This item is the archived peer-reviewed author-version of:

Measuring pre-service teachers' professional vision of inclusive classrooms : a video-based comparative judgement instrument

Reference:

Keppens Karolien, Consuegra Els, Goossens Maarten, De Maeyer Sven, Vanderlinde Ruben.- Measuring pre-service teachers' professional vision of inclusive classrooms : a video-based comparative judgement instrument
Teaching and teacher education - ISSN 0742-051X - 78(2019), p. 1-14
Full text (Publisher's DOI): <https://doi.org/10.1016/J.TATE.2018.10.007>
To cite this reference: <https://hdl.handle.net/10067/1575260151162165141>

Measuring pre-service teachers' professional vision of inclusive classrooms: a video-based comparative judgement instrument

Abstract

This study reports on the development and validation of a video-based measurement instrument to assess pre-service teachers' ($n=278$) professional vision of inclusive classrooms (PVIC). PVIC includes two dimensions of teacher quality that are essential in effective inclusive classrooms: (positive) teacher-student interactions (TSI) and differentiated instruction (DI). It is the first study to measure pre-service teachers' professional vision through a holistic video-based approach using comparative judgement. The validity framework of Chan (2014) was used to investigate construct validity. The results show that the instrument is reliable and valid, and recommendations for use in teacher education programs are formulated.

Keywords: student teachers, professional vision, comparative judgement, measurement, inclusive education

1. Introduction

Schools and teachers worldwide are facing a growing diversity in pupil populations (Cochran-Smith & Power, 2013). This is most visible in primary education as a result of the comprehensive educational system in primary schools and demographic factors, such as growing child poverty and immigrant populations. Diversity has become a reality in primary schools and differences between students are a daily fact for teachers. Teachers are challenged to meet the needs of diverse learners and adapt their teaching to heterogeneous academic abilities, interests, backgrounds, and motivations (Banks et al., 2005). This growing diversity is reflected in educational policies (United Nations, 2007) and reform efforts to establish more inclusive learning environments. However, teacher education programs have not addressed diversity with great success. They offer limited preparation to pre-service teachers in developing the teaching skills necessary to support achievement for all pupils in the classroom (McDonald, 2005). Consequently, it is necessary to support pre-service teachers in their new role as facilitators of inclusion (Mills & Ballantyne, 2010). Therefore, teacher education programs should place sufficient emphasis on helping future teachers to notice and interpret aspects of teaching and learning (van Es & Sherin, 2002) that are important in the context of inclusive classrooms. In this regard, Sherin (2001) introduced the concept of professional vision. Professional vision is referred to as observing what is happening in a classroom through the ability to notice and interpret significant features of classroom interactions (Sherin, 2001). Although professional vision is important for successful teaching, there are currently no reliable and valid tools available for assessing it in the context of pre-service teacher education on inclusive primary education. Generally, professional vision is assessed with classroom videos from secondary schools combined with open questions (van Es & Sherin, 2008) in qualitative research or rating items (Seidel & Stürmer, 2014) in quantitative research (Gold & Holodynski, 2017). These measurement approaches are entirely based on analytic assessment methods by

making use of rubrics to score the rating process. Although analytic assessment methods are predominantly used in measuring professional vision, standardized holistic assessment approaches are promising as an alternative to traditional testing. Holistic assessment involves judging integrated substances, while analytic assessment requires criterion-by-criterion judgements (Sadler, 2009). A possible strategy with regard to use of a more holistic assessment approach can be found in the recent work of Pollitt on comparative judgement (CJ) (Pollitt, 2012a, 2012b). CJ is a measurement strategy in which various assessors independently compare pairs of representations (e.g. videos about classroom teaching) and judge which one is better (Laming, 2003).

In this study, analytic and holistic assessment methods are combined to measure professional vision in a standardized way. Video-based comparative judgement is presented as a complementary approach for assessing professional vision in a more holistic manner. The current study reports on the development and validation of a video-based measurement instrument to assess pre-service teachers' professional vision of inclusive classrooms (PVIC).

2. Theoretical framework

1.1. Modeling professional vision

Over the last decade, research in teacher education has been paying increasing attention to the area of professional vision. This is partly due to current attempts to develop more ways to help pre-service teachers connect theory and practice (Stürmer, Seidel, & Schäfer, 2013b). For instance, many pre-service teachers struggle when they begin their first teaching job, known as “practice shock” (Kelchtermans & Ballet, 2002), and find themselves unable to cope with the complexity of the classroom reality. They experience a discrepancy between what they have learned during their teacher education and the actual classroom (Stokking, Leenders, de Jong, & van Tartwijk, 2003). This is described as the theory-practice gap (Korthagen & Kessels,

1999). A promising approach for the integration of theory and practice is seen in the concept of professional vision (Seidel & Stürmer, 2014).

The concept of professional vision dates back to Charles Goodwin (1994) who defined it as the “socially organized ways of seeing and understanding events that are answerable to the distinctive interests of a particular social group” (Goodwin, 1994, p. 606). Sherin (2001) first introduced Goodwin’s concept in the field of teacher education and defines the concept as teachers’ ability to observe and interpret what is happening in a classroom. Sherin and van Es (2009) describe professional vision as consisting of two subprocesses: (a) noticing and (b) knowledge-based reasoning. Noticing concerns how the teacher decides where to pay attention to at a given moment (Sherin, 2007). It describes a process of selective attention by which teachers identify important classroom events relevant for student learning (van Es & Sherin, 2002). Knowledge-based reasoning refers to the way in which events in the classroom are interpreted based on teachers’ knowledge and understanding (van Es & Sherin, 2008). These researchers have discussed professional vision as a way to attend to and make sense of important features of instruction. It demonstrates a way of thinking and reflecting about instruction and its impact on student learning (van Es & Sherin, 2008).

Sherin’s (2007) application of professional vision to teacher education has been further developed by colleagues in the field (e.g. Stürmer, Könings, & Seidel, 2013a; Stürmer et al., 2013b; Seidel & Stürmer, 2014). Recent research still defines professional vision as consisting of (a) noticing and (b) reasoning (Blomberg, Stürmer, & Seidel, 2011; Stürmer et al., 2013a), which is in analogy with Sherin and van Es (2009). However, they slightly differ in their conceptualization and operationalization of professional vision. According to Seidel and Stürmer (2014), noticing involves pre-service teachers’ identification of classroom events that are significant for effective instructional practice and student learning. Reasoning implies how events in the classroom are interpreted based on knowledge about teaching and learning. These

researchers deviate from the construct as it was originally defined by conceptualizing professional vision merely as knowledge about a pre-set features of effective strategies. Both conceptualizations have in common that professional vision is informed by knowledge of what constitutes teaching and learning (Seidel, Blomberg, & Stürmer, 2010). It is seen as an indicator of the extent to which pre-service teachers are able to apply their conceptual knowledge of teaching and learning to authentic classroom situations (Kersting, Givvin, Sotelo, & Stigler, 2010). In other words, the ability to connect theory with practice. This means that professional vision is a knowledge-based process (Steffensky, Gold, Holodyski, & Möller, 2015). For example, Stürmer et al. (2013a) showed that student teachers acquire declarative pedagogical knowledge in different university courses, but also develop professional vision. Kersting et al. (2010) found a high correlation between teachers' mathematical knowledge for teaching and their ability to analyze videos from mathematics lessons. Accordingly, König et al. (2014) demonstrated that general pedagogical knowledge correlates substantially with the ability to interpret classroom situations. Depending on the focus, professional vision might require different aspects of knowledge. Professional vision with a focus on inclusiveness primarily draws on aspects which can be related to general pedagogical knowledge, as it is applicable in a wide variety of subjects (Voss, Kunter, & Baumert, 2011).

Pre-service teachers who have no or very limited knowledge are often not able to notice and interpret complex events in the classroom (Stürmer et al., 2013a). Studies on teacher cognition (Borko & Livingston, 1989) reveal that experts are better able to selectively notice and interpret classroom situations. Professional vision is therefore a distinctive feature of experts (Sabers, Cushing, & Berliner, 1991). Expert teachers perceive classroom situations more accurately and holistically as compared to novice teachers (Palmer, Stough, Burdenski, & Gonzales, 2005).

1.2. Professional vision of inclusive classrooms

Most studies of professional vision focus either on pedagogical-psychological aspects of teaching, such as classroom management, or content-specific aspects of teaching, such as students' mathematical thinking. In this study, professional vision is investigated in the context of inclusive classrooms by focusing on two dimensions of teacher quality: teacher-student interactions (TSI) and differentiated instruction (DI). These two dimensions are selected since empirical research has demonstrated both to be essential in the context of inclusive classrooms in primary education (e.g. Mitchell, 2014; Stanford & Reeves, 2009; Florian, 2008). In addition, professional vision needs to be directed towards events that can be noticed in video. Therefore, TSI and DI were chosen because of their visibility in video clips concerning inclusive classrooms. Moreover, empirical research shows that both dimensions have substantial impact on students' cognitive, affective and motivational outcomes (e.g. Roorda, Koomen, Spilt, & Oort, 2011; Van Petegem, Aelterman, Van kee, & Rosseel, 2008), and are particularly important for students at risk of academic failure (Hamre & Pianta, 2001; Valiandes, 2015). Research consistently shows that TSI and DI play a large part in eliminating achievement gaps for pupils with ethnic minority backgrounds, low SES or learning difficulties (e.g. Beecher & Sweeny, 2008; Reis, McCoach, Little, Muller, & Kaniskan, 2011).

