Learning Environments Research (2021) 24:315–337
https://doi.org/10.1007/s10984-020-09329-8

ORIGINAL PAPER

Students' perceptions of teaching behaviour in Turkish secondary education: a Mokken Scaling of My Teacher Questionnaire

Sibel Telli[1,2] · **Ridwan Maulana**[2] · **Michelle Helms-Lorenz**[2]

Received: 12 December 2019 / Accepted: 28 July 2020 / Published online: 10 August 2020

Abstract
Teacher behaviour has significant impact on student learning and outcomes and determines the teaching quality in learning environments. The My Teacher Questionnaire (MTQ) has been used to assess students' perceptions of teaching behaviour in national and international studies with well-cited outcomes. In this cross-sectional survey study, we adjusted and shortened the MTQ for diverse settings in Turkish secondary education, using the nonparametric IRT model, Mokken Scaling (MS). The sample consisted of 12,036 students (grade 9–12, age 15–19 years) involving 446 classes/teachers from 24 general public high schools in Turkey. More than half of the students ($n=6544$, 54.40%) were females, while 306 students (2.5%) did not report their gender. The MS polytomous Double Monotonicity Model (DMM) was employed for scaling the individual student data. The ten selected items (MTQ10) showed a strong unidimensional structure ($H=0.61$) with good internal reliability (Cronbach's $\alpha=0.93$, Molenaar Sijtsma $\rho=0.93$) and construct validity. The final structure was tested on three random data sets and convergent validity of the MTQ10 was examined using student engagement in learning. The scale MTQ10 functioned well across various groups (random samples, genders, grades, subjects). Based on these results, MTQ10 reveals strong psychometric quality for the assessment of students' perceptions of teaching behaviour with the potential to deepen our understanding of teaching behaviours and teaching quality in Turkey.

Keywords Effective teaching behaviour · Mokken Scaling (MS) · Nonparametric Item Response Theory (NIRT) · Secondary education · Teacher behaviours · Teaching quality

✉ Sibel Telli
sibeltelli@comu.edu.tr

[1] Department of Mathematics and Science Education, Çanakkale Onsekiz Mart University, Çanakkale, Turkey

[2] Department of Teacher Education, University of Groningen, Groningen, The Netherlands

Introduction

Teaching quality has a pivotal influence on students' academic outcomes (Darling-Hammond 2000; Patrinos and Angrist 2018; UN Sustainable Development Goal (SDG) 4, 2019). Research shows that students' perceptions of their teacher behaviours are quintessential for describing perceived teaching quality in the intertwined dynamic psychosocial structure of learning environments (de Jong and Westerhof 2001; Levy et al. 2003; Maulana and Helms-Lorenz 2016a Seidel and Shavelson 2007). Teaching quality and teacher behaviours have been defined and operationationalised in different ways (Kyriakides et al. 2013; Muijs and Reynolds 2018; Reynolds et al. 2014; Scheerens 2016; Scheerens and Bosker 1997), implying that understanding teacher behaviour is still an ongoing process. This is mainly because of the complex dissonance between the educational setting and interrelated factors of teaching behaviours that should not be treated separately (Kim et al. 2019; Klassen and Tze 2014; Kyriakides et al. 2009; Seidel and Shavelson 2007). Overall, past studies define teaching quality as teachers' behaviour that has significant and positive impacts on student outcomes (Fauth et al. 2019; Hattie 2009; Holzberger et al. 2019; Kyriakides et al. 2020; Lee and Mamerow 2019; Maulana et al. 2015a; Maulana et al. 2016b). Therefore, understanding how students perceive their teachers' behaviour could strengthen teaching quality. However, studies of teacher professional development have revealed that teaching quality is exponentially related to the number of years of job experience; it remains relatively low among early career teachers and generally takes time to develop and reach a sufficient level (Brekelmans et al. 2005). Thus, countries such as Turkey face an additional challenge because of the high proportion of young teachers (TALIS 2018).

The present study was based on a framework from a theory-driven and evidence-based research approach of observable teaching behaviours (Maulana and Helms-Lorenz 2016a). According to this framework, observable teaching behaviours cover the six teaching domains of *Learning Climate (CLM), Classroom Management (ORG), Clarity of Instruction (CLR), Activating Teaching (ACT), Differentiation (DIF)* and *Teaching Learning Strategies (TLS)*. These six domains synthesise various research traditions including teacher effectiveness (Creemers 1994; Scheerens and Bosker 1997), learning environments (Opdenakker et al. 2012) and teacher support (Klem and Connell 2004). Adressing this conceptualisation, My Teacher Questionnarie (MTQ) was initially developed in the Netherlands and has been found useful for measuring perceived teaching behaviour in international research (de Jager et al. 2017; Inda-Caro et al. 2018; Maulana et al. 2019; van de Grift et al. 2017) and for preservice and inservice teacher professional development (Maulana et al. 2015a; Maulana and Helms-Lorenz 2016a; Maulana et al. 2016c; van de Grift 2014a; van de Grift et al. 2014b).

