Measuring teaching skill of South Korean teachers in secondary education

van de Grift, Wim; Lee, Okhwa; Chun, Seyeoung

Published in:
Effective Teaching Around the World

DOI:
10.1007/978-3-031-31678-4_8

Chapter 8
Measuring Teaching Skill of South Korean Teachers in Secondary Education: Detecting a Teacher's Potential Zone of Proximal Development Using the Rasch Model

Wim van de Grift, Okhwa Lee, and Seyeoung Chun

Abstract Many observation instruments are in use to make the skills of teachers visible. These tools are used for assessment, for guidance and coaching, and for policy-oriented research into the quality of education. Depending on the purpose of use of an observation instrument, we not only need more observations about the same teacher, but the observation instrument must also meet higher psychometric requirements. Observation instruments only used to assess sample characteristics, such as the mean and dispersion, require less stringent psychometric requirements than observation instruments that are used to assess individuals. For assessing sample characteristics, it is also not necessary to do more than one observation with each respondent. Observation instruments used for individual assessments that lead to high stake decisions should meet the highest psychometric requirements possible. We can slightly mitigate the psychometric norms attached to an observation tool that is only used for guidance and coaching on the condition that the observed teacher explicitly informed that the observed lesson was representative and that this lesson offered sufficient opportunities to demonstrate all the skills the teacher has. Nevertheless, there are also additional requirements that must be met by observation instruments that are used for guidance and coaching. For good guidance and coaching, it is usually not very useful to tell an observed teacher only what went right or wrong. Teachers need concrete instructions to be able to improve. Many things that have not gone very well are often (and sometimes far) out of the reach of the teacher being observed. Coaching skills that are beyond the reach of the observed person will lead to disappointment rather than to the desired effect. The important thing in

W. van de Grift (✉)
Emeritus Professor of Education, University of Groningen, Groningen, The Netherlands

O. Lee
Department of Education, Chungbuk National University, Cheongju, South Korea

S. Chun
Department of Education, Chungnam National University, Daejeon, South Korea

R. Maulana et al. (eds.), *Effective Teaching Around the World*,
https://doi.org/10.1007/978-3-031-31678-4_8

good guidance and coaching is to ensure that the observed teacher is going to take that very step, that is within his reach, but that he has not just set. Then, of course continue with the next steps, leading to incremental progress. For this, we need to have an insight into the successive difficulty of the different skills of teachers. In the past, we gained some experience with the use of the Rasch model to gain an insight into the successive level of difficulty in the actions of Dutch teachers working in elementary education. These studies are all done with the **I**nternational **C**omparative **A**nalysis of **L**earning and **T**eaching (ICALT) observation instrument. In this chapter, we are trying to make a next step by using the Rasch model for detecting the zone of proximal development of the observed teachers. Another new element in this study is the following: Until now, the ICALT observation instrument has been used mainly in (the culture of) European schools. In this chapter, we focus on Asian secondary education, as it takes shape in South Korea.

Keywords Teaching skill · Zone of proximal development · Rasch model

1 Introduction

Many observation instruments are in use to make the skills of teachers visible (cf. Bell et al., 2018; Dobbelaer, 2019). These tools are used for assessment, for policy-oriented research into the quality of education and for guidance and coaching. For good guidance and coaching, it is usually not very useful to tell an observed teacher only what went right or wrong. Teachers need concrete instructions to be able to improve. Many things that have not gone very well are often (and sometimes far) out of the reach of the teacher being observed. Coaching skills that are beyond the reach of the observed person will lead to disappointment rather than to the desired effect. The important thing in good guidance and coaching is to ensure that the observed teacher is going to take the next step, within his or her reach, that s/he has not yet reached. After that the following steps can be taken, leading to incremental growth. For this, we need to have an insight into the successive difficulty of the different skills of teachers. In this article, we use the Rasch model (Rasch, 1960, 1961) for detecting the potential zone of proximal development of the observed teachers.

The observation instrument we will use is the ICALT instrument. The ICALT observation instrument was developed between 1989 and 1994 for primary education and was initially used by the Education Inspectorate (Van de Grift & Lam, 1998). The instrument, which has also been used by other European education inspectorates (cf. Van de Grift, 2007, 2014), currently has a version consisting of six Likert scales. The six Likert scales contain 32 high inferential items and 120 low inferential examples of good practice. The 152 high and low inferential items are all based on reviews of a large number of studies on the effectiveness of education on student achievement (cf. the references). The 32 high inferential items are the core of the observation instrument. The raw score on the instrument is simply the sum score on these 32 items. These 32 items have an abstract or high inferential

character. An example of a high inferential item is "… promotes learners' self-confidence". In the observation instrument, every high inference item is accompanied by several low inference items. For example, low inferential items that belong to the high inferential item above are "…gives positive feedback on questions and remarks from learners", "…compliments learners on their work", and "…acknowledges the contributions that learners make". The actions in these low inferential items are coded as simply observed or not observed during a lesson. The 120 low inferential items are used in different situations. During the training of the observers, the low inferential items are used to explain the height of the score on the 32 high inferential items. If the score on a high inferential item is low, the scores on the corresponding low inferential items should also be low. When an observer gives a low score on a high inferential item, the scores on the corresponding low inferential items should also be low. Also, the scores of the low inferential items are used when coaching the observed teacher. It has little practical value to use the abstract and high inferential items for that. It is more informative for the observed teacher when the advice based on the low inference items is: 'evaluate whether the lesson aims have been reached' and 'offer weaker learners extra study and instruction time', than the advice based on the high inference item 'adjust instructions and learner processing to inter-learner differences'. The low inference items indicate more concretely what the observed teacher should do. (For more details, see the appendix with the ICALT instrument.)

The first three Likert scales concern the basic skills of teaching: creating a safe and stimulating educational learning climate, organizing the lesson efficiently, and providing clear and structured instruction. The other three Likert scales concern the advanced teaching skills: giving an intensive and activating lesson, tailoring instruction and processing to differences between students and teaching students learning strategies. An observed teacher masters the observed activities from a scale to a more than sufficient extent when the score in that domain is higher than 2.5. (Then ≥65% of the items is scored sufficient.) The six domains of the ICALT instrument show a hierarchical order with increasing difficulty (Van de Grift, 2021). The items from some domains of the observation instrument are relatively easy for teachers to master, for example creating a safe and stimulating learning environment. Other domains are relatively difficult for teachers, for example differentiated teaching and teaching students learning strategies. This hierarchical order in the domains of the ICALT instrument made us wonder whether this order could also be found in the individual items. Therefore we studied in a sample of 400 teachers working with 6–12-year-old students the question whether the 32 individual items meet the requirements of the dichotomous Rasch model. We found a reliable Rasch scale with 31 items for measuring the teaching skills. The simplest items concerned basic skills such as creating a safe learning environment, efficient classroom management and clear and structured instruction. The slightly more difficult items concerned activating learners. The items concerning differentiated instruction were clearly more difficult. The most difficult items were those related to teaching students how to learn. The scale is suitable for distinguishing six zones that give an indication of the zone of proximal development of an observed teacher (Van de Grift et al., 2019).

In 2008, we began studies to determine whether the ICALT observation instrument could also be used reliably and validly with student teachers and beginning teachers in secondary education (Maulana et al., 2015, 2016). In 2015, we started international comparisons of the quality of teaching in various non-Western countries, such as South Korea (Van de Grift et al., 2017) and South Africa (De Jager et al., 2017). In the same period we started analyses in which we investigated whether the Rasch model was applicable to the pedagogical didactic behaviour of teachers in secondary education (Van de Grift et al., 2014; Van der Lans et al., 2017, 2018). The order of the difficulty of the 31 items that fitted the Rasch model appeared to be more or less the same for teachers in secondary education as it was for teachers in basic education. The simplest items concerned basic skills such as creating a safe learning environment, efficient classroom management and clear and structured explanations. The slightly more difficult items concerned activating students. Clearly more difficult were the items about teaching pupils how to learn. In contrast to the situation in primary education, the items that concerned the provision of differentiated instruction proved to be the most difficult in secondary education. The fact that the items providing differentiated instruction were the most difficult for teachers in secondary education probably has to do with the fact that students in primary education are not sorted by skills level as they are in secondary education. In the present publication, we investigate whether this order item difficulties is maintained among secondary school teachers from a completely different culture, the Asian culture.

2 Theoretical and Empirical Background

In this section, we will introduce the idea of "zone of proximal development". After that we will go into some theoretical and empirical backgrounds of

- the relationships between teaching skills and students' learning gain
- the trainability of teaching skills, and
- the relationships between the growth of teaching skills and growth in students' learning gain.

2.1 The Idea of the "Zone of Proximal Development"

Many years ago, the concept "zone of proximal development" was introduced by Vygotsky (1930). Vygotsky was interested in the ontogenetic (and phylogenetic) development of thinking and speech. In his conception the zone of proximal development relates to the difference between what a child can achieve independently

(the so-called actual level of development) and what a child can achieve with guidance and encouragement from a skilled person (the so-called zone of proximal development). Over the years, there has been a lot of discussion about the interpretation of the work of Vygotsky. Part of this discussion has to do with the correct translation of several concepts from Russian into western languages (Lompscher & Rückriem, 2002).

Without going in too much detail, we will interpret in this study this concept as an area of learning that is very near to the actual level of skill of a person. We suppose that students, taught in their zone of proximal development, will learn faster and more effective, than students who are asked to do things that are (too) difficult for them. For example in the teaching of pupils we do not start with an explanation of multiplication before the idea of repeated addition is well understood. We do not start reading comprehension before the child can perform the technical reading process. The zone of proximal development helps to properly determine the upper limit of what a person is already capable of. This is the starting point for feedback and deliberate training and behavioural practice with the aim to raise the upper level of performance to a (slightly) higher level of the proximal development.