1.2.1. Teacher-student interactions (TSI)

Positive teacher-student interaction is one of the most important factors in determining student outcomes (Roorda et al., 2011). TSI is closely related to the concept of emotional support in classrooms (Pianta & Hamre, 2009), which is in turn guided by the attachment theory (Pianta, 1999).

Based on the work of Diana Baumrind (1966) on how responsive an adult is to the individual needs of children and how much authority they assume, Englehart (2009) presented three types of interaction styles in the classroom: permissive (characterized by high warmth and

low control), authoritarian (characterized by low warmth and high control) and authoritative (characterized by high warmth and high control).

The authoritative interaction style best serves the diverse educational needs of students (Englehart, 2009). This means that a teacher provides structure and control in order to establish a respectful relationship and an effective learning climate on the one hand, and takes into account the needs and interests of students for their comfort and personal development on the other hand (Englehart, 2009). Thus, the responsiveness and sensitivity of the teacher is important in meeting the needs of diverse learners. The attachment theory states that sensitive and responsive TSI are the basis for the development of positive interactions in the classroom (La Paro, Pianta, & Stuhlman, 2004). A sensitive and responsive teacher acknowledges students' interests, concerns, and questions, and handles them appropriately (De Vroey, Struyf, & Petry, 2016). The teacher is sensitive and responsive to the social, emotional, and cognitive needs of the student. A responsive and sensitive teacher will act as a secure base from which students can explore the learning environment (e.g. students dare to make mistakes and ask questions in the class). The literature on responsive teaching also regards *caring* as critical for teaching diverse students (Gay, 2002). A caring teacher is understood to be approachable and interested in the personal situation of each child, establishing trust and respect in a caring relationship. Caring about diverse students is also related to caring about their achievement and thus expecting high-level success from them (Gay, 2002). This facilitates the process of student autonomy and enhances student motivation (Englehart, 2009).

1.2.2. Differentiated instruction (DI)

In response to student diversity, differentiated instruction has been adopted to address the individual learning needs and the different abilities of students in the same class. Differentiated instruction aims to maximize students' learning opportunities by supporting them in the learning process. Tomlinson (1999) argues that differentiated instruction is “an

approach to teaching in which teachers proactively modify curricula, teaching methods, resources, learning activities, and student products to address the diverse needs of individual students and small groups of students to maximize the learning opportunities for each student in a classroom” (Tomlinson, 1999, p. 121). Although differentiated instruction is used in different ways, it generally refers to practices of addressing student differences in terms of readiness, background knowledge, language, interest, and learning profile (Hall, 2002; Tomlinson, 1999, 2001, 2004). Tomlinson (2001) distinguishes between differentiation at the level of content, process, and products. Each of these elements of the curriculum can be differentiated. However, it still remains difficult to make statements about practices of differentiated instruction (Coubergs, Struyven, Vanthournout, & Engels, 2017). Building on this, Couborgs et al. (2017) carried out a large-scale study to identify components of differentiated instruction in Flanders (Belgium). They found that flexible grouping and assessment are two components that have a positive effect in differentiated instruction. Flexible grouping refers to cooperative peer learning, and homogeneous and heterogeneous group dynamics (Tomlinson, 2004). Assessment refers to the importance of using student evaluation (e.g. information from tasks/assignments, tests, classroom behavior,...) to get an impression of students’ learning process in order to react responsively to students’ individual needs (Hattie, 2009). Their overall model also emphasizes the need for differentiation with respect to the readiness, interests and learning profiles of students (Couborgs et al., 2017).

1.3. Measuring professional vision

Professional vision can be assessed through analytic or holistic scoring. The current tools used to measure professional vision (such as ‘The Observer’) generally employ analytic assessment methods (Seidel & Stürmer, 2014). These measures of pre-service teachers’ professional vision are based on criteria that make a distinction between different aspects of

teacher knowledge and use a scoring rubric to assist the rating process. For instance, in the instrument of Kersting and colleagues (Kersting, 2008; Kersting et al., 2010) a rubric is applied to teachers' written responses on videos of mathematics classroom situations to score teachers' knowledge of teaching mathematics. In the instrument 'Observer' (Seidel & Stürmer, 2014), assessment criteria are translated in rating items to assess pre-service teachers' professional vision of three teaching and learning components (i.e. goal clarity, teacher support, and learning climate). These pre-defined criteria allow a detailed focus on the different facets of the subject under assessment, and thus offer a frame of reference to judge every subject in the same way (Author et al., 2015). This enhances the reliable scoring of performance assessments (Jonsson & Svingby, 2007).

Although analytic assessment approaches are valuable in measuring professional vision, Bramley (2015) and Pollitt (2012a) propose a more comprehensive alternative for competence assessment, namely comparative judgement. Whereas analytic assessment approaches may lead to more precise and reliable judgements, holistic assessments are a more valid representation of the subject under assessment as a whole (Barkaoui, 2007). In this study, video-based comparative judgement is combined with rating items to assess pre-service teachers' professional vision of TSI and DI. Previous research (e.g. Seidel & Stürmer, 2014; Gold & Holodynski, 2017) proved that rating items are effective to measure pre-service teachers' professional vision in a more standardized way. Besides, the study of Author et al. (2018) provides more insight in the strategy of comparative judgement as a method to study professional vision.

1.3.1. Video representations

Video has become a prominent tool in assessing professional vision in the context of teacher education (Sherin & van Es, 2009). Video provides an excellent means to support the theory-practice connection that is inherent in the concept of professional vision (Brophy, 2004).

This explains why the concept of professional vision has been taken up in research on video-based teacher education. Videos capture the complexity and authenticity of classroom events in real time (Sherin & van Es, 2009) and can be used to avoid overwhelming pre-service teachers with the stress of interacting with a classroom (Sherin, 2004).

Moreover, research indicates that video analysis offers opportunities to study and develop pre-service teachers' professional vision (Stürmer et al., 2013b). More concretely, video supports pre-service teachers' attempts to apply their knowledge about teaching and learning to actual classroom lessons (Kinzer & Risko, 1998). In assessing professional vision, authentic video sequences of classroom situations are used as "prompts" to elicit teacher knowledge (Kersting, 2008). These short sequences display a richness of authentic teaching and learning.

1.3.2. Comparative judgement (CJ)

CJ is a holistic and intuitive measurement strategy that has seldom been used in the educational domain (Heldsinger & Humphry, 2010). CJ is based on Thurstone's method of pairwise comparison (Thurstone, 1927), which states that people are more reliable in comparing things rather than assigning scores to things. In this method, assessors independently compare several representations (objects, student assignments,...) and decide which of them demonstrate the 'best' performance of the given representation (Pollitt, 2012b). In other words, an assessor is asked to compare two student assignments, for example, and decide which of them is better. This rating procedure assists raters to judge the quality of a representation holistically, but also offers support by comparing it against another representation (Author et al., 2017). Moreover, these comparisons produce a measurement scale in which the relative quality of each representation is shown (Pollitt, 2012b). This means that representations can be ranked from worst to best on a scale. This rank-order is based on the comparisons of several assessors, where

the scale serves as the shared consensus of what a good representation encompasses (Pollitt, 2012a).

According to Author et al. (2017), assessors find it easy to make such comparisons because it is less time consuming and more intuitive to compare things rather than to focus on lists of rubrics and criteria. This leads to higher consistency in judgements over different assessors (Bramley, 2007). Moreover, validity is enlarged by the more natural judgement that assessors attempt (Author et al., 2017).

3. Research goal and research questions

The aim of this study is to develop a reliable and valid video-based measurement instrument to empirically assess and describe pre-service teachers' professional vision of inclusive classrooms (PVIC), respectively professional vision of TSI and DI.

The validity of results depending on the instrument is based on the validity framework of Chan (2014). This framework is used in the context of AERA and APA test standards. Construct validity is the central component in the framework. It is defined as the extent to which an instrument measures the concept it intends to measure (Cook & Campbell, 1979) and builds on five sources of evidence (Chan, 2014): (1) the content of the instrument, (2) response processes, (3) internal structure, (4) relations to other variables, and (5) consequences of use in instruction and learning. In this study the first four sources of evidence will be investigated by means of four research questions.

Evidence based on the content of the instrument. (RQ1): Which video clips are best suited to elicit professional vision of TSI and DI?