The Republic of Turkey aims to develop human capital for an improved future and the well-being of its citizens by investing in education (MEB 2017; World Bank 2011). One central way to achieve the goal is to support novice teachers' performance up to the level of experienced teachers by monitoring and coaching their teaching behaviours continuously. However, a valid, reliable, low-cost and user-friendly instrument to assess teaching behaviours is scarce in the Turkish context. The aim of the present study was to adapt the MTQ for use in Turkish secondary education.

Theoretical framework

Teacher behaviours in learning environments

The teacher is a crucial actor in educational settings [UN (United Nation) Sustainable Development Goals (SDG) 4 2019)]. Teaching behaviour makes a difference to students' engagement, learning, achievement and well-being (Martin and Dowson 2009; Muijs et al. 2014; Pineda-Báez et al. 2019; Reeve 2006). Researchers and practitioners agree that teacher behaviours are complex (Hattie 2009; Kyriakides et al. 2009; Muijs et al. 2014). Growing attention has been directed towards identifying components of teacher behaviours that have substantial impacts on students' outcomes (Scheerens 2016; Seidel and Shavelson 2007) and ways to promote sustainable teaching quality in learning environments (Harbour et al. 2015; Kyriakides et al. 2002; Maulana et al. 2019; Panayiotou et al. 2014).

Within the expanding educational research literature, van de Grift et al. (2014b) and Maulana et al. (2015a) unite theory and evidence-based practice to conceptualise and operationalise observable teaching behaviours into six domains. *Learning Climate (CLM)* is characterised by a psychosocially-safe learning environment that stimulates students' learning and development. This includes behaviours such as fostering respect, encouraging self-confidence, facilitating healthy interpersonal relationships, and providing a base for healthy growth. *Classroom Management (ORG)* illustrates teacher behaviours associated with efficient time management for students' activity and minimisation of physical and pschosocial barriers in teaching–learning time, while processing knowledge in an appropriate manner for students' comprehension level. *Clarity of Instruction (CLR)* deals with behaviours such as informing students about the lesson objectives and their expected gain, using multiple instructional strategies in a clear unity, facilitating students' prior knowledge, and checking whether lesson objectives are achieved and if students understand a given task as intended. *Activating Teaching (ACT)* indicates teaching behaviours that facilitate students' active learning. Activating students' knowledge makes them aware of the relevance of content to their learning and their expected performances. *Differentiation (DIF)* covers teaching behaviours related to higher-level operations and strategies at the cognitive and affective levels to support individual student needs to link existing and desired skills for their own learning and metacognition. This serves as a base for students to achieve higher-level cognitive skills. *Teaching Learning Strategies (TLS)* concern teacher behaviours that deliberately demonstrate, teach and scaffold learning processes aimed at to improving self-regulation of learning processes.

Empirical studies show that the six domains of teaching behaviour follow a stage-like order on a unidimensional continuum (van de Grift et al. 2014b; van der Lans et al. 2018, 2019). More-complex teaching behaviours require sufficient experience, practice and knowledge, even though a small number of novice teachers are cabaple of displaying highly-skilled teaching behaviours. The first three teaching behaviours (CLM, ORG, CLR) are viewed as basic competences for teaching, while the other three are viewed as more complex behavioural domains (van de Grift et al. 2014b).

Context of the study: Turkey

The Turkish Ministry of National Education (MEB) is responsible for the educational administration of the national curriculum. The third level of compulsory secondary

education, which is the focus of this study, is the four-year (15–19 age) educational context that prepares students for further study. The schooling at this level consist of 40 class hours per week that vary depending on the track, curriculum and elective courses (EURYDICE 2020). Over the years, significant improvements in education have been made in Turkey (MEB 2019a; TUK 2020). However, a number of educational challenges remain apparent, as revealed by international testing studies (i.e. PISA, TIMSS) (MEB 2019b). This suggests some needed alterations, hard work and roadmaps for developing teaching quality in general and understanding how the students perceive their teachers' behaviours (MEB 2017).