In this study, we are interested in the professional development of teachers. The professional development of teachers differs from ontogenetic theories, but there are related matters. An important related matter is the fact that mastering basic knowledge and skills of teaching is conditional for the mastering of more complex knowledge and skills. Research showed that teaching skills associated with differentiation in teaching are more difficult than those related to activating students are. Activating students is more difficult compared to classroom management skills (Van de Grift et al., 2014, 2019; Maulana et al., 2016). Mastering of the basic skills of teaching seems to be conditional for being able to master other more complex teaching skills. Teachers still having problems with classroom management should not be coached in skills to activate students. They should first be helped with their classroom management problems. The same is for teachers who have problems with giving clear explanations; they are not yet ready for differentiated instruction. They must first learn to explain clearly and in a structured way before they can help pupils with specific learning needs.

The one who is in charge of the guidance or coaching of teachers should consider not only the actual level of development but also the zone of proximal development of teachers. The difference between the teachers actual level of development and the level of performance that he or she achieves in collaboration with the coach, defines the zone of proximal development. Coaching of teachers is maximally productive only when it occurs at a certain point in the zone of proximal development. The zone of proximal development determines the domain of improvements that are accessible to the teacher.

However, determining the zone of proximal development of teachers' teaching skills is not a simple and easy task. It is therefore not surprising that the knowledge about this in the current literature is very scarce.

2.2 Teaching Skills and Students' Learning Gains

Between 1983 and 2008 several reviews are published are published about the relationships between teaching behaviour and student achievement. These research reviews make clear that several teaching behaviours are indeed related to student achievement and learning gains: Setting targets, offering sufficient learning and instruction time, monitoring students' achievements, creating special measures for struggling students, establishing a safe and stimulating educational climate, organizing efficient classroom management, giving clear and structured instruction, organizing intensive and activating teaching, differentiating instruction, and teaching learning strategies. Good readable summaries of various reviews of these studies can be found in Marzano (2003) and Hattie (2009, 2012). More detailed information can be found in the references of this chapter. Several econometric studies indicated also that better teachers have students with more learning gains (Hanushek & Rivkin, 2010; Kane & Staiger, 2008; Rivkin et al., 2005).

Some of these teaching behaviours are susceptible to observation; other behaviours have to be found through interviews. In this study, we concentrate on the issues that can be observed by external observers in classes: establishing a safe and stimulating educational climate, organizing efficient classroom management, giving clear and structured instruction, organizing intensive and activating teaching, adapting instruction, and teaching learning strategies.

An important question is: How malleable and trainable is this behavior? The following paragraph deals with this.

2.3 Trainability of Teaching Skills

Kraft et al. (2018) reviewed 60 American, Canadian, and Chilean empirical studies on the effects of the coaching of teachers and conducted meta-analyses to estimate the mean effect of coaching programs on teachers' instructional practice. There are 55 American, and 5 Canadian and Chilean empirical studies. The mean effect across 60 studies, employing causal research designs was a pooled effect size of 49% of a standard deviation on teachers' instructional practice.

Van den Hurk et al. (2016) studied 110 teachers, working in Dutch elementary education. These teachers had been coached based on a lesson observed with them. After the coaching these teachers showed a skill growth, on several observed aspects of teaching. They found for creating a safe and stimulating climate a growth of 29% of a standard deviation; for efficient classroom management a growth of 37%; for clear and structured instruction a growth of 62%; for activating students 76%; for teaching learning strategies 71%, and for differentiation they found a growth of 51% of a standard deviation. These Dutch results are in agreement with the average effect size found in the American, Canadian and Chilean studies found by Kraft et al. (2018).

The following section handles the relationship between growth in teaching skills and (extra) growth in student achievements.

2.4 Growth of Teaching Skills and Students' Learning Gains

Kraft et al. (2018) found a mean effect of growth in teaching on student achievement of 18% of a standard deviation. Effect sizes were larger (34% of a standard deviation) in smaller programs than in larger programs (10% of a standard deviation). Therefore, it seems that an average growth of 49% of a standard deviation on teachers' instructional practice in USA, Canada and Chile goes along with an average growth of 18% in students' academic achievement.

In several small-scale experiments done in Dutch elementary education (Houtveen & Van de Grift, 2007a, b; Houtveen et al., 2004, 2014) an average effect size of 64% of a standard deviation was found in the growth of teaching skills by specially observed and coached teachers. The students in the experimental groups of these experiments had an extra learning gain of 45% of a standard deviation for decoding, 38% for comprehensive reading and 52% for mathematics. Therefore in these studies, a growth of almost two third of a standard deviation in teaching skill goes along with a growth of student achievement of almost half a standard deviation.

3 Aim of This Study

We have already seen that 31 of the 32 items of the ICALT observation instrument have a hierarchical order. This hierarchical order is very important for accurately tracing the zone of close development of an observed teacher. In this study, we investigate whether the order of item difficulty found among Dutch secondary school teachers is maintained among secondary school teachers from a totally different culture, the South Korean culture.

4 Method

4.1 Sample Characteristics

In South Korea, the teaching skills of a sample of 375 teachers working in 26 secondary schools in the regions Deajeon, Chungnam, Cheongju, and Chungbuk were observed in one real life lesson by specially trained observers. Teachers in the sample were recruited by their voluntary participation in the research project. They were introduced about ICALT and invited by the observers who had been trained with ICALT tool. These data were previously used in Van de Grift et al. (2017). These 375 teachers taught 25 different subjects. The teachers had, on average, 11 years of teaching experience. About 51% of the teachers were female. The average class size was 29 students (see Table 8.1 for more detailed information).

Table 8.1 Sample characteristics (n = 375 teachers)

Subject	% teachers		Years of experience	Class size
Language	17.9	Mean	11.32	29.12
English	20.5	Standard dev.	9.59	7.17
Beta (math, science, information science and so on)	34.7	Minimum	0	10
Else	26.9	Maximum	38	42

This sample of 375 teachers is large enough to estimate proportions in the population of the regions Deajeon, Chungnam, Cheongju, and Chungbuk with a precision of 5% and a confidence interval of 95% (cf. Kirby et al., 2002). These teachers were observed by 40 trained observers; 14 observers observed <5 lessons and 26 observers observed 9–33 lessons. The observers had on average almost 26 years of experience as a teacher.

4.2 Translation of the Observation Instrument and Training of Observers

4.2.1 Translation of the Observation Instrument

The English version instrument was firstly translated into Korean by one of the Korean authors of this chapter. This first translation was back-translated into English from Korean by a native English teacher who were teaching English at a secondary school in South Korea. The back-translated English instrument was examined by both the Dutch ICALT research team and the original Korean translator. Then the Korean version of the instrument had been finalized.

4.2.2 Training of Observers

The observers who participated in this study were trained over the course of two full days. The training involved explanations of the theoretical, empirical and practical backgrounds of the observation instrument, practices with observing two videotaped lessons, and a discussion about how to evaluate teaching behaviours using the associated scoring procedures. Both videotaped lessons were in English.

During the presentation of both video tapes, the observers had to score both high and low inferential items.

After presenting the consensus results of the first video to the observers, discussions were organized between observers who did not agree on one more items. The scores on the low inferential items were used to reach consensus on the scoring of the high inferential items. The scores on the low inferential items are the ‘arguments’ for the score on the high inferential items. These arguments are used during

the discussions. Furthermore, the consensus within the observers and the expert norm was compared, with a cut-off of 0.70. In the current group, the consensus level was 0.82. Only certified observers were invited to observe classrooms.

4.3 Interrater Reliability

It sounds quite simple and reasonable: observers observing the same lesson should reach, working with the same observation instrument, the same conclusion. In order to reach this goal observers should be very consistent with each other in their judgments. Consistency alone is not enough. Observers must also have a high degree of agreement in their scores. Their amount of consensus must also be higher than can be achieved only by guessing.

Several statistics are used to determine whether observers interpret the same event in the same way. Ten Hove et al. (2018) showed that working with the same data, different coefficients show different results. These partially overlapping statistics all have their own merits and advantages, and problems and disadvantages. That is why we use several statistics in this study to obtain an indication of interrater reliability. The results we found with three of these statistics are presented in Table 8.2.

4.3.1 Intra-Class Correlation

We used the intra-class correlation coefficient (ICC; Hallgren, 2012) in order to assess the degree that observers showed consistency in their ratings of teaching skill across the items of the ICALT-scale. According to Cicchetti (1994) the interrater reliability is poor for ICC values less than .40, fair for values between .40 and .59, good for values between .60 and .74, and excellent for values between .75 and 1.0. During the observation training, we used the two video lessons: an English lesson and a geography lesson.

For the English lesson, an ICC of .90 was found, indicating that the observers had a high degree of consistency in their judgements. Studying changes in the ICC when one or more observers were deleted resulted in the conclusion that not inviting two observers should lead to ICC's of respectively .902 and .904. These improvements are not visible when rounded to the second decimal place. Therefore, we had no reason not to invite these observers to continue with this study.

Table 8.2 Coefficients for interrater reliability

	Video English lesson	Video Geography lesson
Intra-class correlation	.90	.95
Percentage agreement	75.14	82.22
Fleiss' κ	.27	.46

For the geography lesson, an ICC of .95 was found, again indicating that the observers had a high degree of consistency in their judgements. In comparison with the first lesson (the English lesson), this is not a major improvement. Looking at the intra-class correlation coefficient, the observers appeared to agree with each other very consistently.

