevoerd. We analyseerden de beoordelingsprocessen van beoordelaars die de interpersoonlijke competentie van een docent-in-opleiding beoordeelden op basis van een specifieke beoordelingsprocedure. Het beoordelingskader dat ten grondslag lag aan deze beoordelingsprocedure was gebaseerd op een zorgvuldig gedefinieerd construct. Niet voor alle relevante docentcompetenties is een dusdanig goed geconceptualiseerd en geoperationaliseerd construct beschikbaar als voor 'interpersoonlijke competentie'. De resultaten van de huidige studie kunnen daarom niet zonder meer geoperationaliseerd worden naar andere beoordelingsprocedures. Beoordelingen gebaseerd op minder goed gedefinieerde constructen leveren mogelijk extra vraagstukken op met betrekking tot beoordelingsprocessen. Beperkingen van dit onderzoek hangen samen met het feit dat de beoordelaars in deze studie de beoordelingsprocedure voor de eerste keer gebruikten, maar één docent-in-opleiding beoordeelden en de percepties van beoordelaars met betrekking tot succesvolle strategieën en bedreigingen werden onderzocht op basis van semi-gestructureerde interviews. Mogelijk zou een grotere variatie in de aard van de beoordelingsprocessen kunnen worden gevonden en zou meer inzicht in 'strategieën' en 'bedreigingen' kunnen worden verkregen door beoordelaars in te schakelen die meer ervaring hebben opgedaan met de huidige beoordelingsprocedure, hun meer docenten-in-opleiding te laten beoordelen en door naast interviews aanvullende methoden te gebruiken om hun percepties te onderzoeken.

In vervolgonderzoek naar de kwaliteit van beoordelingsprocessen van beoordelaars zou gekeken kunnen worden naar de specifieke condities die beoordelaars kunnen stimuleren om de relevante beoordelingsprocessen zoals uiteengezet in deze studie in de praktijk te brengen. Een ander belangrijk thema voor vervolgonderzoek betreft de effecten van competentiebeoordelingen op de professionele ontwikkeling van docenten.

De resultaten van dit onderzoek hebben een aantal implicaties voor de beoordelingspraktijk. De inzichten die dit onderzoek heeft opgeleverd, kunnen een rol spelen bij de training van beoordelaars. Trainingen van beoordelaars zouden expliciet gericht moeten zijn op het vergroten van het bewustzijn van een aantal specifieke deelprocessen in de beoordeling, zoals die geanalyseerd zijn in het onderhavige onderzoek. Het zou een kernpunt in de voorbereiding van beoordelaars op hun taak moeten zijn om beoordelaars een aantal authentieke casussen te laten beoordelen en op basis hiervan de aard van beoordelingsprocessen te bediscussiëren evenals de specifieke manieren waarop beoordelaars geprobeerd hebben om te voldoen aan de relevante indicatoren van validiteit. Het overzicht van succesvolle strategieën die beoordelaars

ondernemen tijdens de beoordeling van docent-competentie (Studie III) kunnen zeker gebruikt worden als onderdeel van de voorbereiding van beoordelaars. Zulke strategieën kunnen worden bediscussieerd in directe relatie tot mogelijke bedreigingen van de validiteit van de beoordeling. Beginnende beoordelaars blijken zich weinig bewust van mogelijke bedreigingen tijdens het proces, en wanneer zij zich er wel bewust van zijn weten ze deze problemen niet altijd op te lossen. Als discussie met een andere beoordelaar onderdeel is van de beoordelingsprocedure kunnen de vier typen collaboratieve beoordelingsprocessen (de verschillende combinaties van aanvullende en uitdagende activiteiten) en hun sterke en zwakke punten geïntroduceerd worden in trainingsprogramma's. Dit maakt de discussie over een valide beoordeling en het voorkomen van construct-irrelevante variantie en construct onderrepresentatie concreter.