Recently, the Teaching and Learning International Survey (TALIS) of 2018 revealed that learning environments are perceived positively by Turkish students and teachers. Nevertheless, teachers reported that they spent 72% of classroom time on actual teaching and learning, which is lower than the OECD average (78%). In addition, teachers did not broadly use effective instructional practices such as student cognitive activation approaches (Burge et al. 2015). Turkish teachers' average age was 36 years, which is below the average of 44 years for the remaining 48 countries. Only 6% of Turkish teachers were aged 50 years and above (OECD average 34%). Alignment of these insights calls for rapid implementation of support programs for novice teachers' professional development and sustainable teaching quality in learning environments.

In Turkey, the majority of studies of teaching behaviours focus on effective, good or ideal teaching from students' perspectives (Telli et al. 2008) and teacher candidates and teachers (Bozkuş and Taştan 2016; Çakmak 2009; Karakelle 2005; Kozikoglu 2017). A recent study focused on effective teaching criteria in subject teaching, such as mathematics (Yıldırım and Yıldırım 2019). Jointly, some comparative studies have involved teachers' behaviour in terms of professional development (Özkan et al. 2019) and teacher questioning styles (Çalık and Aksu 2018). Although the aforementioned studies highlight the importance of positive learning environments and teacher behaviours in general, teaching quality was not an explicit focus in the secondary-education setting. Therefore, little is known about teaching behaviour in secondary education from students' perspectives. Student questionnaires have been recognised in the learning environment literature as highly valuable for tapping into what is happening in the classrooms based on the lens of students (de Jong and Westerhof 2001; Fraser 2012).

The nature of the teaching profession requires practical, yet theory-based, solutions (European Commission 2013; Ingvarson 2019). The present study is particularly important because it attempted to provide evidence regarding the psychometric quality of a student questionnaire that can be used to assess perceived teacher behaviours in secondary schools. In the long term, information gathered in this way could be used to enhance and support teaching quality and to increase the 'true' potential of the teacher's presence in real-time learning.

Research aims

To provide sustainable teaching quality in diverse learning environments, teaching behaviours should be supported and monitored in the professional context. Professional feedback should be provided to improve teaching (i.e. lesson studies, research lessons, professional learning communities). Knowing that higher levels of teaching quality are related to more teaching experience (van de Grift et al. 2014b), and that the Turkish teaching force

is younger than the OECD average (TALIS 2018), a practical, highly-reliable and valid measure is needed to provide prompt professional feedback, in real time, to boost teaching performance. We aimed to develop an instrument that is concise and at the same time adequately represents the construct of effective teaching behaviour. These practical characteristics are highly important in contemporary classroom assessments to maintain sufficient participation rates and reduce response fatigue (Brick 2018; Groves 2006). To our knowledge, a student questionnaire that meets the mentioned characteristics is not available yet in the Turkish context. The present study filled this gap by examining an existing, valid and reliable measure to tap perceived teaching behaviours (Maulana and Helms-Lorenz 2016a) and adapting it for use in the Turkish context. To reach this goal, we applied Mokken Scaling (MS).

Mokken Scaling (MS)

Test construction is based on one of two test theories: classical test theory (CTT) and Item Response Theory (IRT). The present study applied IRT, whose two main models, parametric (IRT, e.g. Rasch) and nonparametric (NIRT, e.g. Mokken), try to explain the structure in the manifest item and test responses by assuming the existence of a latent structure (θ) on which persons and items have a position. In this respect, both have the same assumptions. Hovewer, to do this, the parametric approach defines the shape of the Item Response Function (IRF) and transformations are used that result in measures on an equal interval scale, while the nonparametric approach explores measurement properties by evaluating the relationship between items and the latent structure (θ) being measured (i.e. kernel smoothing, isotonic regression estimation) (Meijer and Baneke 2004; Meijer et al. 2014). Thus, NIRT supports the interpretation of total scores (i.e. sum scores) to meaningfully order persons and items on the latent structure (θ) without any parametric transformations while identifying the unexpected answering behavior in response patterns. Several scholars recognise that these psychometric properties of NIRT are particularly useful in contexts in which the underlying response processes are not well understood, such as non-cognitive data and avoiding misleading results of parametric IRT models (Chernyshenko et al. 2001; Meijer and Baneke 2004). This is important for enhancing our understanding of different learning environments (e.g. multi, hybrid, in-formal) and explore the social, physical, psychological and pedagogical contexts in which learning occurs and which affect students' affective outcomes.