This refers to the content relevance and representativeness of the instrument in relation to the construct that is supposed to be measured (Chan, 2014). This is important with regard to the validity and quality of the selected videos in the sense that they represent discernible

examples of TSI and DI. The complexity of classroom teaching is difficult to capture in a single instrument, therefore it is important to make decisions about which video selections to include in the instrument.

Evidence based on response processes. (RQ2): How do participants experience the use of the instrument in terms of cognitive load?

This refers to the examination of the cognitive processes employed when participants fill in an instrument (Chan, 2014). Cognitive load represents the load that responding to questions imposes on the cognitive system (Paas, Van Merriënboer, & Adam, 1994). Answers to this question are important with regard to mental effort demanded by the instrument and utility in primary teacher education programs. If mental effort is too heavy, this might lead to respondents' attention dropping and being less accurate in their test.

Evidence based on internal structure. (RQ3): To what degree does our model of professional vision (noticing and reasoning) fit the data generated by our measure?

The internal structure refers to the degree to which items represent the construct of the instrument by investigating how items relate to each other (Chan, 2014). First, the aim is to investigate the reliability and validity of comparative judgement as a strategy to assess pre-service teachers' ability to notice classroom events. Second, the aim is to investigate the operationalization of a two-dimensional latent factor for TSI and DI to assess pre-service teachers' reasoning around TSI and DI. In other words, to test whether TSI and DI could each be described by one content factor, operationalized as a two-dimensional model based on TSI and DI as latent factors.

Evidence based on relations to other variables. (RQ4): How is pre-service teachers' ability to notice related to their demographic variables?

This concerns the association between instrument scores and external variables. Research suggests that there is a positive relation between pre-service teachers' educational

training, expertise, and their noticing abilities (Stürmer et al., 2013b). Therefore, the association between pre-service teachers' ability to notice and demographic variables related to their educational training will be investigated.

4. Method

4.1. Research context

With the ratification of the UN convention (United Nations, 2007) on the Rights of Persons with a Disability in 2009, Flanders agreed to develop a more inclusive educational system, supported by the Flemish M-decree 'Measures for students with special educational needs', approved in 2014. Hence, teacher candidates in Flanders - educated at bachelor's level in colleges of higher education (180 ECTS) - are not prepared to teach in inclusive classrooms. Therefore, the Flemish research project – [details removed] - has been introduced to develop and strengthen (pre-service) teachers' competences to create inclusive learning environments. In order to map pre-service teachers' competence development, a video-based measurement instrument has been developed.

4.2. Instrument development and validation process

The development and validation of the instrument is structured around three phases. Figure 1 gives an overview of the different phases included in the development and validation of the instrument and their relation with each other. The first phase is part of the development of the instrument and refers to collecting and selecting video clips that represent discernable examples of TSI and DI. In the second phase, a study with experts was conducted to answer our first research question where we validated the video clips that were selected in phase 1. Experts were selected based on their ability to notice critical classroom events related to TSI and DI. This was in order to develop an expert benchmark for noticing and reasoning

competencies for TSI and DI. In the third phase, a study with pre-service teachers was conducted to investigate the second, third and fourth research question. These research questions are part of the validation of the instrument in which we tested model assumptions and utility in primary teacher education.

<<Figure 1>>

Below, the different phases with their corresponding studies are presented, and the development and validation of the instrument is described. Phase 2 and 3 consist of two studies that test our research questions.

4.2.1 Phase 1: Selecting video clips

The videos were collected in 2014 as part of a previous project. That project gathered and collected 32 hours authentic video data in 8 primary schools in Flanders to determine critical success factors of effective differentiated instruction.

In order to select videos that show TSI and DI as described in the literature, a standardized coding scheme was developed. We first developed a theory-based coding scheme where TSI was defined in relation to the concepts of caring, sensitiveness and responsiveness of the teacher. DI was operationalized in relation to teaching methods, flexible grouping and assessment, which take into account the readiness, interests and learning profiles of students. Second, this coding scheme was used by three independent coders (i.e. researchers from the research project) and the coding results were compared and discussed. The codebook was subsequently adjusted and tested in a second round of coding, discussion and codebook modification. After two cycles of coding, a stable coding scheme with a good inter-rater reliability was achieved. Following this, the coders proceeded to code all remaining videos, and segments that represented TSI and DI were selected. These video segments were edited

into 135 video clips of approximately 2 minutes which were coded independently by the three coders based on a four-point Likert Scale that was added to the coding scheme (i.e., ranging from 0=not presented to 3=strongly presented). Only video clips which strongly represented TSI and DI were used for the instrument. As such, 15 video clips were selected for use in the instrument. Both positive and ambiguous examples of TSI and DI were selected in order to cover positive and improvable features.

4.2.2 Phase 2: Expert study

This study was carried out to investigate the first research question in which the video clips were validated. A digital platform, called D-PAC (Digital Platform for Assessment of Competences) was used to review the videos. Experts were invited to participate in the study on a voluntary basis. They received an online link hosting the D-PAC platform.

4.2.2.1 Participants

A convenience sample of 34 volunteer experts – of which 26 are female – participated in this study. Colleagues to whom the research team has regular access were invited to participate. The criteria for being an expert were: (1) working in the field of teacher education as a teacher educator in diversity courses (n=10), (2) having a strong background in empirical research on teaching and learning or inclusive education (n=15), and (3) working in the educational field as a pedagogical supporter with experience in inclusive classrooms (n=9).

4.2.2.2 Materials and procedure

Two measures were used to validate the video clips: (a) a comparative judgement measure to review the video clips, and (b) open questions to gather positive and negative aspects of each video clip in relation to TSI and DI.

Comparative judgement measure. Experts received an online link to judge 15 video clips via an internet browser in the DPAC-platform. The platform presented each expert with randomly assigned pairs of videos. After watching both videos, experts decided which video

demonstrated the best performance and recorded their choice within the CJ system. Experts were instructed to judge the video clips by focusing on TSI and DI separately. For each comparison, experts were asked to choose which clip is best with regard to TSI on the one hand and DI on the other hand. No other guidance was given in order to allow for holistic judgements. After choosing a representation of a video clip, experts received a new pair to compare.

Open questions. Experts were asked to provide written comments on positive and negative aspects of each video clip with regard to TSI and DI. These written comments were requested after every comparison. Experts could make a comment about what was in their minds when they made their decision before moving on to the next pair of video clips. This was done to gather detailed information on every video clip.

4.2.2.3 Data analysis

The method of Comparative Judgement (CJ) is based on the Bradley-Terry-Luce model (Bradley & Terry, 1952; Luce, 1959), which is comparable to the Rasch model (Andrich, 1978). Applying the BTL-model, the judgements of experts lead to a rank-order and scale of the video, ranging from worst to best (Author et al., 2017). This rank-order is informed by the number of times a video clip ‘beats’ or ‘loses’ from another. The calculation of the position in this rank order is based on the normal distribution (Andrich, 1982). To gain insight into the quality of the video clips, quality measures were performed (Author et al., 2017). First, the Rasch separation reliability was calculated. This index of reliability is referred to as alpha or Scale Separation Reliability (SSR; Bramley, 2015). The SSR represents the extent to which the spread in the results is not caused by measurement error (Andrich, 1982). A high SSR implies that the relative position of the video clips on the scale is fixed, illustrated by a minimal measurement error (Andrich, 1982).

Second, we calculated whether one expert deviates from the others (*assessor Infit*) and whether there is disagreement on certain video clips (*representation Infit*). By using *chi-*

squared (χ^2) *goodness of fit* statistics, the fit of the data to the model was evaluated. The model expectation is the probability that video A will win over video B. In other words, experts' comparisons are statistically translated into odds of winning from one another (Pollitt, 2012a). An *Infit* that lies two standard deviations from the mean indicates a misfit (Pollitt, 2012b). This means that experts consistently judge away from the consensus. This could be because experts have other conceptualizations of what makes a video clip better or worse (Bramley, 2015). A video clip with a high *Infit* could contain something unusual (Bramley, 2015), therefore experts might find them difficult to judge.

Moreover, Nvivo was used to analyse experts' comments on positive and negative aspects of video clips targeting TSI and DI. This resulted in 23 items tapping TSI and 23 items tapping DI.

4.2.3 Phase 3: Student study

A student study was carried out to investigate research question 2, 3 and 4. These research questions were formulated to validate the instrument.

4.2.3.1 Participants

The student study was conducted within two colleges of teacher education in Flanders. Students enrolled in the study followed a three-year curriculum program to become a primary school teacher. Through the head of the department, 746 students received an online link hosting the instrument. They were invited to participate in the study on a voluntary basis. A total of 278 pre-service teachers participated in the study, of which 272 filled in a background survey. The age of the respondents varied from 19 to 48 years, with a mean age of 22 years. The respondents were not divided evenly across both genders: 232 respondents were women, 40 were men. Out of the 272 students, 134 were first year students, 48 second year students, 43 third year students, and 47 students followed an individualized curriculum. In terms of education: 89% followed regular education and 11% followed distance education.