Summary

The research presented in this dissertation concerns validity in teacher assessment. In connection with the growing attention for the quality and professionalization of teachers, assessment of teaching is currently becoming increasingly more common practice. Teaching competences which are required for effective performance in specific teaching contexts are being specified, and procedures to assess the extent to which teachers master these competences are being developed and used. Both the specification of teaching competences and the assessment of teaching are believed to support the professional development of teachers, as they encourage reflection and meaningful dialogue among practitioners with regard to the essence of good teaching. High-quality teacher assessment which also encourages teacher learning should clearly reflect that teaching is complex, context specific, and person-bound. In order to do justice to these characteristics of teaching, teacher assessment is generally based on multiple sources of evidence (e.g., lesson plans, video-taped lessons, written reflections of the teacher on his or her performance).

In the assessment of teachers based on data from different sources, assessors play a crucial role. Even though the assessment procedure used and the criteria employed are important conditions for a valid judgement, the overall quality of the assessment depends heavily on the ability of assessors to use these procedures and criteria to reach a valid judgement. When assessment is based on several types of data, the assessment process is often thought of in terms of two essential tasks: (a) consideration of evidence from separate sources and (b) combination of evidence from different sources to attain an overall judgement. As these tasks require considerable judgement on the part of the assessors, the overall validity of the assessment rests heavily on the quality of assessors' judgement processes.

Two important threats to the validity of judgement processes are construct-irrelevant variance and construct under-representation (Messick, 1989, 1994). *Construct-irrelevant variance* may arise when assessors apply extraneous, irrelevant, or idiosyncratic criteria which are not part of the assessment framework in which they have been trained, and when assessors consider evidence which is simply irrelevant to the construct in question. *Construct under-representation* is characterized by the omission of particular assessment criteria, or evidence, or the

idiosyncratic weighing of certain criteria or evidence which may result in insufficient attention to particular aspects of performance.

In the context of teacher portfolio assessment, the results of studies by Moss et al. (1998) and Schutz and Moss (2004) have shown that assessors cannot avoid forming a pattern out of the data from a teacher portfolio in order to evaluate it. Once a particular pattern has been identified, subsequent information may also be interpreted in terms of this pattern. This strong tendency to seek confirmation of one's initial interpretations can be seen to constitute a bias, which can most certainly affect the validity of the judgement process. This has led Moss et al. (1998) to introduce two interrelated principles underlying a valid assessment process. The first principle is that assessors should search for coherence among the evidence coming from different sources, and explicitly check that all relevant evidence has been taken into account. The second principle is that assessors should be encouraged to actively seek and consider counterevidence, and thus evidence which challenges their developing interpretations, and explicitly entertain alternative interpretations. Compliance with these two principles is assumed to minimize the aforementioned threats to validity, namely, the introduction of construct-irrelevant variance or construct under-representation, or both. It is asserted that discussion between pairs of assessors after they have formed an individual initial judgement may enhance the chances of adherence to the two principles.

Although the validity of teacher assessment rests heavily on the quality of the judgement processes of assessors, few studies have addressed the nature or quality of these judgement processes. Such research could provide greater insight into specific validity issues in teacher assessment and into the specific ways in which assessors can comply with relevant validity criteria (i.e., principles underlying a valid assessment process).

The question that prompted the current research was, *How can validity of teacher assessment be ensured in practice?* The following, more specific, questions were addressed which clearly focused on the nature of the judgement processes of assessors:

(1) To what extent do construct-irrelevant variance and construct under-representation occur in the judgement processes of assessors in the consideration of evidence from separate sources and combination of evidence to attain an overall judgement?

(2) What is the nature of the assessment process when teachers are assessed collaboratively?

(3) What strategies to ensure the validity of the assessment process are performed by assessors in connection with the essential judgement processes of (a) consideration of evidence and (b) combination of evidence to attain an overall judgement?

The research questions were addressed in a series of three studies. In all three studies, we analyzed the judgement processes of assessors who evaluated the interpersonal competence of a student teacher using a specific procedure developed by the author. As interpersonal teacher competence, or the quality of the interpersonal relationships with students, is a major concern for many beginning teachers, this particular competence was selected to be assessed in the present research. The assessment procedure included the following three data sources: (1) the Questionnaire on Teacher Interaction (QTI), which provides a reliable and valid picture of both student and teacher perspectives on the interpersonal relationships between the teacher and his or her pupils, (2) a videotaped lesson selected by the student teacher, and (3) a self-evaluation assignment completed by the student teacher. To help the assessors attain a comprehensive picture of the teacher's interpersonal competence, an assessment framework in which six key components of interpersonal teacher competence were identified was provided.