Mokken Scaling-MS (Mokken 1971) describes the relationship between trait scores and item responses, similar to the way in which IRT models explore the shape of IRF without forcing or matching a particular structure (i.e. logistic ogiveshape) that they do not have (Meijer et al. 2014; Molenaar 2004). Empirical data almost never satisfy the strong IRT model assumptions fully. NIRT (e.g. Mokken scalling) helps to explore the reasons why the data fit the model and it reveals the reasons why a specific logistic IRF model fails to fit the data (i.e. Meijer and Baneke 2004). NIRT also provides information about the psychometric quality of items in a particular population (Meijer et al. 2014). MS is based mainly on Guttman (1945) scaling and, because of its explorative nature, it is described as a probabilistic theory-driven NIRT (van Schuur 2003). MS provides advantages and flexibility to researchers for exploring the nature of data as long as basic ordering requirements are *consensus ad idem*. Additionally, the availability of frequently-updated software R with graphical features and the package *Mokken* (van der Ark. 2007, 2012) supports the popularity of MS among educational

researchers (Wind 2017, 2019). MS uses two NIRT models: the Monotone Homogeneity Model (MHM) based on three assumptions (monotonicity, unidimensionality and local independence); and a general and more strict Double Monotonicity Model (DMM) obtained by adding a fourth assumption, namely, evidence of Invariant Rater Ordering (IRO). Based on the same requirements, Molenaar (1982, 1997) proposed dichotomous and polytomous formulations of these two models by specifing the polytomous DMM with Item Step Response Functions (ISRFs).

Based on the theoretical outline above, we studied effective teaching from the perspective of observable teaching behaviour based on teacher effectiveness and learning environments frameworks (Maulana et al. 2015b; van de Grift 2007). MS polytomous DMM was employed to adopt the MTQ for assessing effective teaching behaviours in a limited time under diverse teaching conditions in Turkey.

Methods

Participants

The sample consisted of 12,036 students (Grade 9–12, age 15–19 years) from 446 classes/teachers from 24 coeducational general public schools accessible for students from various socio-economic backgrounds. Schools were located in two cities (7,995 students, 66.4%) and rural areas (4,041 students, 33.6%) from the highly-populated north-west part of the country (Marmara). This region geographically connects Europe and Asia. Each school participating in the study provided between 8 and 29 classes/teachers ($M = 18.85$, $SD = 4.78$). There were 8,458 students (70.3%) from 296 classes/teachers from one city and its districts and 3,578 students (29.7%), from 150 classes/teachers from the other city and its districts. More than half of the students ($N = 6$, 544, 54.40%) were females, while 306 students (2.5%) did not report their gender. According to national statistics, a total of 1,668,086 students (913,404 are female, 54.76%) attend general public secondary schools (MEB 2019a, p. 129). Therefore, the gender distrubution of our sample was representative of the country. Students are distributed by grades as follows: 4,248 (35.3%) in grade 9, 3,470 (28.8%) in grade 10, 2,905 (24.1%) in grade 11 and 1,413 (11.7%) in grade 12. The distribution by subject taught was: 4,784 students (39.7%) for *Beta Subjects*-science track (biology, chemistry, physic, mathematics); 4,259 students (35.4%) for *Alpha Subjects*-the language track (i.e. English, German, Turkish); 2,567 students (21.3%) for *Gama subjects*-Social sciences; 176 students (1.5%) for *Physical education*; and 220 students (1.8%) for *Music–Art track*. Class size varied from 7 to 39 students ($M = 26.29$, $SD = 6.31$).

Ethics approval was granted by the authorities concerned. Throughout the study, students, teachers and schools were randomly selected on a voluntary basis. All questionnaires were completed during normal class hours (40 min) without the presence of teachers. Data (with multiple measures) were collected in 2017 (9,046 students, 75.2%) and 2018 (2,990 students, 24.8%) during fall (October–December) and spring (March–May) as a part of the International Comparative Analysis of Learning and Teaching (ICALT3) project comparing the perceived teaching quality. This study focused only on modifying the MTQ for the Turkish secondary-education context.

Measures

Two instruments were used. The My Teacher Questionnaire-MTQ (Inda-Caro et al. 2019; Maulana and Helms-Lorenz 2016a) was the main instrument and measured students' perception of teaching behaviour. The Student Engagement Scale (Skinner et al. 2009) was a criterion measure for checking convergent validity considering the theoretical connection between the two constructs (Maulana and Helms-Lorenz 2016a). Response alternatives were on a 4-point Likert scale, with higher responses indicating higher quality levels. Surveys were conducted employing a paper and pencil method.