4.2.3.2 Materials and procedure

Data on pre-service teachers' PVIC were collected by means of two measures in the instrument: (a) a comparative judgement measure to assess pre-service teachers' ability to notice TSI and DI, and (b) rating items to assess pre-service teachers' reasoning of TSI and DI. The instrument starts with general instructions, followed by the comparative judgement measure and rating items. A scale (Paas et al., 1994) to measure pre-service teachers' experiences with the instrument in terms of cognitive load was added to the instrument. Each of these will be described in turn.

Ability to Notice. Pre-service teachers' received an online link to view 15 validated video clips. A design was constructed in which pre-service teachers were asked to compare 10 pairs of randomly assigned video clips. An equal view algorithm was implemented to make sure that every student was able to see all the video clips once. Two randomly chosen videos were presented side by side on a computer screen. Pre-service teachers were instructed to judge the video clips in terms of TSI and DI. For each comparison, pre-service teachers were asked to choose which clip is best in terms of TSI on the hand and DI on the other hand. This was done by using the following prompt: 'specify in which video the teacher works more inclusive in his or her interactions with the students (video A or B)' and 'specify in which video the teacher works more inclusive in his or her instruction and teaching methods (video A or B)'. This method requires pre-service teachers to first identify classroom situations related to TSI and DI before selecting the best video. Pre-service teachers need to build upon their ability to notice relevant teaching practices in comparing video clips in terms of TSI and DI. In other words, it reflects the ability to identify and differentiate between relevant events related to TSI and DI without making any further judgements (Stürmer et al., 2013b). Moreover, an ability measure was implemented that controls for the extent to which a student has only difficult

comparisons compared to only easy comparisons. A correction is applied in which the same ability scores are generated for videos of comparable quality (Author et al., 2017).

Reasoning. After the comparisons, pre-service teachers were asked to indicate how decisive provided aspects were when comparing pairs of video clips. Therefore, arguments (i.e. rating items) were constructed (based on the comments of the expert study) to measure pre-service teachers' reasoning of TSI and DI. Pre-service teachers were asked to motivate why they chose one video above another by choosing out of a list of pre-defined arguments. This means that rating items were constructed to motivate pre-service teachers' noticing decisions (see Table 5 for item examples).

They had to rate each item separately on a five-point Likert scale (ranging: 0=not important, 1=slightly important, 2=moderately important, 3=important, 4=very important). They were presented to students at the end of the 10 comparisons, after having observed all the video clips. First the items were shown for TSI, followed by the items presenting aspects of DI. The more students rely on certain aspects (rating items) in their reasoning (or the more important an argument was in their judgement for choosing the best video), the closer they relate to experts. Higher mean scores (range 0-4) for individual rating items indicate that these aspects were important for pre-service teachers in choosing the best video clip.

Cognitive load. The self-report Invested Mental Effort scale of Paas et al. (1994) was used to measure pre-service teachers' experience with the instrument in terms of cognitive load. It is a single-item with a nine-point Likert scale (ranging from 'very, very low mental effort' to 'very, very high mental effort'). This scale was presented to students directly after the reasoning items so that they could rate the level of perceived mental effort required on this scale.

4.2.3.3 Data analysis

To answer our second research question, descriptive analyses of the cognitive measure were performed in SPSS to capture participants experiences with the instrument in terms of mental effort.

To test the model of professional vision, our third research question, a stepwise approach was adopted. First, we conducted analyses in R to validate pre-service teachers' noticing measure. The Bradley-Terry-Luce model (Bradley & Terry, 1952; Luce, 1959) was used. It was assumed that pre-service teachers' ability to notice could be measured efficiently through the method of Comparative Judgement. Based on the Joint Maximum Likelihood procedure, two types of parameters (Judge Infit and Representation Infit) can be estimated, producing an ideal likelihood for the data.

Second, it was assumed that pre-service teachers' reasoning could be measured as a continuous two-dimensional latent variable for TSI and DI. Therefore, several steps were involved: (1) an EFA to identify TSI as one factor and DI as one factor, (2) a CFA to examine the stability of the exploratory factor structure, and (3) a reliability analysis to determine the internal consistency of the factors. Two separate one-dimensional CFA's were done as previous research on experts' professional vision of TSI and DI indicated that both dimensions are conceptual and empirical separable (Author et al., 2018). Following the recommendations of Hair, Black, Babin, Anderson and Tatham (2006), all items with loadings of .50 and less were excluded from further analysis. For the CFA, several fit indices were calculated to determine the adequacy of the fitted model: (a) the χ^2 and p-value, (b) the comparative fit index (CFI), (c) the Tucker-Lewis index (TLI), (d) the root mean square error of approximation (RMSEA), and (e) the standardized root mean square residual (SRMR). Following Hu and Bentler (1999), cutoff values of $\leq .06$ and $\leq .08$ for RMSEA and SRMR respectively indicate a good fit. CFI and TLI scores $\geq .90$ indicate adequate fit, while scores of $\geq .95$ indicate a good fit. To test the

internal consistency of the instrument, Cronbach's alpha was calculated. Factors with an α of 0.80 are considered reliable (Tabachnick & Fidell, 2007).

To answer the fourth research question on the relation between pre-service teachers' ability to notice and other variables, SPSS was used for multiple linear regression analyses of cross-sectional data.

5. Results

5.1. Evidence based on the content of the instrument (RQ1)

The comparative judgement measure produces a rank order of the video clips, ranked from worst to best. Table 1 shows the rank order of videos targeting TSI and DI as a result of the pairwise comparison of experts and the according scale scores expressed in logit scores. This means that, for instance, Video 3 is judged as the worst concerning both TSI and DI. Because the rank order is based on the decisions of several experts, the scale represents the shared consensus of what a good video clip comprises (Pollitt, 2012a).

<< Table 1. Rank order of the video clips targeting TSI and DI >>

Next, the reliability of experts' judgements in relation to TSI and DI was examined. Based on the Scale Separation Reliability (analogous to Cronbach's alpha), the results indicate that experts' reliability of judgements in relation to TSI was good (SSR = .70). Experts' reliability of judgements in relation to DI was even higher (SSR = .78). The alpha shows that experts were consistent with each other in terms of rank order.

Furthermore, we calculated whether experts deviated from the consensus (Judge Infit). The results showed that judge 18 and judge 23 exceeded the criterion of an Infit value higher than 2 (see Table 2). To a significant degree, these judges interpreted the video clips targeting TSI differently from others. Therefore, data from these two misfitting judges was removed from

the file in order to use it as a benchmark for the measurement of professional vision of pre-service teachers. A difference in reliability was found after removing these misfitting judges. The SSR increased from 0.70 to 0.74. Regarding the video clips targeting DI, only judge 18 exceeded the Infit criterion of a value of 2. After removing this misfitting judge, reliability increased from 0.78 to 0.79.

<< Table 2. Judge fit measures (Outfit and Infit) >>

Next, we calculated whether there is disagreement among experts about some video clips (Representation Infit). For TSI this is summarized in Figure 2. As all dots are within the range indicated by the red and green line, we can conclude that there is no video clip for which a significant amount of disagreement is present.

<<Figure 2>>

Figure 3 shows that there is one video clip (video 11) that leads to more inconsistent judgements about DI (see dot above the red line). This means that experts find it difficult to consistently judge this video clip in terms of DI. The criterion value for a video clip to be considered misfitting is conventionally taken to be 2. Video 11 is situated very close above the cut-off of 2. However, video 11 represents characteristics of TSI and DI that were not covered by any other video clip. Therefore, it is still acceptable to retain the video clip.

<<Figure 3>>

5.2. Evidence based on response processes (RQ2)

The results from the mental effort scale (range 1 ‘very, very low mental effort’ to 9 ‘very, very high mental effort’) confirms that the instrument does not involve too much complexity ($M=5.14$, $SD=1.28$). The average mean score (5.14) indicated that pre-service teachers experienced “neither low nor high mental effort” during the completion of the instrument. This suggest that our instrument is perceived as stimulating and activating, on the one hand, and does not involve too much complexity, on the other hand. It implies that the instrument does not overwhelm pre-service teachers with high mental effort (Seidel & Stürmer, 2014). Moreover, the perceived amount of mental effort invested in this study is the same across teacher education type (distance education $M = 5.16$, regular education $M = 5.14$) and gender (men $M = 5.21$, women $M = 5.13$). These mean scores are comparable with mean scores reported in previous research on teacher education students (e.g. De Leeuw & Mayer, 2008). In addition, the completion time of the instrument was about 40 minutes.

5.3. Evidence based on internal structure (RQ3)

5.3.1 Ability to Notice

The reliability of measuring pre-service teachers’ noticing through CJ was examined. The results showed a high reliability for pre-service teachers’ judgements targeting TSI ($SSR = 0.78$) and DI ($SSR = 0.86$).