The results of all three studies are based on the same sample of assessors judging the same student teacher. Twenty-four individuals who were at the time teacher educators (N=4) or teachers supervising the training of prospective secondary school teachers (N=20) were selected to participate as assessors in the present research. All of the individuals had experience with the evaluation of student teachers or the provision of feedback on the performance of student teachers. Prior to their participation in the present study, they all followed a two-evening training programme. During the first evening, the procedure to assess the interpersonal competence of student teachers was introduced and the assessment framework was discussed. During the second evening, the assessors conducted an entire assessment using an authentic case (i.e., data provided by a student teacher). Following completion of the assessment task, the experiences of the assessors were exchanged in a plenary discussion; specific attention was paid to the difficulties encountered.

In **Chapter 2** (Study I) we reported on a small-scale study in which the procedure to assess the interpersonal competence of student teachers was developed and validated. The validity of the

judgements of four assessors who evaluated the interpersonal competence of the same student teacher was explored using the concepts of construct-irrelevant variance and construct under-representation. The research question of this first study was the following: To what extent do construct-irrelevant variance and construct under-representation occur in the judgement processes of assessors in the consideration of evidence from separate sources and combination of evidence to attain an overall judgement? To investigate the amount of construct-irrelevant variance and construct under-representation in the consideration and combination of evidence, assessors' *notes* (i.e., evidence noted for the three separate data sources) and overall judgements in the form of *statements* in an evaluative summary were compared to an expert judgement. The results showed the amount of construct-irrelevant variance and construct under-representation in these notes and statements to be small. That is, in both their notes and their statements, the assessors hardly referred to evidence which was irrelevant to interpersonal competence. In addition, they included all relevant aspects of interpersonal competence as operationalized in the assessment framework underlying the procedure. Small differences between the four assessors were found to stem primarily from the process of combining evidence from separate sources to attain an overall judgement. That is, the four assessors noted more or less the same evidence for the separate sources, but relatively greater variance was found in their statements. On the basis of these results it was concluded that the validity of the assessment of interpersonal competence — using the present assessment procedure — could be improved specifically with regard to the judgement process of the combination of evidence from separate sources into an overall judgement. The consideration or recognition of the relevant data did not pose specific problems, as the assessors in the present study noted more or less the same data for the different components of interpersonal competence and for the different sources of data.

Although the results with regard to construct-representation in the notes and statements of assessors presented in Study I may be regarded as indicators of the validity of the judgement processes of these assessors, they do not provide practical guidance for the specific ways in which the quality of the relevant judgement procedures can be ensured or improved. In **Chapter 3** (Study II), the judgement processes of twenty-four assessors were analyzed in greater detail. We aimed at answering the following research question: What is the nature of the assessment process when teachers are assessed collaboratively? We particularly aimed to examine the extent to which these processes reflected the two principles of assessment