Consistency in ratings is the tendency for one observer to increase, or decrease as another observer increases or decreases. The covariance between the observers plays a very important role in this statistic. This has the disadvantage that strict observers can have high correlations with more indulgent observers, while strict observers nevertheless give more insufficient scores than more lenient observers. That is why we also computed the percentage of agreement between the observers.

4.3.2 Agreement Percentage

A simple and popular method for calculating inter-assessor reliability consists in calculating the percentage agreement of the observers. This is done by adding up the number of items that received identical ratings by the observers and dividing that number by the total number of items rated by observers (Stemler, 2004). The consensus percentage among observers was 75.1% for the English lesson and 82.2% for the geography lesson. This means that the exact agreement on the question sufficient or insufficient was on average over 75% and 82%. This result indicates that the average agreement percentage of the observers is satisfactory.

The highest agreement percentages are found for both the most difficult and most easy items. The relatively low agreement percentages are found around the sum score of the scale. As we will see in paragraph 5.4, the items with the lowest percentages of consensus are exactly in the area of current development of the observed teacher. It is hardly surprising that the exact marking of the skill of the observed teacher causes relatively most consensus problems between the observers.

Several researchers are of the opinion that the percentage of agreement should be corrected for the chance of accidental agreement (Cohen, 1960; Kundel & Polansky, 2003; Landis & Koch, 1977). This is the subject of the following section.

4.3.3 Fleiss' κ

Fleiss' κ is a measure of the agreement between more than two observers, where agreement due to chance is factored out (Cohen, 1960; Fleiss & Cohen, 1973; Fleiss, 1981). Fleiss' κ varies from −1 (perfect disagreement), 0 (no different to change) to 1 (perfect agreement). According to Landis and Koch (1977) the inter-rater reliability is poor for values less than .00, slight for values between .0 and .20, fair for values between .21 and .40, moderate for values between .41 and .60, substantial for values between .61 and .80, and almost perfect for values between .81 and 1.0. These intervals for Fleiss' κ are cited as norms in many articles (e.g. Viera & Garret, 2015). Landis and Koch (1977), however, are much more modest in their

article. They are looking for a "consistent nomenclature". They call their intervals arbitrary. The intervals can be seen as "benchmarks" for the discussion about one of their tables in their article (Landis & Koch, 1977, 165). In their article, Landis and Koch do not provide any empirical arguments for their intervals and their indications of the strength of the agreement.

Falotico and Quatto (2015) found that Fleiss' κ statistic behaves inconsistently in cases of strong agreement between observers, since this statistic assumes lower values than it would have been expected. In the formula for Fleiss' κ all items are assessed equivalent. However, in a Rasch scale, the items are not equivalent. Some items are at the beginning of the dimension and are dominated by many teachers. The consensus between observers will be high in that part of the scale. The same applies to the items at the end of the dimension of a scale. Here too the consensus will be high, because many teachers do not meet these items. However, exactly at the point where the current skill of the observed teacher lies, the consensus will be relatively low. If it is important to control for chance, then there must also be a control for the skill level of an observed teacher, otherwise the Fleiss will underestimate.

It would be useful if an empirical study were to be conducted, in which the 'standards' of Landis and Koch would be validated. This is also done by Lipsey (1990) for the standards that Cohen (1967) proposed for effect size differences.

We started the observation training with video about an English lesson. On this video, we found a Fleiss' κ of .27, indicating a fair agreement (according to Landis and Koch) between the observers. For the geography lesson, we found a Fleiss' κ of .46, indicating a moderate agreement (according to Landis and Koch) between the observers. In view of the discussion above, we are inclined that the Fleiss' kappa's, we found make it clear in any case, that the agreement found between the observers is not based on chance only.

We found that after the training the observers grew in their mutual consistency and their degree of agreement. The extent to which their agreement could be explained by chance alone decreased after the training.

Furthermore, we found that observers were very consistent with each other in their judgments. The observers also had a high degree of agreement in their scores. Their amount of consensus was higher than can be achieved by guessing alone.

Each of the observers was invited to participate in this study. We may conclude that these results are sufficient to set up a study into the characteristics of the frequency distribution in the sample.

For a study in which we want to determine the area of immediate development of individual teachers, the ICC is sufficiently high, but it is also important that the percentage of agreement of the items in the middle of the Rasch scale is at least 70%.

4.4 The Fit of the Rasch Model

In a Guttman (1950) scale, items are arranged in such an order that an individual who responds correctly on a particular item also respond correctly on items of lower rank-order. With the perfect Guttman scale one is able to predict with the raw score

alone, which items were responded correctly or not. To measure a person's ability, Guttman scale is very helpful for finding a person's zone of proximal development. This "deterministic" Guttman model, however, works fine for constructs that are strictly hierarchical and highly structured. In most social science contexts however, data from respondents often do not closely match Guttman's deterministic model. That is why Guttman's deterministic model is brought within the probabilistic framework of the Rasch model. The Rasch model (Rasch, 1960, 1961) offers unique possibilities for arranging items and persons on a single dimension. Item difficulty parameters and abilities of persons can be estimated independently and find their location on the same dimension. The Rasch model requires the data of a scale to satisfy three assumptions:

- the scale should be unidimensional,
- the items of the scale should be local stochastic independent, and
- the item characteristic curves should be parallel.

We therefore checked whether the evaluations of the observers made with this instrument met these assumptions.

In most cases, a measurement scale is only used to determine the score of a person, because we are interested in the sample mean. In our case however, we are less interested in the average score of a sample. In our study, we are concerned with the scores of individual teachers in order to be able to coach them. This means that we have to set higher requirements in the quality of the individual items. That means also that we cannot work with global testing alone. We also need to map the quality of individual items. This requires tests that provide a detailed picture of the functioning of the individual items. Therefore, model-data fit analyses will be carried out using several different statistical programs.

Another reason for using different analysis techniques is that many analysis techniques do not really provide **the** proof, or **the** hard evidence for unidimensionality, local independence or parallelism of item characteristic curves.

4.4.1 Unidimensionality

The assumption of unidimensionality states that observations can be ascribed to a single latent construct, in our case: teaching skill observable in the classroom. The unidimensionality assumption of a (Rasch) scale is difficult to confirm or to disconfirm (DeMars, 2010). Nevertheless, we can use several procedures to test whether it is likely that a set of items form a unidimensional scale.

Confirmatory Factor Analysis

A possible procedure is using confirmatory factor analysis (CFA) with a one-factor model. For this analysis, we used the program Mplus 7.4 (Muthen & Muthen, 1998–2015). The usual χ^2-based test for model fit is substantially affected by

sample size (Marsh et al., 1988). Because we have a large sample of observations, we use the Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI). Both indices are less vulnerable to sample size. Furthermore, we consider the Root Mean Square Error of Approximation (RMSEA) to assess model fit. The norms for acceptable fit are CFI and TLI > .90 and RMSEA < .08 (Chen et al., 2008; Hu & Bentler, 1999; Marsh et al., 2004; Kline, 2005; Tucker & Lewis, 1973; Cheung & Rensvold, 2002).

Table 8.3 shows that both the Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI) for the dichotomised 32 items are above the norm of .90 and the Root Mean Square Error of Approximation (RMSEA) is below the norm of .08, which is an indication for unidimensionality.

In order to determine whether the one-factor model is an optimal model, we investigated whether a four-factor model that corresponds to the areas of proximal development found (cf. Table 8.9) might be a better alternative. This was not the case. Both the CFI and the TLI of this four-factor model were unacceptably low (respectively .728 and .708) and the RMSEA of this four-factor model was .132, which is unacceptably high (cf. Table 8.3).

A Scree Plot of Eigenvalues

Another way to check whether the 32 items of the teaching skill together form a unidimensional latent construct is using a "graphical test" by making a scree plot of the eigenvalues based on the correlation matrix of items. The eigenvalues of the factor analysis are plotted in Fig. 8.1.

The first eigenvalue (11.23) is considerably larger than the second (1.86) and third (1.49) eigenvalues. These results indicate that the scree plot clearly shows one dominant factor, which indicates that the assumption of unidimensionality seems to be reasonable.

Factor analysis is an analysis technique that stems from the classical test theory. Factor analysis is based on the factor loadings of the items. In the Rasch model, not so much the factor loadings as the item difficulties play a central role. That is why we need to extend the research into unidimensionality of the Rasch scale with a technique that has been specially developed for the Rasch model. We will use Andersen's (1973, 1977) log-likelihood ratio test. This analysis technique developed by Andersen also offers excellent possibilities to trace the items that cause disruptions of the unidimensionality.