Next, we calculated whether pre-service teachers deviated from each other (Judge Infit). The results showed that only 6 students (2%) exceeded the criterion. To a significant degree, these students interpreted the video clips related to TSI differently from the others. Regarding the video clips related to DI, 7 students (2.4%) had different interpretations of the videos. This is in line with the proportions of misfitting judgements in comparability studies, which is

usually less than 5% (Bramley, 2007). Table 3 shows the summary statistics for the Judge Infit for TSI and DI.

<< Table 3. Judge Infit summary statistics >>

The results show that the Infit is for 95% of the students situated between 0.772 and 1.234. This is between the grade boundary of 0 and 2 and indicates that the video clips are judged consistently by the students.

Furthermore, we calculated whether there were certain misfitting video clips (Representation Infit) targeting TSI and DI. Only Video 3 could be seen as a misfitting video clip in terms of TSI. However the low SD (0.012) indicates that this video clip cannot be considered as misfitting. This means that it was clear for students to make their judgements, as the videos contained no features that caused the students to vary in their decisions. Table 4 shows the different summary statistics for the Representation Infit for TSI and DI.

<< Table 4. Representation Infit summary statistics >>

The summary table shows that the Infit is for 95% of the video clips situated between the grade boundary of 0 and 2, with a mean square of 1.0. This indicates that there were no specific difficulties for students to judge or rank order the videos.

5.3.2 Reasoning

To investigate the two-dimensional structure for TSI and DI reasoning, an Exploratory Principal Axis Factoring on 23 items for TSI and 23 items for DI was performed. Based on the analysis, 7 items targeting TSI and 6 items targeting DI were deleted due to low factor loadings. Table 5 presents the results of this Exploratory Factor Analysis.

<< Table 5. Standardized factor loadings of the two-dimensional model indicating the professional vision of TSI and DI.>>

Second, a Confirmatory Factor Analysis was conducted to evaluate the stability of the two-dimensional structure. The final model comprises 16 items tapping TSI and 17 items tapping DI. The results show a good fit between the hypothesized model and the observed data, providing evidence for the two-dimensional model of TSI and DI. Table 6 summarizes the fit statistics for TSI and DI. The results suggest that all items load significantly onto the latent factor for TSI and DI.

<< Table 6. Goodness-of-fit indicators for the latent variables >>

Next, a reliability analysis was performed on the complete data set ($n=278$) to examine the internal consistency of the two latent factors. The two-dimensional factor for TSI and DI was found to be highly reliable, with a Cronbach's alpha (α) of .892 for TSI and .894 for DI, respectively.

5.4. Evidence based on relations to other variables (RQ4)

A regression analysis was done with pre-service teachers' ability to notice TSI and DI as dependent variables and their socio-demographic variables as independent variables. Pre-service teachers' noticing outcome was calculated by comparing their ranking of the videos to the aggregated ranking of experts, producing a misfit-score on pre-service teachers' ability to notice both TSI and DI. This misfit outcome calculates the difference between the expected outcome (expert ranking) to the observed ranking (student ranking) of each comparison. This results in a logit score per video clip which indicates the probability that a video clip will 'win'

compared to a reference video clip from the expert ranking. This misfit outcome describes the extent to which pre-service teachers' rank order corresponds to the rank order of experts. This means that expert ratings are used as a norm to measure pre-service teachers' ability to notice, since they are characterized by high level and integrated noticing abilities (Palmer et al., 2005).

Only two significant predictors were found in the relation between pre-service teachers' noticing and socio-demographic variables. First, a significant relationship was found between pre-service teachers' ability to notice TSI and their highest diploma ($F_{1,265} = 5.809$, $p < 0.05$). The regression coefficient associated with this predictor indicates that there is a positive relationship between the dependent and independent variable, $\beta_1 = 0.355$. Table 7 gives an overview of the different mean scores of pre-service teachers' noticing misfit for TSI and their highest diploma.

<< Table 7. Mean scores of the 'noticing' misfit for TSI and highest diploma >>

Pre-service teachers' misfit of TSI is lower with students that already earned a diploma at a college or university compared to students with only a secondary education diploma. This means that pre-service teachers with a higher education diploma are better able to notice TSI. These students are often older, more grown-up and have more experience in critical thinking and observing. This is in line with Sabers et al. (1991) who argue that more experienced teachers are better able to notice complex events in the classroom.

Another significant relationship was found between pre-service teachers' noticing misfit of TSI and pre-service teachers' study year ($F_{1,265} = 5.513$, $p < 0.05$). The regression coefficient that is associated with this predictor indicates that there is a negative relationship between the dependent and independent variable, $\beta_1 = -0.089$. Table 8 gives an overview of the different mean scores of pre-service teachers' noticing misfit for TSI in different study years.

<< Table 8. Mean scores of the ‘noticing’ misfit for TSI in different study years >>

Pre-service teachers’ noticing misfit decreases from the first year to the third year. This means that pre-service teachers in their last study year deviate less from experts compared to pre-service teachers in the first grade. This suggests that pre-service teachers are better able to notice events in the classroom when they are more competent in teaching. This is possibly because pre-service teachers’ competences increase during their teacher education training.

6. Discussion

6.1. General discussion

The growing diversity in pupil populations poses enormous challenges to the professional development of pre-service teachers, especially since they experience a theory-practice gap when entering the classroom (Korthagen & Kessels, 1999). Therefore, professional vision is introduced as a promising approach for the integration of theory and practice (Seidel & Stürmer, 2014). Professional vision is generally defined as the ability to notice relevant classroom situations and to reason about these situations on the basis of knowledge about teaching and learning (Sherin & van Es, 2009). This study aimed to develop a valid and reliable video-based measurement instrument to empirically assess pre-service teachers’ professional vision of inclusive classrooms (PVIC). PVIC focusses on two dimensions of effective inclusive classrooms: teacher-student interactions (TSI) and differentiated instruction (DI).

In the literature, noticing and reasoning are described as interconnected, which often results in one measurement approach for both. However, the ability to notice is considered crucial in the context of inclusive education, as it implies noticing new classroom events in terms of students’ needs (van Es & Sherin, 2002). In contrast to other research, this study

developed an individual measure of pre-service teachers' ability to notice classroom events in relation to inclusive classrooms. This measure is based on comparative judgement (CJ), which requires respondents to compare pairs of videos and decide which video is best in terms of TSI and DI (Thurstone, 1927). Our results confirmed the reliability and validity of the video-based measure of pre-service teachers' ability to notice events in inclusive classrooms (TSI and DI).

Moreover, the results of this study revealed differences in noticing abilities between students in different study years of their teacher education program. Students in their last year are better able to notice events in the classroom related to TSI compared to students in the first year. This difference in noticing abilities between students in different grades implies that professional vision can be developed during teacher education training. This is in line with findings of van Es and Sherin (2002), which indicate that pre-service teachers' ability to notice classroom events develops over time. Moreover, Stürmer et al. (2013a) argue that professional vision, and perception in particular, is a skill that can be learned. Therefore, teacher education should support pre-service teachers in learning to notice classroom events. Although such ability is something that can be developed during teacher education training, the results from this study shows that final year students' ability to notice classroom events related to TSI and DI is still different from experts. This indicates that professional vision is an ongoing process and can be further developed during teachers' careers through professional development initiatives.

Despite the relevance of noticing abilities within the domain of professional vision, the reasoning component of professional vision cannot be underestimated. In this study, video clips of TSI and DI combined with rating items served as a measure to assess pre-service teachers' reasoning. The videos include several facets of TSI and DI, however, we tested whether TSI and DI could each be described by a one-dimensional latent factor, operationalized as a two-dimensional model for TSI and DI. This is in contrast to other existing studies on measuring

professional vision of other teaching components (e.g. Seidel & Stürmer, 2014). For example, Seidel and Stürmer (2014) developed an instrument ('The Observer') to measure pre-service teachers' professional vision of goal clarity, teacher support, and learning climate. They distinguished three different aspects of reasoning: (a) description, (b) explanation, (c) prediction. Examples of these reasoning facets are the ability to link perceived events to conceptual knowledge (i.e. explanation) or the ability to predict the consequences of observed events in terms of student learning (i.e. prediction). However, the quantitative study of Seidel and Stürmer (2014) reported high correlations between the different aspects of reasoning, indicating that they are highly interrelated.

Therefore, we decided to test whether TSI and DI could each be described by one content factor instead of distinguishing between different aspects of reasoning. To test this two-dimensional model of professional vision (of TSI and DI), EFA, CFA and reliability analysis were performed so that one factor for each could be identified. Our results confirmed that TSI and DI can both be described by one content factor. Goodness of fit estimates were calculated, indicating good fit. This is similar to the recent work of Gold and Holodyski (2017), who found that PVCMM (professional vision of classroom management) could be best described by one content-related factor instead of their two-dimensional model addressing qualitative aspects of knowledge-based reasoning (description, interpretation).