mentioned by Moss et al. (1998): (a) seeking coherence among the available sources of evidence and checking whether all relevant evidence has been taken into account, and (b) challenging developing interpretations with counterevidence and explicitly considering alternative interpretations. Engagement in these two 'key assessment processes' is assumed to minimize the introduction of construct-irrelevant variance and construct under-representation. As it is assumed that discussion between assessors may enhance the chances of engagement in these processes, we examined the processes which assessors engaged in when *collaboratively* (i.e., pairwise) assessing the interpersonal competence of a student teacher. The research question concerned the nature of the assessment process when teachers are assessed collaboratively. The following more specific questions were addressed: (a) Which communicative activities do assessors engage in while collaboratively assessing teacher competence? (b) To what extent are the aforementioned key processes represented in the observed collaborative judgement processes? and (c) To what extent does the representation of the key processes vary across the different pairs of assessors? The results of the study showed that the assessors clearly undertook communicative activities which reflected the key assessment processes of (a) seeking coherence among available sources of evidence and checking whether all of the relevant evidence has been taken into account and (b) challenging developing interpretations with counterevidence or perspectives of the other assessor. Inspection of the overview of communicative activities showed the two key assessment processes to be reflected particularly in the activities of 'providing additional evidence' (*addition*) and 'challenging an interpretation via presentation of counterevidence or alternative interpretation' (*challenging*). It seems, nevertheless, highly difficult for assessors to engage in *both* key processes. That is, the collaborative judgement processes of only few collaborative pairs were found to involve both *addition* and *challenging*. The discussions of the other pairs were characterized by either a predominantly *challenging* pattern of collaborative assessment, a predominantly *addition* pattern of collaborative assessment, or neither *addition* nor *challenging*. Although judgement processes involving both *addition* and *challenging* may be assumed to produce the most valid judgements, the results showed that even the quality of the collaborative assessment process for the few pairs who were able to realize such a process could be further improved. The discussion activities did not always produce clear conclusions, and some of the discussion activities were even ignored at times, meaning that the counterevidence presented and the alternative interpretations put forth were not always reflected in the final judgements of these pairs. Even collaborative pairs who were engaged in both key assessment processes eventually showed a disposition to consider

only confirmatory evidence as opposed to actively entertaining counterevidence or alternative interpretations. Based on these results, it was concluded that a judgement process in which initial interpretations are actively challenged with counterevidence or alternative interpretations, which is assumed to produce the most valid judgements, is difficult to realize. As such a process proved difficult to realize for collaborative pairs, we believe that taking all of the relevant evidence into account may be even harder for individual assessors.

Chapter 4 (Study III) addresses the occurrence of construct-irrelevant variance and construct under-representation from the perspective of the assessors. The experiences of assessors (N=22) with the judgement processes of consideration of evidence from separate sources were investigated using semi-structured interviews. The following research question was addressed: Which strategies to ensure the validity of the assessment process are performed by assessors in connection with the essential judgement processes of consideration of evidence from separate sources and combination of evidence to attain an overall judgement? For each judgement process, perceptions of the assessors were categorized in terms of the validity criteria of (a) no construct-irrelevant variance and (b) no construct under-representation and their indicators. For the consideration of evidence from separate sources, the following indicators were distinguished: (a) evidence refers only to the relevant competences: no extraneous, irrelevant criteria are added, (b) all of the relevant competences and components of these are covered; all relevant evidence is considered. For the combination of evidence from different sources to attain an overall judgement, the indicators were the following: (a) all of the evidence noted for the separate sources is also clearly reflected in the overall judgement, (b) both confirmative evidence and counterevidence are considered in the overall judgement, (c) no idiosyncratic weighing of evidence occurs, and (d) all conclusions are supported by clear argumentation which also draws upon the original data. Perceptions of assessors with regard to the consideration and the combination of evidence were categorized according to the afore-mentioned indicators. In such a manner, an overview was provided of 'successful strategies' and 'problems' for the consideration of evidence from separate sources and the combinations of evidence into an overall judgement.

The overview of strategies presented shows that the assessors reported a great diversity of successful strategies pertaining to the validity indicators distinguished. However, the different strategies were generally mentioned only by small numbers of assessors. The results show that more strategies were mentioned for the process of consideration of evidence

from separate sources than for the process of combining this evidence in an overall judgement. In the process of consideration of evidence from separate sources, nearly all assessors used strategies aimed at the consideration of all relevant evidence; in the process of combination of evidence to attain an overall judgement, few assessors explicitly aimed at also systematically reflecting all relevant evidence in the overall judgement. From these results it may be concluded that assessors show more awareness of construct under-representation with regard to consideration of evidence than with regard to the process of combination of evidence to attain an overall judgement.