Table 8.3 Confirmatory factor analyses

	CFI	TLI	RMSEA
Norms for model fit	>.90	>.90	<.08
Results of the intended one-factor model	.964	.961	.048
Results of an alternative four-factor model	.728	.708	.132

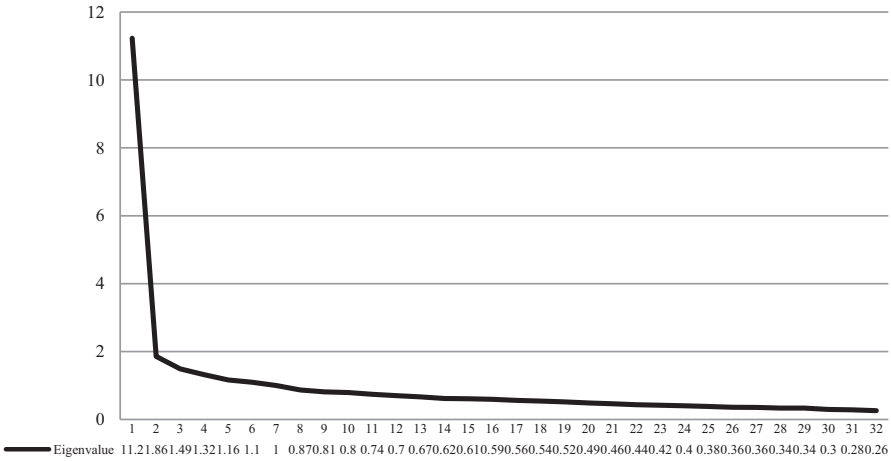

Fig. 8.1 Scree plot of eigenvalues

Table 8.4 Anderson's log likelihood ratio test for different teacher characteristics

	Anderson's χ^2	df	p-value
Gender	45.781	31	.042
Gender leaving out item 27	39.282	30	.120
Teaching experience (<5 years of experience and ≥5 years of experience)	35.812	31	.253
β-Subject matter (math, science, information science and so on) versus language, English and other subject matters	37.526	31	.195
Class size (<30 students and ≥30 students)	41.174	31	.105

Anderson's Log Likelihood Ratio Test

A third way to test the assumption of unidimensionality is to check whether variables other than the intended latent dimension, observable teaching skill, affect the item difficulty parameters. This is also important, because the observation instrument must be suitable for use with teachers who have different characteristics like gender and teaching experience, or work with different subject matters or different class sizes. We used Andersen's (1973, 1977) log-likelihood ratio test that is implemented in the eRm R-package (Mair & Hatzinger, 2007) to compare the difficulty parameters b for each item and to compute Anderson's log-likelihood ratio χ^2 test. Results are shown in Table 8.4.

Andersen's log-likelihood ratio test results showed that the difficulty parameters of

- male and female teachers,
- beginning and experienced teachers,
- teachers teaching beta-subject matter (math, science, information science and so on) on the one side and alfa and gamma subjects such as language, English and other subject matters on the other side,
- teachers working in small or large classes were invariant.

When we apply a general norm of .05 for the p-value of Andersen's log-likelihood ratio test, we found a small incident with item 27: "The teacher teaches students how to simplify complex problems". This item has a bit different item difficulty for female and male teachers.

4.4.2 Local Stochastic Independence

Local stochastic independence is one of the underlying assumptions of the Rasch model. The variable measured with a Rasch scale explains why the observed items are related to another. This assumption means that the observed items of a Rasch scale are conditionally independent of each other given the score on the latent variable that is measured by the Rasch scale. The assumption of local stochastic independence involves that the correlations between the items disappear when the effect of the intended latent variable (teaching skill) has been partialled out. We will use one overall procedure to test whether the 32 items meet this assumption and two item-specific procedures to detect the item pairs susceptible to local dependency.

Confirmatory Factor Analysis with all Residual Correlations Fixed at 0

Firstly we used confirmatory factor analysis (with the Mplus 7.4 program) to check the item correlations after the effect of the latent skill was partialled out. We formulated a one-factor model in which all residual correlations were set at zero.

Table 8.5 shows that both the Comparative Fit Index (CFI) and the Tucker-Lewis index (TLI) are above .90 and the Root Mean Square Error of Approximation (RMSEA) is below the norm of .08, which can be interpreted as an overall indication of local stochastic independence.

Computing Correlations Between the Residues of 32 Items

Using the Mplus 7.4 program, we computed (for the one-factor-model with free residual correlations) the residual correlations of the pairs of items after the effect of the intended latent variable (teaching skill) has been partialled out.

It turned out that 354 out of 496 residual correlations were below .10. A total of 141 residual correlations were between .10 and .30. Only one residual correlation was above .30. The residual correlation between item 22 (The teacher clearly

Table 8.5 Confirmatory factor analyses on 32 dichotomous items and 1 factor residual correlations set at 0

Model fit for residual correlations set at 0	CFI	TLI	RMSEA
Norm	>.90	>.90	<.08
Result	.945	.945	.057

specifies the lesson aims at the start of the lesson) and item 23 (The teacher evaluates whether the lesson aims have been reached) was .318. The residual correlation between item 22 and item 23 goes together with an R squared of .101.

Cohen (1988) evaluates an R below .10 as negligible and an R between .10 and .30 as a small effect. With the exception of the residual correlation between item 22 and item 23, these results might be interpreted as an indication of the local independence of the items.

Chen and Thissen's LDχ2 Index

Chen and Thissen (1997) proposed a standardized index, the LDχ^2 index, to establish whether there is a violation of the assumption of local stochastic independence for pairs of items. A value of <5 means that there is little likelihood of local dependence. Values between 5 and 10 form a "grey area". When the Chen-Thissen LD χ^2 has a value >10, it indicates possible local dependence. We computed Chen-Thissen's LDχ^2 with the program IRTPRO (Cai et al., 2005–2013). Results show that some pairs of items indicate possible local dependence (LDχ^2 > 10):

- LDχ^2:10.1: item 2 "maintains a relaxed atmosphere" with item15 "gives a clear explanation of how to use didactic aids and how to carry out assignments"
- LDχ^2:12.1: item 5 "ensures the lesson proceeds in an orderly manner" with item18 "stimulates learners to think about solutions"
- LDχ^2:10.4: item 9 "presents and explains the subject material in a clear manner" with item 24 "offers weaker learners extra study and instruction time"
- LDχ^2:10.6: item14 "teaches in a well-structured manner" with item 17 "stimulates the building of self-confidence in weaker learners"
- LDχ^2:11.5: item 22 "clearly specifies the lesson aims at the start of the lesson" with item 23 "evaluates whether the lesson aims have been reached".

According to this index, we have five pairs of items with possible local dependence. Only the relatively high LDχ^2:11.5 of the last pair of items (22/23) is in agreement with the actual correlation (.318) we have computed between the residuals of these items.

4.4.3 Parallelism of Item Characteristic Curves

Within the Rasch model, the probability of a positive score on an item should depend on the ability of a person, in our case the teacher. When the probability of a positive score on an item is plotted against the skill of teachers, the result would be a smooth S-shaped curve, called the item characteristic curve. The items in the scale should have a stable sequence for each ability group. This means that the item characteristic curves of the items should ideally be parallel. Examining whether certain items have too flat item or too steep characteristic curves, is important, because

these items function differently for people with different skills. We used various procedures to check whether this was the case for the 32 items in the scale.

Anderson's Log Likelihood Ratio Test for Teachers with Low and High Scores

Firstly, we used Andersen's (1973, 1977) log-likelihood ratio test to examine the equality of the item parameters of teachers with a high and low skill level. We used the eRm R-package (Mair & Hatzinger, 2007) to compare the difficulty parameters (b) for each item and to compute Anderson's log-likelihood ratio χ^2 test. Results are shown in Table 8.6.

Results show that with all 32 items Anderson's log likelihood ratio χ^2 test is 74.25 with 31 degrees of freedom and a p-value of .000, indicating a misfit. Leaving out item 17, 20, 31 show that the χ^2 is relatively small, given the number of degrees of freedom (28). The p-value is now .08, also indicating a reasonable fit. The misfitting items are: "item 17, stimulates the building of self-confidence in weaker students", "item 20, let students think aloud", and "item 31, encourages students to think critically". Following this test results, the other 28 items should have about the same difficulty parameters for teachers with a high and a low level of teaching skill. This is a first indication of parallelism of these 28 item characteristic curves.

The Slopes of the Item Characteristic Curves

Another way for testing parallelism is computing the actual slope of each item characteristic curve. We used the LTM R-package (Rizopoulos, 2006) for estimating the slope of the item characteristic curve of each item. The slopes and their standard errors are found in Table 8.7.

The average slope (also called as a parameter in the IRT terms) is 2.01. The rule of thumb for parallelism of item characteristics curves may be that a deviation of approximately two standard errors is too large. Slope parameters that are more than about two times their standard error (S.E.) higher than the average slope parameter are too steep. Slope parameters that are more than about two times their standard error (S.E.) smaller than the average slope parameter are too flat.

The slope of item 9 ("presents and explains the subject material in a clear manner") is rather steep (3.17). The slopes of item 20, 22, and 31 are rather flat. These

Table 8.6 Anderson's log likelihood ratio test for teachers with low and high scores

	Anderson ICC χ^2	df	p-value
32 items	74.246	31	.000
31 items, excluding item 20	60.864	30	.001
30 items, excluding item 20 and 31	43.913	29	.037
29 items, excluding item 20, 31 and 17	39.388	28	.075

Table 8.7 Slopes of the item characteristic curves

Item	Slope (a)	s.e.
1	1.6675	.2613
2	1.5649	.2549
3	1.7911	.2404
4	2.0916	.2506
5	1.6913	.2648
6	2.7299	.3595
7	1.8832	.2504
8	1.5001	.2323
9	3.1681	.4919
10	1.7026	.2284
11	2.7722	.3663
12	2.8507	.3694
13	1.5471	.2094
14	2.5233	.3236
15	1.9324	.2454
16	1.7006	.2201
17	1.5155	.1800
18	2.5843	.3004
19	2.2616	.2672
20	1.1702	.1740
21	1.7998	.2503
22	1.2960	.2066
23	1.8420	.2252
24	2.2747	.2686
25	2.5902	.3003
26	2.4196	.2770
27	1.7772	.2098
28	1.8963	.2351
29	2.3890	.2774
30	1.7950	.2296
31	1.2427	.1637
32	2.2183	.2510

items are respectively “let students think aloud”, “clearly specifies the lesson aims at the start of the lesson”, “encourages students to think critically”.