6.2. Limitations and suggestions for future research

There are some limitations to the present study. First, the measurement instrument developed was tested for pre-service teachers in primary teacher education programs. Testing the validity and reliability of the instrument in a university context could further improve its applicability. Furthermore, it would be valuable to test the instrument for applicability among in-service teachers.

A second limitation to this study is that the measurement instrument was not validated in terms of consequences of use (Chan, 2014). Educational consequences refer to the effects that competence assessment have on learning and instruction. In the context of this study, more evidence is needed to clarify the extent to which the video-based instrument has an impact on learning and teaching.

A third limitation concerns our focus on two dimensions of inclusive classrooms, namely TSI and DI. Also other factors are particularly important in such classrooms. For example, literature on inclusiveness also discusses approaches such as teacher's diagnostic competencies (Dean & Burns, 2002), culturally relevant pedagogy (Ladson-Billings, 1995), exploiting diversity (Johnson & Johnson, 1989), collaborative teaming (King-Sears, Janney, & Snell, 2015) and issues of power and agency (De Vroey, Struyf & Petry, 2016) as they intersect with students' racial/ethnic and socio-economic background as crucial to effective inclusive teaching.

A fourth limitation of our study concerns the conceptualization of pre-service teachers' reasoning. Seidel and Stürmer (2014) distinguish three different levels within reasoning: (a) description, (b) explanation, and (c) prediction. In this study, we do not purposefully differentiate between these different aspects of reasoning to ensure an economically manageable assessment tool (i.e. time to fulfill the instrument).

As our results are promising, the availability of this video-based instrument opens perspectives for future research. First, the instrument can be used to study pre-service teachers' PVIC. Such studies would enable researchers to identify pre-service teachers' competence development regarding inclusive education. This information could help improve teacher education training regarding inclusive education. As such, the video-based instrument opens avenues for teacher educators to assess pre-service teachers' development of PVIC. Moreover, the instrument can be used as a learning tool for teacher education programs. Pre-service

teachers can learn by observing classroom videos as a form of professional learning and development. For example, pre-service teachers' struggling with differentiating the curriculum can improve their skills by observing teachers in a real and relevant inclusive learning environment. Being taught about a particular instructional theory does not mean it can be easily applied in the classroom. Observation, on the other hand, can bring actual practice to the forefront. This offers an opportunity for dialogue and reflection among pre-service teachers about the teaching behaviour in the video.

Furthermore, intervention studies can be set up in which the instrument is used in a pretest-posttest design. In this respect, the effectiveness and impact of teacher education trainings or interventions can be studied. This research is critical since studies show that teacher education programmes often fail to prepare student teachers for inclusive classrooms (McDonald, 2005).

Finally, in this study only demographic variables have been related to pre-service teachers' professional vision. In future studies, additional dimensions of teacher competencies could be added to investigate correlations between pre-service teachers' PVIC to other variables, such as pre-service teachers' beliefs and self-efficacy with regard to teaching in inclusive classrooms. Moreover, future research could delve into the association between pre-service teachers' PVIC on the one hand, and their behavior on the other hand. This would help to develop a comprehensive understanding of pre-service teachers' competencies regarding inclusive education.

6.3. Significance of research

Although research on pre-service teachers' professional vision has increased over the years, most studies have used small-scale qualitative designs. This study addresses this gap by developing a quantitative and holistic measurement instrument designed to empirically assess

pre-service teachers' PVIC. Moreover, by investigating PVIC, this study enhances empirical understanding of pre-service teachers' knowledge of inclusion and diversity in the classroom, which remains an 'undiscovered' domain.

7. References

- Author et al. (2016) [details removed for peer review]
- Author et al. (2017) [details removed for peer review]
- Author et al. (2018) [details removed for peer review]
- Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, 2(3), 451-462. <http://dx.doi.org/10.1177/014662167800200319>
- Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika*, 47(1), 105-113. <http://dx.doi.org/10.1007/BF02293856>
- Banks, J., Cochran-Smith, M., Moll, L., Richert, A., Zeichner, K., LePage, P., ... McDonald, M. (2005). Teaching diverse learners. In L. Darling-Hammond, & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 232-274). San Francisco: Jossey-Bass.
- Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, 12(2), 86–107. <http://dx.doi.org/10.1016/j.asw.2007.07.001>
- Baumrind, D. (1966). Effects of authoritative parental control on child behavior. *Child Development*, 37(4), 887-907. <http://dx.doi.org/10.2307/1126611>
- Beecher, M., & Sweeny, S. M. (2008). Closing the achievement gap with curriculum enrichment and differentiation: One school's story. *Journal of Advanced Academics*, 19(3), 502-530. <https://doi.org/10.4219/jaa-2008-815>
- Blomberg, G., Stürmer, K., & Seidel, T. (2011). How pre-service teachers observe teaching on video: Effects of viewers' teaching subjects and the subject of the video. *Teaching and Teacher Education*, 27(7), 1131-1140. <http://dx.doi.org/10.1016/j.tate.2011.04.008>
- Borko, H., & Livingston, C. (1989). Cognition and improvisation: Differences in mathematics instruction by expert and novice teachers. *American Educational Research Journal*, 26(4), 473-498. <http://dx.doi.org/10.3102/00028312026004473>

- Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3-4), 324-345. <http://dx.doi.org/10.2307/2334029>
- Bramley, T. (2007). Paired comparison methods. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 264–294). London: QCA.
- Bramley, T. (2015). *Investigating the reliability of adaptive comparative judgment*. Retrieved from Cambridge, UK <http://www.cambridgeassessmentjobs.org/Images/232694-investigating-the-reliability-of-adaptive-comparative-judgment.pdf>.
- Brophy, J. (2004). *Using video in teacher education*. Amsterdam, Netherlands: Elsevier
- Chan, E. K. H. (2014). Standards and guidelines for validation practices: development and evaluation of measurement instruments. In B. D. Zumbo, & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 9-24). Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-07794-9_2
- Cochran-Smith, M., & Power, C. (2013). Preparing teachers for diversity: A global issue. *Mondi Migranti*, 5(2), 81-98.
- Cook, T. D., Campbell, D. T., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston, MA: Houghton Mifflin.
- Coubergs, C., Struyven, K., Vanthournout, G., & Engels, N. (2017). Measuring teachers' perceptions about differentiated instruction: The DI-Quest instrument and model. *Studies In Educational Evaluation*, 53, 41-54. <http://dx.doi.org/10.1016/j.stueduc.2017.02.004>
- Dean, V. and Burns, M. (2002). Inclusion of Intrinsic processing difficulties in LD diagnostic models: A critical review. *Learning Disability Quarterly*, 25(3), 170-176. <http://dx.doi.org/10.2307/1511300>
- De Leeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology*, 100, 223-234. <http://dx.doi.org/10.1037/0022-0663.100.1.223>.
- De Vroey, A., Struyf, E., & Petry, K. (2016). Secondary schools included: A literature review. *International Journal of Inclusive Education*, 20(2), 109-135. <http://dx.doi.org/10.1080/13603116.2015.1075609>

- Englehart, J. M. (2009). Teacher–student interaction. In L. J. Saha, & A. G. Dworkin (Eds.), *International handbook of research on teachers and teaching* (pp. 711–722). Boston, MA: Springer
- Florian, L. (2008). Inclusion: Special or inclusive education: future trends. *British Journal of Special Education*, 35(4), 202-208. <http://dx.doi.org/10.1111/j.1467-8578.2008.00402.x>
- Gay, G. (2002). Culturally responsive teaching in special education for ethnically diverse students: Setting the stage. *International Journal of Qualitative Studies in Education*, 15(6), 613-629. <http://dx.doi.org/10.1080/0951839022000014349>
- Gold, B., & Holodynski, M. (2017). Using digital video to measure the professional vision of elementary classroom management: Test validation and methodological challenges. *Computers & Education*, 107, 13-30. <http://dx.doi.org/10.1016/j.compedu.2016.12.012>
- Goodwin, C. (1994). Professional vision. *American Anthropologist*, 96(3), 606-633. <http://dx.doi.org/10.1525/aa.1994.96.3.02a00100>
- Hair, J., Black, W., Babin, B., Anderson, R., & Tatham, R. (2006). *Multivariate data analysis* (6th ed.). Uppersaddle River, NJ: Pearson Prentice Hall
- Hall, T. (2002). *Differentiated instruction*. Wakefield, MA: National Center on Accessing the General Curriculum (NCAC).
- Hamre, B. K., & Pianta, R. C. (2001). Early teacher–child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development*, 72(2), 625-638. <http://dx.doi.org/10.1111/1467-8624.00301>
- Hattie, J. (2009). The black box of tertiary assessment: An impending revolution. In L.H. Meyer, S. Davidson, H. Anderson, R. Fletcher, P. Johnston, & M. Rees (Eds.), *Tertiary assessment & higher education student outcomes: Policy, practice & research*. Wellington, New Zealand: Ako Aotearoa.
- Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, 37(2), 1-19. <http://dx.doi.org/10.1007/BF03216919>
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: a Multidisciplinary Journal*, 6(1), 1-55. <http://dx.doi.org/10.1080/10705519909540118>
- Johnson, D. W., & Johnson, R. T. (1989). *Cooperation and competition: Theory and research*. Edina, MN: Interaction Book Company.

- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130-144. <http://dx.doi.org/10.1016/j.edurev.2007.05.002>
- Kelchtermans, G., & Ballet, K. (2002). The micropolitics of teacher induction. A narrative-biographical study on teacher socialisation. *Teaching and Teacher Education*, 18(1), 105-120. [http://dx.doi.org/10.1016/S0742-051X\(01\)00053-1](http://dx.doi.org/10.1016/S0742-051X(01)00053-1)
- Kersting, N. (2008). Using video clips of mathematics classroom instruction as item prompts to measure teachers' knowledge of teaching mathematics. *Educational and Psychological Measurement*, 68(5), 845-861. <http://dx.doi.org/10.1177/0013164407313369>
- Kersting, N. B., Givvin, K. B., Sotelo, F. L., & Stigler, J. W. (2010). Teachers' analyses of classroom video predict student learning of mathematics: Further explorations of a novel measure of teacher knowledge. *Journal of Teacher Education*, 61(1-2), 172-181. <http://dx.doi.org/10.1177/0022487109347875>
- King-Sears, M. E., Janney, R., & Snell, M. (2015). *Collaborative teaming*. Baltimore: Brookes.
- Kinzer, C. K., & Risko, V. J. (1998). Multimedia and enhanced learning: Transforming preservice education. In D. Reinking, M. C. McKenna, L. D. Labbo, & R. D. Kieffer (Eds.), *Handbook of literacy and technology: Transformations in a post-typographic world*. New York, NY: Routledge.
- König, J., Blömeke, S., Klein, P., Suhl, U., Busse, A., & Kaiser, G. (2014). Is teachers' general pedagogical knowledge a premise for noticing and interpreting classroom situations? A video-based assessment approach. *Teaching and Teacher Education*, 38, 76-88. <http://dx.doi.org/10.1016/j.tate.2013.11.004>
- Korthagen, F., & Kessels, J. (1999). Linking theory and practice: changing the pedagogy of teacher education. *Educational Researcher*, 28(4), 4-17. <http://dx.doi.org/10.3102/0013189X028004004>
- Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, 32(3), 465-491. <http://dx.doi.org/10.3102/00028312032003465>
- La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the prekindergarten year. *The Elementary School Journal*, 104(5), 409-426. <http://dx.doi.org/10.1086/499760>
- Laming, D. (2003). *Human judgment: The eye of the beholder*. Andover, MA: Cengage Learning EMEA.

- Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review*, 66(2), 81.
<http://dx.doi.org/10.1037/h0043178>
- McDonald, M. A. (2005). The integration of social justice in teacher education: Dimensions of prospective teachers' opportunities to learn. *Journal of Teacher Education*, 56(5), 418-435. <http://dx.doi.org/10.1177/0022487105279569>
- Mills, C., & Ballantyne, J. (2010). Pre-service teachers' dispositions towards diversity: Arguing for a developmental hierarchy of change. *Teaching and Teacher Education*, 26(3), 447-454. <http://dx.doi.org/10.1016/j.tate.2009.05.012>
- Mitchell, D. (2014). *What really works in special and inclusive education: Using evidencebased teaching strategies*. London and New York: Routledge.
- Paas, F. G., Van Merriënboer, J. J., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills*, 79(1), 419-430.
<http://dx.doi.org/10.2466/pms.1994.79.1.419>
- Palmer, D. J., Stough, L. M., Burdenski, T. K., & Gonzales, M. (2005). Identifying teacher expertise: An examination of researchers' decision making. *Educational Psychologist*, 40(1), 13-25. http://dx.doi.org/10.1207/s15326985ep4001_2
- Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38(2), 109–119. <http://dx.doi.org/10.3102/0013189X09332374>
- Pianta, R. C. (1999). *Enhancing relationships between children and teachers*. Washington, DC: American Psychological Association.
- Pollitt, A. (2012a). Comparative judgement for assessment. *International Journal of Technology and Design Education*, 22(2), 157-170. <http://dx.doi.org/10.1007/s10798-011-9189-x>
- Pollitt, A. (2012b). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 19(3), 281-300.
<http://dx.doi.org/10.1080/0969594X.2012.665354>
- Reis, S. M., McCoach, D. B., Little, C. A., Muller, L. M., & Kaniskan, R. B. (2011). The effects of differentiated instruction and enrichment pedagogy on reading achievement in five elementary schools. *American Educational Research Journal*, 48(2), 462-501.
<http://dx.doi.org/10.3102/0002831210382891>
- Roorda, D. L., Koomen, H. M., Spilt, J. L., & Oort, F. J. (2011). The influence of affective teacher–student relationships on students' school engagement and achievement: A

- meta-analytic approach. *Review of Educational Research*, 81(4), 493-529.
<http://dx.doi.org/10.3102/0034654311421793>
- Sabers, D. S., Cushing, K. S., & Berliner, D. C. (1991). Differences among teachers in a task characterized by simultaneity, multidimensional, and immediacy. *American Educational Research Journal*, 28(1), 63-88.
<http://dx.doi.org/10.3102/00028312028001063>
- Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, 34(2), 159-179.
<http://dx.doi.org/10.1080/02602930801956059>
- Seidel, T., & Stürmer, K. (2014). Modeling and measuring the structure of professional vision in preservice teachers. *American Educational Research Journal*, 51(4), 739-771.
<http://dx.doi.org/10.3102/0002831214531321>
- Sherin, M. G. (2001). Developing a professional vision of classroom events. In T. Wood, B. S. Nelson, & J. Warfield (Eds.), *Beyond classical pedagogy: Teaching elementary school mathematics* (pp. 75–93). Hillsdale, NJ: Erlbaum
- Sherin, M. G. (2007). The development of teachers' professional vision in video clubs. *Video Research in the Learning Sciences*, 383-395.
- Sherin, M. G., & Han, S. Y. (2004). Teacher learning in the context of a video club. *Teaching and Teacher Education*, 20(2), 163-183. <http://dx.doi.org/10.1016/j.tate.2003.08.001>
- Sherin, M. G., & Van Es, E. A. (2009). Effects of video club participation on teachers' professional vision. *Journal of Teacher Education*, 60(1), 20-37.
- Stanford, B., & Reeves, S. (2009). Making it happen: Using differentiated instruction, retrofit framework, and universal design for learning. *Teaching Exceptional Children Plus*, 5(6), 1–9. Retrieved from <http://escholarship.bc.edu/education/tecplus/vol5/iss6/art4>
- Stokking, K., Leenders, F., De Jong, J., & Van Tartwijk, J. (2003). From student to teacher: Reducing practice shock and early dropout in the teaching profession. *European Journal of Teacher Education*, 26(3), 329-350.
<http://dx.doi.org/10.1080/0261976032000128175>
- Stürmer, K., Könings, K. D., & Seidel, T. (2013a). Declarative knowledge and professional vision in teacher education: Effect of courses in teaching and learning. *British Journal of Educational Psychology*, 83(3), 467-483. <http://dx.doi.org/10.1111/j.2044-8279.2012.02075.x>

- Stürmer, K., Seidel, T., & Schäfer, S. (2013b). Changes in professional vision in the context of practice. *Gruppendynamik und Organisationsberatung*, 44(3), 339-355.
<http://dx.doi.org/10.1007/s11612-013-0216-0>
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson/Allyn & Bacon.
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4), 273–286. <http://dx.doi.org/10.1037/h0070288>
- Tomlinson, C. A. (1999). Mapping a route toward differentiated instruction. *Educational Leadership*, 57, 12–17. Retrieved from http://pdonline.ascd.org/pd_online/diffinstr/el199909tomlinson.html
- Tomlinson, C. A. (2001). Differentiated instruction in the regular classroom: What does it mean? How does it look? *Understanding our gifted*, 14(1), 3-6.
- Tomlinson, C. A. (2004). Differentiation in diverse settings: A consultants experiences in two similar school districts. *The School Administrator*, 61(7), 28.
- United Nations. (2007). *Convention on the Rights of Persons with Disabilities*. Retrieved from <http://www.un.org/disabilities/convention/conventionfull.shtml>.
- Valiandes, S. (2015). Evaluating the impact of differentiated instruction on literacy and reading in mixed ability classrooms: Quality and equity dimensions of education effectiveness. *Studies in Educational Evaluation*, 45, 17-26.
<http://dx.doi.org/10.1016/j.stueduc.2015.02.005>
- van der Schaaf, M., Stokking, K., & Verloop, N. (2005). Cognitive representations in raters' assessment of teacher portfolios. *Studies in Educational Evaluation*, 31(1), 27-55.
<http://dx.doi.org/10.1016/j.stueduc.2005.02.005>
- van Es, E. A., & Sherin, M. G. (2002). Learning to notice: Scaffolding new teachers' interpretations of classroom interactions. *Journal of Technology and Teacher Education*, 10(4), 571-596.
- van Es, E. A., & Sherin, M. G. (2008). Mathematics teachers' "learning to notice" in the context of a video club. *Teaching and Teacher Education*, 24(2), 244-276.
<http://dx.doi.org/10.1016/j.tate.2006.11.005>
- Van Petegem, K., Aelterman, A., Van Keer, H., & Rosseel, Y. (2008). The influence of student characteristics and interpersonal teacher behaviour in the classroom on student's wellbeing. *Social Indicators Research*, 85(2), 279-291.
<http://dx.doi.org/10.1007/s11205-007-9093-7>