The assessors in the present study not only described successful strategies underlying a valid assessment process, but also described several problems encountered during the assessment process. These problems clearly reflect *threats* to the validity of the assessment. The numbers of assessors who mentioned these threats were, however, small. For a number of threats mentioned, a successful strategy to overcome these specific threats was introduced; however, the greater part of the threats mentioned remained unresolved. (i.e., no successful strategies for overcoming these threats were employed by the assessors). On the basis of the present results it was concluded that the validity of the assessment can be improved when assessors are able to employ a broad repertoire of strategies which underlie a valid assessment. When assessors are able to recognize threats to the validity of their judgement processes and, in addition, are aware of a range of potential successful strategies to overcome such threats, they might be more likely to use these strategies during the assessment process, which may result in a more valid assessment.

In **Chapter 5** we summarize the main conclusions of the three studies. We discuss the nature of the judgement processes of assessors and the specific difficulties in these processes which may affect the validity of the overall judgement. The results show the essential assessment process of combination of evidence to attain an overall judgement to be more challenging for assessors than the process of consideration of evidence from separate sources. The results of Study I show the differences between assessors to stem primarily from the combination of evidence. The results of Study II and Study III indicate that the assessors in the present study put relatively much emphasis on the consideration of evidence. The results of Study II show that the judgements — in the form of evaluative summaries — of five of the twelve collaborative pairs did not truly go beyond the level of specific data. That is, in their collaborative judgement processes, these assessors took the approach of summarizing the relevant data, and

did not clearly interpret these data so as to come up with a holistic judgement of the data. In addition, the results of Study III show that the assessors considered more strategies pertaining to the consideration of evidence than to the combination of evidence in an overall judgement. With regard to this latter process, assessors should aim to reflect all relevant evidence in the final judgement; they should not only reflect evidence which clearly fits initial judgements, but should also involve and explain counterevidence and support the final judgement by clear argumentation representing all relevant evidence. All assessors in the present study, however, showed a tendency to consider only confirmatory evidence for their initial interpretations, as opposed to also considering counterevidence or alternative interpretations. The results of Study II show that even when assessor pairs *did* discuss counterevidence or alternative interpretations, or both, this did not always produce clear conclusions, indicating that counterevidence found and alternative interpretations put forth were not reflected in the argumentation of assessors and, therefore, in the final judgement. The results of Study III show that a great many assessors used the discussion with another assessor primarily to seek confirmation of their initial judgement, and did not aim at challenging each other's or their own judgements.

In our study, we suggested that it is important for assessors to be aware of their own judgement processes and the specific threats to validity in these processes. Such awareness is likely to enable assessors to take conscious actions during the assessment process to overcome such threats to validity and, in this connection, to reach a sound judgement. With respect to the process of combination of evidence to attain a sound overall judgement, awareness of the process particularly involves being able to elicit the line of argumentation underlying the overall judgement. Being able to elicit the line of argumentation makes it possible to evaluate the extent to which assessors comply with the validity indicators summarized in Study III (i.e., all relevant evidence is clearly reflected in the overall judgement; both confirmative and counterevidence are considered; no idiosyncratic weighing of evidence occurs; and all conclusions are supported by clear argumentation which also draws upon the original data).

The limitations of the research pertain to the specific assessment context in which the three studies were conducted. We analyzed the judgement processes of assessors who judged the interpersonal competence of a student teacher using a specific assessment procedure. The assessment framework central to this procedure was based on a well-defined construct. Given that such a well-conceptualized and well-operationalized construct is not available for all relevant teacher competences as was the case for 'interpersonal competence', the present findings cannot necessarily be generalized to other assessment procedures. Assessment based

on less well-defined constructs might bring about additional issues in the judgement processes of assessors. Limitations to the present research are that the assessors in the present study were using the assessment procedure for the first time, the assessors only assessed one student teacher, and perceptions of assessors with regard to successful strategies and threats underlying a valid assessment were examined using semi-structured interviews. The analysis of the judgement processes of assessors more experienced with the assessment procedure, the assessment of more (student) teachers, and the use of additional methods to examine assessors' activities and perceptions, might elicit greater variance in the nature of assessors' judgement processes and might further elicit the successful strategies and threats underlying valid assessment.