4.4.4 Conclusions About the Fit of the Rasch Model

At the moment there is no simple approach to test whether a dataset satisfies the assumptions of the Rasch model. Therefore, we have used several different procedures, implemented in several different statistical packages. The use of many

procedures brings along that always one or more items give significant misfit. Some items however, produced several times a misfit:

- Item 9 "presents and explains the subject material in a clear manner" had a too high LDχ^2 (10.4) with item 24 and had a slope of the item characteristic curve that was too steep (3.17).
- Item 20 "lets learners think aloud" disturbed the parallelism of the item characteristic curves with both a significant result on the Andersen's log likelihood ratio test for high and low scorers, and a too flat slope parameter (1.70).
- Item 22 "clearly specifies the lesson aims at the start of the lesson" showed a too high residual correlation (.318) with item 23, a too high LDχ^2 (11.5) with item 23, a too flat slope parameter (1.30), and a significant result on the Andersen's log likelihood ratio test for high and low scorers.
- Item 31 "encourage learners to think critically" had significant result on the Andersen's log likelihood ratio test for high and low scorers and a too flat slope parameter (1.24).

These four items will bring along some problems in determining the zone of proximal development of individual teachers. Therefore, we will remove item 9, 20, 22 and 31 from the scale.

4.5 The Person Fit

Thus far, attention was given to items that disturb the fit of the Rasch model. Now the person fit is considered. There are persons having unexpected item score patterns, that should not be expected when the data fit the Rasch model. In the deterministic Guttman model, persons should not respond correctly to difficult items when they respond wrongly to easier items. In the Rasch model, this requirement is somewhat more relaxed, but the number of Guttman errors should remain within certain limits. This is especially true when we want to use a person's score to detect a person's zone of proximal development. Several statistics are used to test a person's fit (Mousavi et al., 2016). In this study, we will use the G-normed-statistic (Meijer, 1994).

4.5.1 Meijer's G-Normed-Index

The simple G-statistic counts the number of (0, 1) pairs given that the items are ordered in decreasing proportion-correct scores order. The size of the G-statistic depends on the amount of (pairs of) items. The G-normed-statistic was created to bind the G-statistic between zero and one by dividing it by its maximum (Van der Flier, 1982; Meijer, 1994; Tendeiro, 2014). We used the Per Fit R-package (Mousavi et al., 2016) to compute the G-normed-statistic for each observed teacher. Table 8.8

Table 8.8 Meijer's G normed index (average: .21; standard deviation: .18)

G normed index	<.30	.30–.50	>.50
% of observed teachers	72.7	21.4	5.9

presents the results. In an empirical study of Van der Lans et al. (2016) the norm of .30 is proposed for this person fit index.

In 5.9% of the cases the G-normed-index is above 50%, 21.4% of the observed teachers have a G- normed-index between .30 and .50, and 72.7% of the teachers have a G-normed index of <.30.

In the existing statistical literature, we did not find a norm for the G-normed-statistic yet. If we accept the proposal of Van der Lans et al. (2016), a GFI of .30 and more seems too high to be used as a cut-off. This means that we should be careful to use the results for finding a person's zone of proximal development in about 27% of the cases.

Most of these teachers with a high (>.30) G-normed-index are found by four observers who observed each around 20 teachers and by three other observers who observed just one or two teachers. These seven observers have on average five years less experience as a teacher than the other observers do. This difference is significant ($p = .000$). To avoid that this difference affects the result significantly, it is important that these teachers were observed (several) more times, before we could estimate their zone of proximal development more precisely. Another, perhaps simpler approach could be to develop a variant of the G-normed index that can be used in the training of observers. It is also important that observers themselves have sufficient experience in teaching. In the future it might be important to exclude novice teachers from acting as observers in research.

5 Results

Based on results above, we found that the ICALT observation scale with 28 items fulfil the criteria of the Rasch model. In the next part of this chapter, we will present the items, their difficulty parameters and the person parameters of each observed teacher.

5.1 Item Difficulties and Person Parameters

We used the eRm R-package (Mair & Hatzinger, 2007) to compute the difficulty parameter b for each of the dichotomized 28 selected items. Table 8.9 shows our version of a slightly changed Wright map. In column, two and three the items are presented in the order of their difficulty parameter (b) with their standard errors (S.E.).

Table 8.9 Wright map for the ICALT28-scale (N = 375 Korean secondary school teachers)

Item	b	se	Warm's θ	se	Frequency	Cumulative frequency
			−4.477	1.480	1.1	1.1
			−3.313	.878	.3	1.3
			−2.736	.700	.5	1.9
			−2.331	.608	1.1	2.9
			−2.011	.551	1.6	4.5
			−1.740	.512	2.9	7.5
Maintains a relaxed atmosphere	**−1.656**	.163				
Ensures the lesson proceeds in an orderly manner	**−1.549**	.159				
			−1.501	.484	2.7	10.1
Shows respect for students in his/her behaviour and language	**−1.447**	.156				
Uses the time for learning efficiently	**−1.348**	.153				
			−1.284	.463	3.2	13.3
			−1.083	.448	3.5	16.8
			−.894	.436	1.9	18.7
Gives interactive instructions	**−.837**	.141				
			−.713	.428	2.4	21.1
Promotes students' self-confidence	**−.656**	.139				
Provides effective classroom management	**−.636**	.138				
Presents and explains the subject material in a clear manner	**−.557**	.137				
			−.538	.421	2.4	23.5
Encourages students to do their best	**−.519**	.137				
Monitors to ensure students carry out activities in the appropriate manner	**−.442**	.136				
Teaches in a well-structured manner	**−.423**	.136				
			−.367	.417	4.0	27.5
Engages all students in the lesson	**−.348**	.135				
Stimulates the application of what has been learned	**−.310**	.135				
Offers activities and work forms that stimulate students to take an active approach	**−.291**	.134				
			−.198	.415	4.3	31.7
Gives a clear explanation of how to use didactic aids and how to carry out assignments	**−.089**	.133				
			−.030	.415	4.3	36.0

(continued)

Table 8.9 (continued)

Item	b	se	Warm's θ	se	Frequency	Cumulative frequency
Stimulates the use of control activities	**.002**	.133				
During the presentation stage, checks whether students have understood the subject material	**.056**	.132				
			.138	.417	4.8	40.8
Evaluates whether the lesson aims have been reached	**.182**	.132				
Fosters mutual respect	**.218**	.132				
			.309	.420	3.5	44.3
Asks questions which stimulate students to reflect	**.470**	.132				
			.483	.426	4.0	48.3
Teaches students to check solutions	**.632**	.133				
			.663	.433	3.5	51.7
Stimulates students to think about solutions	**.723**	.133				
			.850	.444	4.3	56.0
Teaches students how to simplify complex problems	**.852**	.134				
			1.048	.457	3.2	59.2
			1.259	.474	2.9	62.1
Adjusts the processing of subject matter to relevant inter-student differences	**1.211**	.138				
Adjusts instruction to relevant inter-student differences	**1.309**	.139				
Asks students to reflect on practical strategies	**1.369**	.140				
Stimulates the building of self-confidence in weaker students	**1.369**	.140				
			1.488	.497	2.9	65.1
			1.742	.527	4.0	69.1
			2.032	.568	3.2	72.3
			2.376	.629	3.2	75.5
Offers weaker students extra study and instruction time	**2.716**	.170				
			2.813	.725	5.6	81.1
			3.434	.909	4.8	85.9
			4.659	1.525	14.1	100.0

The item sequence is more or less similar to the item sequence found in previous studies with Dutch teachers in secondary education (Van de Grift et al., 2014; Van der Lans et al., 2016, 2017). The easiest items are the items about a safe learning climate and efficient classroom management. These items are followed in difficulty with items about the quality of basic instruction. Next items on the dimension are about activating students, teaching learning strategies, and the dimension end with differentiation of teaching, which are the most difficult ones. We will use this ordering in categories of items as indications of the zones of proximal development.

There is one important exception in this ordering. In the previous Dutch study, the item 'fosters mutual respect' has a difficulty parameter that is much lower than in the current Korean study (cf. Van de Grift et al., 2014; Van der Lans et al., 2016, 2017).

The person parameters were estimated using Warm's weighted likelihood estimates (Warm, 1989). This procedure is less biased in comparison with the traditional maximum likelihood estimates method (Hoijtink & Boomsma, 1995) and has the advantage that it also can be used to estimate the skills of people with a zero and a maximum score. We used the program WINMIRA (Von Davier, 1994) to compute the person parameters Warm's weighted likelihood estimates. Table 8.9 shows in column four and five the Warm's θ and the standard error and some information on the frequency distribution is found in column six and seven.

5.2 Warm's θ and some Teacher, Class and School Characteristics

Table 8.10 presents some descriptive information about the characteristics of the frequency distribution of Warm's θ.

The average score is 1.03 with a standard deviation of 2.09. Both skewness and kurtosis are <1.0, which is in indication for an approximately normal distribution. Nevertheless we can observe in Table 8.11 that the amount of teachers with a perfect score (θ = 4.66) is rather high (14%).

Table 8.11 presents some details about relationships of teachers, classrooms and schools and the skill of teachers. We found no significant differences between male and female teachers, teachers teaching α-γ- and β-subject matters or teachers working in general and vocational schools, or working in public or private schools. There was no significant relationship between the years of experience of a teacher and teaching skill. We found a significant, but small, negative correlation of −.25 between class size and the skill shown by teachers: Teachers show lower skill in large classrooms. Furthermore, we found a significant difference between the skill of teachers in lower and upper secondary education. The difference is 55% of a standard deviation in the advantage of the teacher in lower secondary education.

