Assessing the interpersonal competence of beginning teachers: The quality of the judgement process

Mirjam Nijveldt[a],[*], Douwe Beijaard[b], Mieke Brekelmans[c],
Nico Verloop[a], Theo Wubbels[c]

[a]*ICLON Graduate School of Education, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands*
[b]*University of Groningen, The Netherlands*
[c]*Utrecht University, The Netherlands*

Abstract

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 beginning teacher. The validity of the assessment procedure was found to be satisfactory and the use of multiple assessment instruments to clearly have added value.
© 2006 Elsevier Ltd. All rights reserved.

Keywords: Teacher–pupil relationship; Authentic assessment; Teacher competence

1. Introduction

The quality of interpersonal relationships with pupils is a concern for many beginning teachers (e.g., Evans & Tribble, 1986; Veenman, 1984). Particularly in secondary education, beginning teachers express a strong need for professional assistance in this domain. In the present study, a procedure to assess the interpersonal competence of beginning 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

*Corresponding author. Tel: + 31 71 527 4698; fax: + 31 71 527 5342.
  E-mail address: nijveldt@iclon.leidenuniv.nl (M. Nijveldt).

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 beginning 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 analysed 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 paper, an interpersonal perspective (see Wubbels & Brekelmans, 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. 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 beginning teachers within the domain of interpersonal competence, as the starting point for the selection of the assessment instruments to be used.

The QTI (Wubbels & Brekelmans, 2006) was selected for use as it provides a reliable and valid picture of both the pupil and teacher perspectives on the *stabilized* interpersonal relationships between the teacher and his or her pupils. Comparison of the pupil and teacher perspectives also provides information on the extent to which the teacher is aware of pupil 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 & Brekelmans, 2006) was adopted to develop an assessment framework to analyse situated teacher behaviour and identify the key components of interpersonal competence (i.e., specific competences). The eight sector labels in the MITB 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
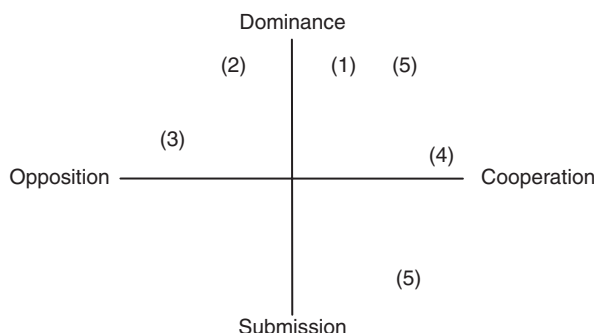
Fig. 1. The five competences positioned according to the opposition-cooperation (i.e., Proximity) and submission-dominance (i.e., Influence) dimensions of the MITB. (1) Providing guidance/clear structure; (2) setting norms and standards; (3) correcting undesirable pupil behaviour; (4) paying attention to pupils; (5) giving pupils responsibility and freedom.

our definition of teacher competence, the following competences were distinguished (with the relevant sector from the MITB noted in parentheses):

1. providing guidance/clear structure (DC leadership);
2. setting norms and standards (DO strict);
3. correcting undesirable pupil behaviour (OD admonishing);
4. paying attention to pupils (CD helpful/friendly and CS understanding);
5. giving pupils responsibility and freedom (SC student freedom).

In Fig. 1, the five competences are depicted according to the submission-dominance (i.e., Influence) and opposition-cooperation (i.e., Proximity) dimensions of the MITB (cf. Fig. 2 from Wubbels & Brekelmans, 2006). As can be seen, the sectors OD (admonishing), DO (strict), DC (leadership), CD (helpful/friendly), CS (understanding) and SC (student 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 counterexamples of the competences of (1) providing guidance/clear structure and (4) paying attention to pupils.

Given that interpersonal competence involves 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 behaviour and (b) the ability to monitor one's professional development.

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 (Frederiksen & Collins, 1989; Linn, Baker, & Dunbar, 1991; Haertel, 1990; 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 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 analysed 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.

4. Method

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.

4.2. The beginning teacher case

Using the QTI, the first author of this article selected a female beginning teacher with 9 months of experience teaching part-time and a Tolerant interpersonal profile in the opinion of her pupils (see Wubbels & Brekelmans, 2006). This teacher was judged to

resemble the average beginning teacher because most beginning teachers have been found to be of the Tolerant or Uncertain/Tolerant type.

A lesson taught by the beginning teacher was next videotaped for observation purposes (i.e., presentation to the four assessors) and self-evaluation purposes (i.e., presentation to the teacher herself). Finally, the beginning teacher was asked to complete a self-evaluation assignment (see Section 4.3.3). In such a manner, the four assessors assessed the same beginning teacher.