Voss, T., Kunter, M., & Baumert, J. (2011). Assessing teacher candidates' general pedagogical/psychological knowledge: Test construction and validation. *Journal of Educational Psychology*, 103(4), 952–969. <http://dx.doi.org/10.1037/a0025125>

8. Appendix

List of tables

Table 1. Rank order of the video clips targeting TSI and DI

Table 1

Rank order of the video clips targeting TSI and DI

Ranking TSI			Ranking DI		
Script	TrueScore	seTrueScore	Script	TrueScore	seTrueScore
Video 3	-4,08	1,75	Video 3	-4,31	1,70
Video 10	-0,65	0,54	Video 1	-1,60	0,60
Video 1	-0,55	0,48	Video 5	-1,22	0,57
Video 4	-0,51	0,51	Video 15	-0,61	0,53
Video 9	-0,44	0,48	Video 10	-0,41	0,51
Video 6	-0,31	0,45	Video 9	-0,39	0,56
Video 15	-0,30	0,50	Video 2	-0,09	0,53
Video 8	-0,20	0,49	Video 11	0,15	0,49
Video 5	-0,15	0,54	Video 4	0,40	0,49
Video 2	0,64	0,50	Video 12	0,53	0,53
Video 11	0,64	0,50	Video 14	0,96	0,59
Video 14	0,76	0,61	Video 8	1,11	0,49
Video 13	0,85	0,51	Video 7	1,20	0,52
Video 12	0,95	0,54	Video 13	1,21	0,54
Video 7	2,12	0,72	Video 6	1,26	0,47

Table 2. Judge fit measures (Outfit and Infit)

Table 2

Judge fit measures (Outfit and Infit)

TSI			DI		
Judge	Outfit	Infit	Judge	Outfit	Infit
1	1.69	1.39	1	0.65	0.76
2	0.41	0.45	2	1.28	1.29
3	0.55	0.63	3	0.83	0.91
4	1.02	1.02	4	1.24	1.26
5	0.52	0.61	5	0.46	0.53
6	1.06	1.09	6	0.86	0.98
7	0.39	0.45	7	0.52	0.81
8	1.56	1.56	8	1.05	1.05
9	0.71	0.90	9	0.55	0.93
10	0.64	0.84	10	0.77	0.93
11	1.51	1.54	11	0.87	0.93
12	0.37	0.51	12	0.61	0.75
13	1.39	1.49	13	1.18	1.04
14	0.92	0.94	14	0.59	0.68
15	1.24	1.15	15	1.86	1.64
16	2.34	1.66	16	0.24	0.30
17	0.76	0.76	17	1.85	1.85
18	6.01	2.56	18	4.65	3.25
19	0.75	0.86	19	0.36	0.58
20	0.78	0.81	20	1.68	1.57
21	0.70	0.96	21	1.04	1.52
22	0.76	0.91	22	0.46	0.63
23	2.67	2.23	23	0.45	0.60
24	1.78	1.73	24	2.44	1.82
25	0.88	0.92	25	0.61	0.78
26	0.47	0.56	26	0.53	0.71
27	0.47	0.65	27	0.55	0.81
28	0.48	0.70	28	0.36	0.51
29	0.20	0.25	29	0.60	0.78
30	1.09	1.21	30	0.60	0.75
31	0.31	0.36	31	0.24	0.24
32	0.34	0.39	32	0.81	0.85
33	0.48	0.50	33	2.00	1.88
34	0.61	0.84	34	0.57	0.82

Note. Bold numbers indicate misfitting judges.

Table 3. Judge Infit summary statistics

Table 3

Judge Infit summary statistics

TSI			DI		
	Outfit	Infit		Outfit	Infit
Mean	1.003	1.003	Mean	0.999	0.998
SD	0.118	0.115	SD	0.143	0.134
Lower	0.766	0.772	Lower	0.712	0.730
Upper	1.239	1.234	Upper	1.286	1.267

Note. Upper and lower represents the boundary of a 95% confidence interval.

Table 4. Representation Infit summary statistics

Table 4

Representation Infit summary statistics

TSI			DI		
	Outfit	Infit		Outfit	Infit
Mean	1.002	1.000	Mean	1.002	1.000
SD	0.068	0.012	SD	0.064	0.016
Lower	0.865	0.977	Lower	0.875	0.968
Upper	1.139	1.024	Upper	1.130	1.031

Note. Upper and lower represents the boundary of a 95% confidence interval.

Table 5. Standardized factor loadings of the two-dimensional model indicating the professional vision of TSI and DI

Table 5

Standardized factor loadings of the two-dimensional model indicating the professional vision of TSI and DI

TSI		DI	
Item	Factor loading	Item	Factor loading
The students receive encouragement	0.571	There are flexible group formations in the classroom	0.391
The teacher stimulates interaction between students	0.346	The students receive feedback on their learning process	0.562
The teacher pays attention to social-emotional needs of students	0.414	The teacher uses a clear language	0.523
The teacher uses humor in his/her communication with students	0.278	The lesson is structured	0.595
The students have a voice in the classroom	0.359	The students are working actively during the lesson	0.527
The teacher radiates calmness	0.548	Students work together through cooperative learning	0.495
The students receive individual attention	0.41	The teacher takes into account the learning pace of the students	0.604
The teacher pays attention to students' talents	0.577	The teacher pays attention to the level and capabilities of individual students	0.578
The teacher stimulates self-reflection and self-evaluation	0.472	There are variations in assignments and teaching methods	0.629
The teacher establishes an effective class management	0.618	The teacher addresses the interests of students	0.576
There is positive affection between the teacher and the students	0.542	The classroom style supports an inclusive learning environment	0.395
The teacher uses clear communication	0.666	Students can use technology in the classroom	0.383
The students receive support	0.583	The teacher uses a playful approach	0.607
The teacher is available to the students	0.648	The teacher uses activating teaching methods	0.591
Students receive feedback on their learning process	0.594	The students receive 'choice options' in the classroom	0.457
The teacher poses high expectations	0.316	The students have a voice in the classroom	0.612
The students are involved during the lesson	0.649	The teacher acts as a coach	0.594
The students are treated equally by the teacher	0.552	The students are able to use resources (e.g. cubes)	0.524
The teacher gives responsibilities to the students	0.525	The teacher pays attention to different learning styles and learning profiles of students	0.564
There is a calm atmosphere in the classroom	0.547	The students receive support from the teacher	0.602
There is a safe classroom environment	0.678	The students are motivated	0.556
The students can be themselves in the classroom	0.739	The students are working autonomously	0.455
The students receive compliments	0.636	The students have different responsibilities, functions and roles during cooperative learning assignments	0.522

Note. Factor loading in bold when item retained in component.

Table 6. Goodness-of-fit indicators for the latent variables

Table 6

Goodness-of-fit indicators for the latent variables

Latent variable	df	χ^2	CFI	TLI	RMSEA	SRMR	N items
TSI	120	1304,62***	0.917	0.901	0.060	0.052	16
DI	136	1535,96***	0.922	0.908	0.058	0.054	17

Note. *** $p \leq .001$

Table 7. Mean scores of the ‘noticing’ misfit for TSI and highest diploma

Table 7

Mean scores of the ‘noticing’ misfit for TSI and highest diploma

Independent variable (diploma)	Mean (SD)
Secondary education diploma	1.71 (0.73)
Higher education diploma	1.35 (0.56)

Table 8. Mean scores of the ‘noticing’ misfit for TSI in different study years

Table 8

Mean scores of the ‘noticing’ misfit for TSI in different study years

Independent variable (study year)	Mean (SD)
First year	1,76 (0.76)
Second year	1,68 (0.60)
Third year	1,57 (0.81)
Individualized curriculum (combination of second and third year)	1,50 (0.59)

List of figures

Figure 1. Schematic overview of the development and validation of the instrument

Figure 2. Representation Infit of TSI

Figure 3. Representation Infit of DI