Table 8.10 Relations between teacher and school characteristics and Warm's θ

	n	average	standard deviation	effect size	significant	R with θ	significant
Theta-score of all 28 ICALT-items	375	1.03	2.09				
Male	183	.95	2.27	.077	.470		
Female	192	1.11	1.90				
Years of experience	369					.095	.069
Subject α-γ	245	1.06	2.08	.033	.700		
Subject β	130	.98	2.11				
Class size	351					−.246	.000
Lower secondary	154	1.69	2.19	.551	.000		
Upper secondary	221	.58	1.88				
General	361	1.04	2.11	.053	.852		
Vocational	14	.93	1.15				
Public	223	.91	1.76	.135	.213		
Private	151	1.19	2.48				
Student's academic engagement	375	3.10	.69			.68	.000

Table 8.11 Areas of proximal development

Zone	Warm's θ			Description	% lessons
1			<−1.0	Safe climate and efficient classroom management	16.8
2	−1.0	–	0.0	Basic tasks of teaching and activating students	19.2
3	0.0	–	1.0	Teaching how to learn	20.0
4	1.0	–	3.0	Differentiating teaching	25.1
5	3.0		4.00	Satisfies the basic and (almost all) advanced teaching skills	4.8
6			>4.0	Satisfies all teaching skills	14.1
					100.0

5.3 Predictive Value of the Scale

In order to study the predictive validity of the Rasch scale we developed a simple scale for measuring the students' academic engagement.

The scale consists of three items that reflect increasing student involvement: 'the learners are fully engaged in the lesson', 'the learners show that they are interested' and 'the learners take an active approach to learning'. The students' academic engagement scale has a range of 1–4. We found an average score of 3.10 with a standard deviation of .69 (cf. Table 8.10). The theta-score of the 28-ICALT-scale had a correlation of .68 with the students' academic engagement scale. So the better the teaching skill, the better the students were involved in the lesson. This is an indication of the predictive validity of the ICALT28-scale.

5.4 A Proposal for Detecting a person's Zone of Proximal Development

The raw score of a perfect Guttman scale predicts which items are responded correctly or not. This is very helpful and very precise for finding a person's zone of proximal development. The stochastic character of a Rasch scale, however, brings along several uncertainties in finding a person's zone of proximal development. We have already seen in Table 8.8 that 27% of the observed teachers have severe deviations from the perfect Guttman model. But even when the items have Q-indices (Rost & Von Davier, 1994) nicely near zero and when we wait for more observations for persons with high G-normed-indices (Meijer, 1994), we still have concerns with finding the exact zone proximal development of the observed teachers. The reasons for these concerns are found in the stochastic character of a Rasch scale. Therefore, we will propose an overall procedure with areas of proximal development, based on the meaning of the items. In order to reduce uncertainties in finding a person's zone of proximal development we will use 'areas of proximal development', instead of separate items.

The easiest items are the items about safe learning climate and efficient classroom management. These sets of items are followed in difficulty with a group of items about the quality of basic instruction. Items that are more difficult are about activating students, teaching learning strategies, and the group of items about differentiation of teaching, are the most difficult ones. Inspecting Table 8.9 makes clear that more or less the same ordering is found in the Rasch scale. We will use this ordering in domains of items as indications of the zones of proximal development. Our proposal is laid down in Table 8.11.

Next sections give some descriptions of these areas of proximal development. The scores are clustered in six categories. We used the Warm's θ scores: below −1; −1–0; 0–1; 1–3; 3–4; and above 4. These are all intervals of just one interval point on the Warm's θ scale. Only one interval is larger (1–3) larger. This had to do with the most difficult item. This is of course an arbitrary format, but it guarantees a simple application. The meaning of the categories is just the concept that fits with the meaning of the items within each category. The meaning of the categories corresponds with the complexity level of the teaching skill ranging from low complexity to high complexity. We will present the percentage of lessons we found for each domain.

5.4.1 Safe Climate and Efficient Classroom Management

In 16.8% of the observed lessons, the θ-score is below −1.0. In these lessons, creating a safe learning climate and in maintaining an orderly classroom management was not sufficient. E.g., the atmosphere in the classroom is not relaxed, the lesson does not proceed in an orderly manner and the time for learning is not used efficiently. When there were no special events during the lesson or special other reason

for this low score, than it is clear that the zone of proximal development of teachers within this group is working on a safe climate and an orderly classroom management.

5.4.2 Basic Tasks of Teaching and Activating Students

In 19.2% of the lessons, the θ-score lies between −1.0 and 0.0. These lessons could be improved by e.g. giving more structured and more interactive instructions.

5.4.3 Teaching Students How to Learn

In 20.0% of the lessons, the θ-score is between 0.0 and 1.0. In these lessons, the basic skills of teaching (creating a safe and stimulating educational climate, an orderly classroom management, and clear and activating instruction) are sufficient.

These lessons could be improved by teaching students how they can learn things: The teacher can improve the lesson by e.g. asking questions that stimulate students to reflect and to check solutions.

5.4.4 Differentiating Teaching

In 25.1% of the lessons, the basic tasks of teaching, activating students, and teaching students how to learn things are observed to be sufficient. These lessons have θ-scores between 1.0 and 3.0. These lessons can be improved by adjusting instruction and the processing of subject matter to relevant inter-student differences. One of the most difficult tasks for the teachers in this zone of proximal development is offering weaker students extra study and instruction time.

5.4.5 Lessons Satisfying All Basic and Almost All Advanced Teaching Skills

In 4.8% of the lessons, a θ-score between 3.0 and 4.0 is found. Teachers reveal in these lessons all basic skills and most advanced teaching skills.

5.4.6 Lessons Satisfying all Teaching Skills

In 14.1% of the lessons, all 28 teaching skills were exhibited. This is a rather high percentage. The percentage of 14% perfect scores could be a reason to add some more important items with higher difficulty to this scale. We know that the current

version of the ICALT observation instrument can be supplemented with additional items about differentiation.

These somewhat arbitrary areas are mostly important for giving a θ-score a meaning in terms of the skills of teachers. The θ-score is the actual level of development, and the domain (cf. Table 8.9) specifies the zone of proximal development. The limits used for these domains are of course somewhat arbitrary. When a lesson gets a score that is just below the upper limit of one of the different domains, it is probably wise to shift the zone of proximal development to the next area. To give an example: A teacher with a score of Warm's θ = .85 (cf. Table 8.9) does not really have to wait until he masters the last item of teaching how to learn, before he can start differentiation of his instruction.

6 Conclusions

In this study, we reported the development of a 28-item-scale for observing teaching skills that fulfils the assumptions of the dichotomous Rasch model.

We discovered that the order of item difficulty found among Dutch secondary school teachers is in general maintained among secondary school teachers from a totally different culture, the South Korean culture. There is one important exception in this ordering. In the previous Dutch study, the item 'fosters mutual respect' has a difficulty parameter that is much lower than in the current Korean study. This is probably due to the fact that the word 'respect' in Asian cultures has a more stringent meaning than in many Western European cultures. This makes it necessary to conduct further and more detailed research into cultural differences in the quality of teaching skill.

The scores on the scale had predictive value for the engagement of students. In subsequent studies it should be determined whether the scale also has a predictive value for the performance of the students.

With this study, we have developed an observation tool with which we can not only determine the current level of development of a teacher, but we also can give an indication of the zone of proximal development of the observed teacher. The latter in particular is very important. It simply does not help enough if we tell a teacher what his or her score is and what s/he does not do well. The 'trick' is to help a teacher by pointing out activities that s/he does not do, but that are within her or his reach. This ICALT observation instrument offers the possibility to coach teachers and guide them in matters that they are not yet doing.

Appendix

ICALT Lesson Observation Form (international comparison of learning and teaching)			
Country	…	Date observation (dd-mm-yyyy)	…
School name	…	Class …	
Location	…	No of learners present …	
Level of education	0=primary education 1=secondary education	Name observer …	
School denomination	0=public 1=private	Gender observer M / F	
Subject matter	…	Years of teaching experience observer	…
Name teacher	…	Occupation observer	0=school teacher 1=university teacher 2=other Please specify…
Gender teacher M / F Years of teaching experience teacher	…		
Observe the following behaviours and events:			
Rate[1] Please circle the appropriate answer: 1= mostly weak; 2=more often weak than strong; 3= more often strong than weak; 4= mostly strong **Observed**[2] Please circle the appropriate answer: 0= no, I have not observed this; 1= yes, I have observed this			

Domain	**Indicator: The teacher…**		**Rate**[1]	Examples of good practice: *The teacher…*	Observed[2]
Safe and stimulating learning climate	**1**	**…shows respect for learners in his/her behaviour and language**	**1 2 3 4**	*…lets learners finish their sentences*	*0 1*
				…listens to what learners have to say	*0 1*
				…does not make role stereotyping remarks	*0 1*
	2	**…maintains a relaxed atmosphere**	**1 2 3 4**	*…addresses learners in a positive manner*	*0 1*
				…uses and stimulates humour	*0 1*
				…accepts the fact that learners make mistakes	*0 1*
				…shows compassion and empathy for all learners present	*0 1*
	3	**…promotes learners' self-confidence**	**1 2 3 4**	*…gives positive feedback on questions and remarks from leaners*	*0 1*
				…compliments learners on their work	*0 1*
				…acknowledges the contributions that learners make	*0 1*
	4	**…fosters mutual respect**	**1 2 3 4**	*…stimulates learners to listen to each other*	*0 1*
				…intervenes when learners make fun of someone	*0 1*
				…keeps (cultural) differences and idiosyncrasies in mind	*0 1*
				…stimulates solidarity between learners	*0 1*
				…encourages learners to experience activities as group events	*0 1*