4.3. Assessment procedure

4.3.1. Questionnaire on teacher interaction (QTI)

The first instrument administered to assess the interpersonal competence of the beginning teacher was the QTI, which provides information on (a) pupil 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 pupil perceptions with the profile of the average beginning teacher, as provided in the QTI report. The assessors were also asked to compare the pupil perceptions of the beginning teacher with the beginning teacher's own self-perceptions in order to attain an indicator of the beginning teacher's awareness of her own interpersonal competence.

4.3.2. Observation instrument

The second instrument used to assess the interpersonal competence of the beginning teacher involved the observation of actual situated behaviour. The five competences described above provided the framework for the assessors to analyse 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 beginning 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 beginning teachers in particular. Illustrative examples and counterexamples are presented for each of the five competences in Table 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 correcting disruptive pupil behaviour: (a) How does the teacher deal with undesirable situations? and (b) How does the teacher give a sequel to his or her interventions?

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-reflection. The teacher was asked to analyse both the results of the QTI and the videotaped lesson. Next, the teacher was asked to provide: (a) an overview of her

Table 1
Examples and counterexamples for five competences

1. Providing guidance/clear structure
Examples. The teacher:
- is able to make pupils active learners;
- keeps plenary moments functional;
- realizes smooth transitions between parts of the lesson.

Counterexamples. The teacher:
- appears insecure;
- allows pupils 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
Examples. The teacher:
- makes clear what behaviour is expected of the pupils;
- acts as a role model;
- formulates rules positively ("please do" as opposed to "do not").

Counterexamples. The teacher:
- sets rules which pupils perceive as unreasonable;
- communicates distrust in pupils;
- accompanies the explication of norms and standards with aggressive behaviour.

3. Correcting undesirable pupil behaviour
Examples. The teacher:
- shows a broad repertoire of low-intrusive corrections;
- aims corrections at pupil's behaviour and not pupil's personality;
- checks whether pupils respond to his or her corrections.

Counterexamples. The teacher:
- looks away while correcting pupil's behaviour;
- shows "more of the same" corrections;
- does not grant pupils a new chance after correction.

4. Paying attention to pupils
Examples. The teacher:
- shows personal interest in pupils;
- is patient with pupils;
- is approachable for pupils.

Counterexamples. The teacher:
- shows clear preferences in his or her treatment of individual pupils;
- tries to act as "pupil-with-the pupils";
- addresses pupils in a disrespectful manner.

5. Giving pupils responsibility and freedom
Examples. The teacher:
- gives the pupils an appropriate level of responsibility;
- gives the pupils 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 pupils perceive as insecure or weak;
- sticks too rigorously to his or her own ideas;
- cuts any and every initiative on the part of pupils short.

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: (a) Does the teacher show an awareness of her strengths and her weaknesses? and (b) Is the teacher able to formulate an adequate analysis of her behaviour?

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 beginning teacher. This report is presented in Table 2.

The judgement process was next analysed 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 beginning teacher's self-evaluation.
2. The overall judgement of the beginning 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 beginning teacher's self-evaluation and (b) the combination of evidence to attain an overall judgement.

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

Table 2
Expert report on the beginning 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 analyse her own behaviour or formulate concrete plans for improvement. We nevertheless consider this adequate for a beginning teacher. With respect to the Proximity domain, the teacher shows a basic ability to pay attention to pupils. Nevertheless, improvement is possible with regard to paying more personal attention to pupils and dividing one's attention more equally across the pupils. 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 pupil and beginning teachers as their problems mainly arise in this domain. Both behaviour and reflection within this domain were thus judged to be sufficient.

Table 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.

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 analysed using the coding scheme outlined in Table 3.

5. Results

5.1. Overview of judgements

In Table 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 beginning 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 beginning teacher's interpersonal competence.

5.1.1. Influence
As can be seen from Table 4, assessors A, C and D judged the beginning teacher's competence within the domain of influence to be sufficient. The assessors acknowledge the

Table 4
Summary of overall judgements provided by assessors

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

QTI results in connection with the videotaped lesson (i.e., in connection with the situated behaviour of the beginning teacher). *Assessor A* noted the following example of competence: "The beginning teacher has no difficulties in beginning to speak, and pupils listen rather quietly to her plenary introduction." The following counterexamples were also noted: "An inadequate repertoire of corrections with a higher intensity, the beginning teacher's recurrent invisibility and a lack of overview over the classroom affect her control over the classroom." *Assessor C* stated that the beginning teacher "obviously has difficulties providing leadership: the structure is not clear for the pupils, it is not clear what is allowed and not allowed, and disruptive pupil 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, pupils 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: "Pupils 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 pupils. The teacher does not set rules and does not correct pupil behaviour." Assessor D further notes that "nevertheless, it never becomes really chaotic." *Assessor B* was the only assessor to assess the beginning teacher's competence as insufficient in the domain of influence and based this conclusion on the results of the QTI which showed the beginning 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 beginning 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 beginning teacher appears to be secure during the videotaped lesson.