In future research into the quality of assessors' judgement processes, the specific conditions which enable assessors to engage in the relevant judgement processes can productively be investigated. Another essential theme for further research is the consequences of assessment for the professional development of teachers.

Based on the insights generated by the present research, practical implications are described for the design of assessment procedures and practices, and for assessor preparation programmes. Assessor training should explicitly be aimed at increasing awareness of relevant judgement processes, as analysed in the present research. A vital part of assessor preparation programmes would be to have assessors judge a number of authentic cases and to discuss the nature of their judgement processes and the specific ways in which assessors tried to comply with relevant indicators of validity. The overview of successful strategies undertaken during the assessment of teacher competence (Study III) can certainly be considered in assessor preparation programmes. Such strategies could be discussed in explicit relation to potential threats to validity. When discussion with another assessor is a part of the assessment procedure, the overview of the communicative activities undertaken during the collaborative assessment of teacher competence (Study II) can also be introduced into teacher assessment preparation programmes. The four types of collaborative assessment (the various combinations of addition and challenging activities or the lack of these), together with their relative strengths and weaknesses, should be considered in such a programme. This would allow the discussion of such validity-related issues as construct-irrelevant variance and construct under-representation to be made more tangible.

Publications

Scientifc publications

Nijveldt, M.J., Beijaard, D., Brekelmans, J.M.G., Verloop, N., & Wubbels, Th. (2006). Assessing the interpersonal competence of beginning teachers: the quality of assessors' judgement processes. *International Journal of Educational Research, 43*, 89-102.

Manuscripts submitted for publication

Nijveldt, M.J., Brekelmans, J.M.G., Wubbels, Th., Verloop, N., & Beijaard, D.(submitted). The nature of the collaborative judgement process in the assessment of student teachers.

Nijveldt, M.J., Beijaard, D., Brekelmans, J.M.G., Wubbels, Th., & Verloop, N. (resubmitted to Studies in Educational Evaluation). Assessment of teacher competence: Assessors' perceptions of their judgement processes.

Presentations

Nijveldt, M.J., Beijaard, D., Brekelmans, J.M.G., Verloop, N., & Wubbels, Th. (2006). *Beoordelen van interpersoonlijke docentcompetentie: de aard en kwaliteit van beoordelingsprocessen*. Paper presented at a symposium at the Onderwijs Research Dagen, Amsterdam.

Nijveldt, M.J., Beijaard, D., Brekelmans, J.M.G., Verloop, N., & Wubbels, Th. (2006). *Assessing beginning teachers' interpersonal competence: The quality of assessors' judgement processes*. Paper presented in a symposium conducted at the annual meeting of the AERA, San Francisco, USA.

Nijveldt, M.J., Beijaard, D., Brekelmans, J.M.G., Verloop, N., & Wubbels, Th. (2004). *Designing and developing an assessment procedure on beginning teachers' interpersonal competence*. Paper presented in a symposium at the Northumbria / EARLI SIG Assessment Conference, Bergen, Norway.

Nijveldt, M.J., Beijaard, D., Brekelmans, J.M.G., Verloop, N., & Wubbels, Th. (2004). *Kwaliteitsvraagstukken in de ontwikkeling van een procedure voor het beoordelen van interpersoonlijke docentcompetentie*. Paper presented at Onderwijs Research Dagen, Utrecht.

Nijveldt, M.J., Beijaard, D., Brekelmans, J.M.G, Verloop, N., & Wubbels, Th. (2003). Validity and reliability in a teacher assessment procedure: Optimising assessors' decision-making. Poster presented at the ISATT Biannual Conference, Leiden

Curriculum Vitae

Mirjam Nijveldt was born on the 26th of October 1979, in Gouda, the Netherlands. She attended secondary education at the Groene Hart Lyceum in Alphen aan den Rijn, were she graduated in 1998. From 1998 to 2002 she studied Educational Science at Utrecht University. Her master's thesis concerned the design and evaluation of a self-assessment procedure for teaching competences in higher education. After her graduation, she started as a PhD student at ICLON Graduate School of Teaching at Leiden University. Her research focussed on validity in the assessment of student teachers. She followed master classes on teaching and teacher eduaction, qualitative analysis and new modes of assessment. Mirjam is currently employed at the ILS Graduate School of Education at Radboud University in Nijmegen.