(continued)

ICALT Lesson Observation Form (international comparison of learning and teaching)					
Efficient organi-sation	**5**	**…ensures the lesson proceeds in an orderly manner**	**1 2 3 4**	*Learners enter and settle in an orderly manner*	*0 1*
				…intervenes timely and appropriately in case of disorder	*0 1*
				…safeguards the agreed rules and codes of conduct	*0 1*
				…keeps all leaners involved in activities until the end of the lesson	*0 1*
				…makes sure that learners know what to do if they need help with their work and explains clearly when they can ask for help	*0 1*
				…makes sure learners know what to do when they have finished their work	*0 1*
	6	**…monitors to ensure learners carry out activities in the appropriate manner**	**1 2 3 4**	*…checks whether learners have understood what they have to do*	*0 1*
				…provides feedback on learners' social functioning whilst carrying out a task	*0 1*
	7	**…provides effective classroom management**	**1 2 3 4**	*…explains clearly which materials can be used*	*0 1*
				The materials for the lesson are ready for use	*0 1*
				Materials are geared at the right level and developmental stage of the learners	*0 1*
	8	**…uses the time for learning efficiently**	**1 2 3 4**	*… starts the lesson on time*	*0 1*
				… does not waste time at the beginning, during, or at the end of the lesson	*0 1*
				…prevents any unnecessary breaks from occurring	*0 1*
				…does not keep learners waiting	*0 1*

Clear and structured instructions	**9**	**…presents and explains the subject material in a clear manner**	**1 2 3 4**	*…activates prior knowledge of learners*	*0 1*
				…gives staged instructions	*0 1*
				…poses questions which learners can understand	*0 1*
				…summarises the subject material from time to time	*0 1*
	10	**…gives feedback to learners**	**1 2 3 4**	*…makes clear **whether** an answer is right or wrong*	*0 1*
				*…makes clear **why** an answer is right or wrong*	*0 1*
				…gives feedback on the way in which learners have arrived at their answer	*0 1*
	11	**…engages all learners in the lesson**	**1 2 3 4**	*…creates learners assignments which stimulate active participation*	*0 1*
				…asks questions which stimulate learners to reflect	*0 1*
				…makes sure that learners listen and/or continue working	*0 1*
				…allows for 'thinking time' after asking a question	*0 1*
				…also invites learners to participate who do not volunteer to do so	*0 1*
	12	**…during the presentation stage, checks whether learners have understood the subject material**	**1 2 3 4**	*…ask questions which stimulate learners to reflect*	*0 1*
				…checks regularly whether learners understand what the lesson is about	*0 1*
	13	**…encourages learners to do their best**	**1 2 3 4**	*…praises learners who do their best*	*0 1*
				…makes clear that all learners should do their best	*0 1*
				…expresses positive expectations about what learners are going to achieve	*0 1*

(continued)

ICALT Lesson Observation Form (international comparison of learning and teaching)					
	14	**…teaches in a well-structured manner**	**1 2 3 4**	*The lesson is built up in terms of clear stages and transitions between stages*	*0 1*
				The lesson builds up logically, going from the simple to the complex	*0 1*
				Activities and assignments are connected to the materials presented during the presentation stage	*0 1*
				The lesson offers a good variety of presentation, instruction, controlled practice, free practice, and so forth.	*0 1*
	15	**…gives a clear explanation of how to use didactic aids and how to carry out assignments**	**1 2 3 4**	*…makes sure that all learners know what to do*	*0 1*
				…explains how lesson aims and assignments relate to each other	*0 1*
				…explains clearly which materials and sources can be used	*0 1*
Intensive and activating teaching	**16**	**…offers activities and work forms that stimulate learners to take an active approach**	**1 2 3 4**	*…uses diverse forms of conversation and discussion*	*0 1*
				…offers controlled (pre-)practice	*0 1*
				…lets learners work in Group	*0 1*
				…uses Information and Communication Technology (ICT, e.g., digiboard, beamer)	*0 1*
				…employs a variety of instruction strategis	*0 1*
				…varies assignments	*0 1*
				…varies lesson material	*0 1*
				…uses materials and examples from daily Life	*0 1*
				…asks a range of questions	*0 1*
	17	**…stimulates the building of self-confidence in weaker leaners**	**1 2 3 4**	*…gives positive feedback on questions from weaker learners*	*0 1*
				…displays positive expectations about what weaker learners have to achieve	*0 1*
				…compliments weaker learners on their works	*0 1*
				…acknowledges the contributions made by weaker learners	*0 1*

	18	**…stimulates learners to think about solutions**	**1 2 3 4**	*…shows learners the path they can take towards a Solutions*	*0 1*
				…teaches strategies for problem-solving and referencing	*0 1*
				…teaches learners how to consult sources and reference works	*0 1*
				…offers learners checklists for problem-solving	*0 1*
	19	**…asks questions which stimulate learners to reflect**	**1 2 3 4**	*…waits long enough to give all learners the chance to answer a question*	*0 1*
				…encourages learners to ask each other questions and explain things to each other	*0 1*
				…asks learners to explain the different steps of their strategy	*0 1*
				…checks regularly whether instructions have been understood	*0 1*
				…asks questions which stimulate reflection and learner feedback	*0 1*
				…checks regularly whether learners understand what the lesson is about	*0 1*
	20	**…lets learners think aloud**	**1 2 3 4**	*…provides the opportunity for learners to think aloud about solutions*	*0 1*
				…asks learners to verbalise solutions	*0 1*
	21	**…gives interactive instructions**	**1 2 3 4**	*…promotes the interaction between learners*	*0 1*
				…promotes the interaction between teacher and learners	*0 1*
	22	**…clearly specifies the lesson aims at the start of the lesson**	**1 2 3 4**	*…informs learners at the start of the lesson about the lesson aim*	*0 1*
				…clarifies the aims of assignments and their learning purpose	*0 1*

(continued)

ICALT Lesson Observation Form (international comparison of learning and teaching)					
Adjusting instructions and learner processing to inter-learner differences	**23**	**…evaluates whether the lesson aims have been reached**	**1 2 3 4**	*…evaluates whether the lesson aims have been reached*	*0 1*
				…evaluates learners' performance	*0 1*
	24	**…offers weaker learners extra study and instruction time**	**1 2 3 4**	*…gives weaker learners extra study time*	*0 1*
				…gives weaker learners extra instruction Time	*0 1*
				…gives weaker learners extra exercises/practices	*0 1*
				…gives weaker learners 'pre- or post-instruction'	*0 1*
	25	**…adjusts instructions to relevant inter-learner differences**	**1 2 3 4**	*…puts learners who need little instructions (already) to works*	*0 1*
				…gives additional instructions to small groups or individual learners	*0 1*
				…does not simply focus on the average learner	*0 1*
	26	**…adjusts the processing of subject matter to relevant inter-learner differences**	**1 2 3 4**	*…distinguishes between learners in terms of the length and size of assignments*	*0 1*
				…allows for flexibility in the time learners get to complete assignments	*0 1*
				…lets some learners use additional aids and means	*0 1*
Teaching learning strategies	**27**	**…teaches learners how to simplify complex problems**	**1 2 3 4**	*…teaches learners how to simplify complex problem*	*0 1*
				…teaches learners how to break down complex problems into simpler Jones	*0 1*
				…teaches learners to order complex problem	*0 1*
	28	**…stimulates the use of control activities**	**1 2 3 4**	*…pays attention to prediction strategies for reading*	*0 1*
				…lets learners relate solutions to the context of a problem	*0 1*
				…stimulates the application of alternative strategis	*0 1*
	29	**…teaches learners to check solutions**	**1 2 3 4**	*…teaches learners how to estimate outcomes*	*0 1*
				…teaches learners how to predict outcomes	*0 1*
				…teaches learners how to relate outcomes to the practical context	*0 1*

	30	**…stimulates the application of what has been learned**	**1 2 3 4**	*… stimulates the conscious application of what has been learned in other (different) learning contexts*	*0 1*
				…explains to learners how solutions can be applied in different situations	*0 1*
				…relates problems to previously solved problem	*0 1*
	31	**…encourages learners to think critically**	**1 2 3 4**	*…asks learners to provide explanations for occurrences*	*0 1*
				…asks learners for their opinion	*0 1*
				…asks learners to reflect on solutions or answers given	*0 1*
				…asks learners to provide examples of their own	*0 1*
	32	**…asks learners to reflect on practical strategies**	**1 2 3 4**	*…asks learners to explain the different steps of the strategy applied*	*0 1*
				…gives an explicit explanation of possible (problem-solving) strategies	*0 1*
				…asks learners to expand on the pros and cons of different strategies	*0 1*
	Indicator: The learners…		**Rate[1]**	Examples of good practice: *Learners…*	Observed[2]
Learner enga-gement	**33**	**…are fully engaged in the lesson**	**1 2 3 4**	*…pay attention during instructions are given*	*0 1*
				…participate actively in conversations and discussions	*0 1*
				…ask questions	*0 1*
	34	**…show that they are interested**	**1 2 3 4**	*…listen actively when instructions are being given*	*0 1*
				…show their interest by asking follow-up questions	*0 1*
	35	**…take an active approach to learning**	**1 2 3 4**	*…ask follow-up questions*	*0 1*
				…show that they take responsibility for their own learning process	*0 1*
				…work independently	*0 1*
				…take the initiative themselves	*0 1*
				…use their time efficiently	*0 1*

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*, 123–140.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42*, 69–81.

Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2018). Qualities of classroom observation systems. *School Effectiveness and School Improvement, 30*, 3–29. https://doi.org/10.1080/09243453.2018.1539014

Cai, L., Thissen, D., & Du Toit, S. (2005–2013). *IRTPRO 2.1*. Scientific Software International.

Chen, W. -H., & Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Journal of Educational and Behavioral Statistics 22*(3), 265–289.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research, 36*(4), 462–494.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modelling, 9*(2), 233–255.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290.

Cohen, J. A. (1960). Coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Cohen, J. (1967). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

De Jager, T., Coetzee, M.J., Maulana, R., Helms-Lorenz, M., & Van de Grift, W. (2017). Profile of South African secondary-school teachers' teaching quality: Evaluation of teaching practices using an observation instrument. *Educational Studies, 43*(4), 410–429.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Dobbelaer, M. J. (2019). *The quality and qualities of classroom observation systems*. Ipskamp printing.

Falotico, R., & Quatto, P. (2015). Fleiss' kappa statistic without paradoxes. *Quality & Quantity, 49*, 463–470.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). John Wiley.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*, 613–619.

Guttman, L. A. (1950). The basis for scalogram analysis. In S. A. Stouffer, Guttman, L. A., & E. A. Schuman Measurement and prediction. Volume 4 of studies in social psychology in world war II. : Princeton University Press

Hallgren, K. V. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23–34.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review: Papers & Proceedings, 100*(2), 267–271.

Hattie, J. A. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hattie, J. A. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge.

Hoijtink, H., & Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 53–68). Springer.

Houtveen, A. A. M., & Van de Grift, W. J. C. M. (2007a). Effects of metacognitive strategy instruction and instruction time on reading comprehension. *School Effectiveness and School Improvement, 18*(2), 173–190.

Houtveen, A. A. M., & Van de Grift, W. J. C. M. (2007b). Reading instruction for struggling learners. *Journal of Education for Students Placed at Risk, 12*(4), 405–424.

Houtveen, A. A. M., Van de Grift, W. J. C. M., & Creemers, B. P. M. (2004). Effective school improvement in mathematics. *School Effectiveness and School Improvement, 15*(3–4), 337–376.

Houtveen, A. A. M., Van de Grift, W. J. C. M., & Brokamp, S. K. (2014). Fluent reading in special elementary education. *School Effectiveness and School Improvement, 25*(4), 555–569.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (NBER Working Paper No. 14607).

Kirby, A., Gebski, V., & Keech, A. C. (2002). Determining the sample size in a clinical trial. *The Medical Journal of Australia, 177*(5), 256–257.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). Guilford.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research, 88*(4), 547–588.

Kundel, H. L., & Polansky, M. (2003). Measurement of observer agreement. *Radiology, 228*, 303–308.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Sage.

Lompscher, J., & Rückriem, G. (2002). Editorial. In L. S. Vygotskij (Ed.), *Denken und Sprechen*. Beltz Verlag.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modelling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*, 1–20.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness of fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*, 391–410.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers of overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modelling, 11*, 320–341.

Marzano, R. J. (2003). *What works in schools: Translating research in action*. ASCD.

Maulana, R., Helms-Lorenz, M., & Van de Grift, W. (2015). Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement, 26*, 169–194.

Maulana, R., Helms-Lorenz, M., & Van de Grift, W. (2016). The role of autonomous motivation for academic engagement of Indonesian secondary school students: A multilevel modelling approach. In R. B. King & A. B. I. Bernardo (Eds.), *The psychology of Asian learners. A festschrift in honor of David Watkins* (pp. 237–252). Springer.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18*, 311–314.

Mousavi, A., Tendeiro, J. N., & Younesi, J. (2016). Person fit assessment using the PerFit package in R. *The Quantitative Methods for Psychology, 12*(3), 232–242.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Muthén and Muthén.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Denmarks Paedagogiske Institut.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, IV* (pp. 321–334). University of California Press.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools and academic achievement. *Econometrica, 73*(2), 417–458.

Rizopoulos, D. (2006). ltm: An R-package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1–25.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4), 1–11.

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2018). On the usefulness of interrater reliability coefficients. In M. Wiberg, S. A. Culpepper, R. Janssen, J. Gonzáles, & D. Molenaar (Eds.), *Quantitative psychology. The 82nd annual meeting of the psychometric society, Zurich, Switzerland* (pp. 67–76). Springer.

Tendeiro, J. N. (2014). Package 'PerFit' (published online). In R. Cran (Ed.), *The comprehensive R network*. Retrieved from: http://cran.r-project.org/web/packages/PerFit/PerFit.pdf

Tucker, L., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1–10.
Van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology, 13*(3), 267–298.
Van de Grift, W. (2007). Quality of teaching in four European countries: A review of the literature and an application of an assessment instrument. *Educational Research, 49*(2), 127–152.
Van de Grift, W. J. C. M. (2014). Measuring teaching quality in several European countries. *School Effectiveness and School Improvement, 25*(3), 295–311.
Van de Grift, W. (2021). Het coachen van leraren (1/2) [the coaching of teachers]. *Basisschoolmanagement, 2*, 24–29.
Van de Grift, W. J. C. M., & Lam, J. F. (1998). Het didactisch handelen in het basisonderwijs [Teaching in primary education]. *Tijdschrift voor Onderwijsresearch, 23*(3), 224–241.
Van de Grift, W., Helms-Lorenz, M., & Maulana, R. (2014). Teaching skills of student teachers: Calibration of an evaluation instrument and its value in predicting student academic engagement. *Studies in Educational Evaluation, 43*, 150–159.
Van de Grift, W. J. C. M., Chun, S., Maulana, R., Lee, O., & Helms-Lorenz, M. (2017). Measuring teaching quality and student engagement in South Korea and The Netherlands. *School Effectiveness and School Improvement, 28*(3), 337–349.
Van de Grift, W. J. C. M., Houtveen, A. A. M., Van den Hurk, H. T. G., & Terpstra, O. (2019). Measuring teaching skills in elementary education using the Rasch model. *School Effectiveness and School Improvement, 30*, 455–486. https://doi.org/10.1080/09243453.2019.1577743
Van den Hurk, H. T. G., Houtveen, A. A. M., & Van de Grift, W. J. C. M. (2016). Fostering effective teaching behaviour through the use of data-feedback. *Teaching and Teacher Education, 60*, 444–451.
Van der Lans, R. M., Van de Grift, W. J. C. M., Van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation, 50*, 88–95.
Van der Lans, R. M., Van de Grift, W. J. C. M., & Van Veen, K. (2017). Individual differences in teacher development: An exploration of the applicability of a stage model to assess individual teachers. *Learning and Individual Differences, 58*, 46–55.
Van der Lans, R. M., Van de Grift, W. J. C. M., & Van Veen, K. (2018). Developing an instrument for teacher feedback: Using the Rasch model to explore teachers' development of effective teaching strategies and behaviors. *The Journal of Experimental Education, 86*(2), 247–264.
Viera, A. J., & Garret, J. M. (2015). Understanding interobserver agreement: The kappa statistic. *Family Medicine, 37*(5), 360–363.
Von Davier, M. (1994). *WINMIRA 200.1*. IPN.
Vygotsky, L. S. (1930). *Mind and society*. Harvard University Press.
Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450.

Prof. Dr. Wim van de Grift (1951) is emeritus professor in Educational Sciences at the University of Groningen. He was director of the Teacher Training Institute of the University of Groningen and was scientific advisor of the Inspectorate of Education in the Netherlands. Van de Grift's research is aimed at the development and testing of theories on the professional development of teachers. This research program focusses on the following questions: How do teaching skills develop during the teaching career? Which factors influence the development of teaching skills? What is the influence of the teachings skills of teachers on students' academic engagement and students' achievements?

Van de Grift studied psychology at Utrecht University and obtained in 1987 his doctoral degree at Leiden University with a dissertation on 'The role of the school leader in educational innovations'.

Work:

1978–1989: University of Amsterdam and Utrecht University.

1989–2016: Inspectorate of Education (Ministry of Education, Culture and Science).

2008–2016: University of Groningen.

2017–now: Director of his own company specialized in observing and coaching teachers.

Emeritus Prof. Okhwa Lee (since 2022 March). Department of Education of Chungbuk National University, South Korea. Prof. Ok-hwa Lee is a specialist in educational technology and a practitioner of teacher education. She has been a pioneer of the e-learning, technology applications in education and educational reform through smart education in Korea. She was a member of the Presidential Educational Reform Committee and the Presidential e-Government Committee of the Republic of Korea, also consulting members for various ministries regarding educational applications of technology. She has rich experiences of international collaborations with the Europe Erasmus mobility with Finland Sweden, Estonia, Netherlands and etc., long history of research collaboration with USA, Australia, Thailand and etc. Recently she collaborated with developing countries through the Korean government ODA (Official Development Assistant) programs to Sudan, Nigeria, Nicaragua, Vietnam, Ethiopia, Cambodia, Myanmar, and etc. Her work through ODA focused on teachers' capacity development of teaching skills using technology.

Prof. Seyeoung Chun is Professor Emeritus of Education at Chungnam National University, one of the major national universities in Daejeon, Korea. He received his education and Ph.D. from Seoul National University, South Korea, and has been actively engaged in education policy research and has held several key positions such as Secretary of Education to the President and CEO of KERIS. He founded the Smart Education Society in 2013, and has led many projects and initiatives for the paradigm shift of education in the digital era. Since his early career at the Korean National Commission for UNESCO, he has participated in many international cooperation projects and worked for several developing countries such as Nicaragua, Honduras, Cambodia, etc. *Education Miracle in the Republic of Korea* is the latest book to be published as a summary of his academic life.