5.1.2. Proximity

While assessors A, C and D all rated the competence of the beginning 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 beginning teacher's situated behaviour. Assessor A stated that "although [the beginning teacher] clearly gets pupil recognition, acts respectful and states in her self-evaluation to be very satisfied with her relationship with the pupils, to me she appears primarily distant and detached. She fails to see pupils' signals, ignores difficult pupils and does not have any chats with pupils." Along these lines, *assessor C* stated that she does "not see much

interaction with pupils. Moreover, interaction always seems to be initiated by the pupils instead of by the teacher. Nonetheless, the beginning teacher is patient, responds to questions, rewards and walks around the classroom during coursework." *Assessor D* noted that "pupils consider this teacher 'a good sport'. The atmosphere is open, friendly and trusting." As a counterexample of competence, assessor D noted that the beginning teacher "frequently pays more attention for her own work than to pupils and that attention is paid to only those pupils who ask for it." *Assessor B* assessed the beginning teacher's competence within the domain of proximity to be insufficient and observed that "although [the beginning teacher] is patient, praises pupils frequently, and is helpful, I perceive a considerable distance between teacher and pupils. I cannot sense a good rapport. For example, when pupils enter the classroom, the teacher does not call the pupils 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.

5.1.3. Reflection

Just as for the domains of influence and proximity, assessors A, C and D judged the beginning teacher's competence within the domain of reflection to be sufficient but assessor B did not. The four assessors generally agreed that although the beginning teacher showed awareness of her strengths and weaknesses, she was not able to analyse 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 beginning teacher's self-reflection that "no weaknesses in the domain of the relationship with the pupils whatsoever are mentioned" and clearly stated that she considered this "rather problematic." She concluded that the beginning 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 beginning teacher herself expresses doubts in her self-evaluation. The teacher could be reluctant to change her behaviour." Nevertheless, assessor C thinks that the beginning 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 beginning teacher could not adequately analyse her own behaviour and therefore could not formulate concrete solutions for problems. According to this assessor, this is simply unacceptable for even a beginning teacher. Just as assessor C, assessor B found it difficult to assess the beginning teacher's potential for future professional development. Assessor B observed that "[the beginning teacher] often looks for excuses for not realizing her ideal."

5.2. The validity of the judgements

In Table 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. 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*

Table 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*		

was found to occur for the competences of providing guidance/clear structure (1) and paying attention to pupils (4). The notes of three of the assessors, assessors 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. Assessor 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 counterexamples 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

underrepresentation in her overall judgements because competences 1 and 4—which were underrepresented in her original notes—were sufficiently represented in her overall judgement although the statements referring to the relevant competences were rather abstract.

6. Conclusions and discussion

The purpose of the present study was to develop and validate an assessment procedure to evaluate the interpersonal competence of beginning teachers. In the present paper, the validity of the judgement process was considered in particular. The results for four judges who evaluated the same beginning teacher using the assessment procedure showed the validity of their judgements to be satisfactory. The amount of construct-irrelevant variance and construct underrepresentation found for the assessment processes of (a) noting evidence and (b) combining evidence from different sources was small. Although our enquiry was restricted to the *validity* of the judgements provided, the results of this study nevertheless suggest that the interrater reliability of the assessment procedure was reasonable: Three of the numeric scores assigned by the four assessors were largely in agreement with each other.

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., pupil perceptions of their relations with a teacher) should be combined with evidence from the observation of situated teacher behaviour. Pupil 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 beginning 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 beginning 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-reflection 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 beginning teacher may not be revealed by two successive administrations of the QTI while behavioural observation and self-reflection 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 beginning 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 beginning teachers using the present assessment procedure would nevertheless provide further insight into the utility and validity of the procedure. Particularly challenging beginning 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.

Acknowledgement

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 (Eds.), *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*(2), 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*(2), 81–85.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27–32.

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

Haertel, E. H. (1990). Performance tests, simulations, and other methods. In L. M. Darling-Hammond, & J. Millman (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers.* London: Sage.

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

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(2), 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*(9), 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*(2), 139–161.

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/.

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*(2), 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*, this issue, doi:10.1016/j.ijer.2006.03.003.