Dankwoord

Verschillende mensen hebben in verschillende fasen een belangrijke bijdrage geleverd aan de totstandkoming van dit proefschrift.

Ten eerste wil ik de lerarenopleiders en begeleiders-op-school noemen die betrokken waren bij de ontwikkeling van de beoordelingsprocedure. Marijke Hermans, André Sistermans, Peter Kop, Doesjka Nijdeken, Frans Teitler, Mirjam Meinema en Esther de Vrind: ik heb veel van jullie geleerd en ben jullie ontzettend dankbaar voor jullie inzet. Ook de IPP-groep van het IVLOS was in deze fase belangrijk voor mij: hartelijk dank voor jullie hulp!

Een groot aantal opleiders en begeleiders was vervolgens bereid om getraind te worden als beoordelaar. Charlotte van der Ham, Inger Klijnhout, Rob Anders, Jeroen Verweij, Cris Bertona, Toos van der Valk, Thea van Bokhoven, Hans Vosse, Tineke Harmzen, Astrid Buys, Hans Hulshof, Stephan van de Laak, Wouter de Lange, Daphne Ouwerkerk, Henk van Rhijn, Bernard Schut, Elise Storck, Ted Vernooij, Arnout Vos, Leo Zwart en Pieter Manders: jullie inspanningen hebben prachtige onderzoeksdata opgeleverd. In dit opzicht wil ik ook de docenten-in-opleiding bedanken die bereid waren om als 'case' te fungeren.

Collega's in het land, mede-promovendi, en in het bijzonder mijn collega's van het ICLON zijn in alle fasen van mijn promotietraject op vele manieren onmisbaar geweest. Heel veel dank voor jullie betrokkenheid, voor jullie feedback op stukken in allerlei stadia van wording en voor jullie hulp op alle mogelijke gebieden — van meedenken over de opbouw van een presentatie tot het afnemen van interviews aan toe! Daarnaast kijk ik met heel veel plezier terug op de met jullie doorgebrachte koffiepauzes, borrels, etentjes, masterclasses, summerschool, VPO dagen en natuurlijk de nationale en internationale congressen als de ORD, ISATT, EARLI Assessment conference en AERA.

Een aantal collega's wil ik ten slotte in het bijzonder noemen. Anne Marie Uhlenbeck, Mirjam Bakker en Dineke Tigelaar, met jullie deelde ik de interesse in beoordelen. Fijn om jullie zo dicht in de buurt te hebben: ik heb heel veel aan jullie gehad! Ben Smit, dank voor de prettige ondersteuning gedurende het gehele traject en voor je fantastische hulp bij de analyses van hoofdstuk 3.

ICLON

Leiden University Graduate School of Teaching

PhD dissertation series

Hoeflaak, A. (1994). *Decoderen en interpreteren: Een onderzoek naar het gebruik van strategieën bij het beluisteren van Franse nieuwsteksten.*

Verhoeven, P. (1997). *Tekstbegrip in het onderwijs klassieke talen.*

Meijer, P.C. (1999). *Teachers' practical knowledge: Teaching reading comprehension in secondary education.*

Zanting, A. (2001). *Mining the mentor's mind: The elicitation of mentor teachers' practical knowledge by prospective teachers.*

Uhlenbeck, A.M. (2002). *The development of an assessment procedure for beginning teachers of English as a foreign language.*

Oolbekkink-Marchand, H.W. (2006). *Teachers' perspectives on self-regulated leaning: An exploratory study in secondary and university education.*

Henze, F.A. (2006). *Science teachers' knowledge development in the context of educational innovation.*

Mansvelder-Longayroux, D.D. (2006). *The learning portfolio as a tool for stimulating reflection by student teachers.*

Meirink, J.A. (2007). *Individual teacher learning in a context of collaboration in teams.*

Nijveldt, M.J. (2007). *Validity in teacher assessment: An exploration of the judgement processes of assessors.*