The MTQ contains 41 items measuring perceived behaviour in the six domains: Learning Climate (CLM) (5 items, e.g. "My teacher answers my questions", $\alpha = 0.75$); Classroom Management (ORG) (8 items, e.g. "My teacher applies clear rules", $\alpha = 0.83$); Clarity of Instruction (CLR) (7 items, e.g. "My teacher explains the purpose of the lesson clearly", $\alpha = 0.86$); Activating Teaching (ACT) (10 items, e.g. "My teacher encourages me to think", $\alpha = 0.86$); Differentiation (DIF) (4 items, e.g. "My teacher knows what I have difficulty with", $\alpha = 0.79$); and Teaching Learning Strategy (TLS) (7 items, e.g. "My teacher teaches me to check my solutions", $\alpha = 0.85$). Prior studies have shown that the items of the MTQ (Maulana et al. 2016c; van de Grift et al. 2014b) can be ordered in a unidimensional structure (Maulana et al. 2015a, 2019; van der Lans et al. 2019) and is valid and reliable across countries (Maulana et al. 2019).

Students' engagement was assessed using 10 items in two scales: Behavioural Engagement-BEHE (5 items; e.g. "In this class I pay attention", $\alpha = 0.84$), and Emotional Engagement-EMEN (5 items; e.g. "In this class I feel good.", $\alpha = 0.80$.)

Translation process

Following International Test Commission (ITC 2018) guidelines, instruments were translated separately from English into Turkish by two native Turkish speakers majoring in English as a Foreign Language (Translation-1). The translations were then double checked, proofread and finally back translated by three different independent experts who were qualified and experienced in these languages and knowledgeable about the instrument development and adaptation (Back translation-2). Translated items were checked for the content and the appropriateness of the translation. Concurrently, a Turkish secondary-school language teacher with over 15 years of teaching experience reviewed the measures for the semantic structure (Committee approach-3).

Through Translation-1 and Back translation-2, MTQ items were independently double checked with the original Dutch version (source for the translation) by a native Dutch speaker and a multilingual teacher educator. This combination was preferred for maximising the suitability of the test adaptation and recognising the differences (i.e. linguistic, cultural, psychological) and equivalence (Grisay 2003; van de Vijver and Tanzer 2004).

Data analysis

Descriptive statistics were generated for items and subscales. Construct validity of the MTQ involved (1) data examination, (2) scaling as recommended by Sijtsma and van der Ark (2017) and (Wind 2017) and (3) predictive validity (Crișan et al. 2020). For

the student engagement measure, Principal Component Analysis (PCA) (varimax rotation) and Confirmatory Factor Analysis (CFA) were performed with the ML estimator for both models using the R package 'lavaan'. Model fit was checked using the Comparative Fit Index (CFI), Tucker–Lewis Index (TLI), Standardised Root Mean Residual (SRMR) and Root Mean Square Error of Approximation (RMSEA). RMSEA < 0.05 and TLI and CFI close to 0.95 were considered to indicate good fit (Schreiber et al. 2006). Missing data were reported and deleted listwisely for both measures. Analysis was performed using the programs SPSS25 and R (version 3.6.1) and MSP5 for Windows (Groningen:ProGamma).

Results

(1) Data examination for the *MTQ* involved, first, the Kolmogorov–Smirnov test [df (9415) = 0.00. $p = 0.005$] which indicated that the data did not follow a normal distribution, but was skewed to the left in all cases. Second, the Graded Response Model (GRM), an extended IRT model for ordered polytomous observed variables, was applied to understand the response behaviours and how the set of items performed (Samejima 1968; Perner and Imiya 2005). The MTQ items were visualised using R package 'psychotree' to explore the unidimensionality assumption further (Maulana et al. 2015a, 2019; van der Lans et al. 2019). The total information estimated by this model indicated the presence of a nonnormal distribution with the highest frequency towards the maximum scores (3–4) in the data (Fig. 1).

Subsequently, the NIRT approach was carried out to identify the items which satisfied the four assumptions of *Unidimensionality (UD)* (all items are related to a single latent variable-θ), *Monotonicity (M)* (as person locations on the latent variable increase the probability for correct response, X = 1, does not decrease), *Local independence (LI)* (answers on items depend solely on the latent trait and not on some other characteristics of the individual or its environment), and *Non-intersecting ISRFs* (the conditional probability for a

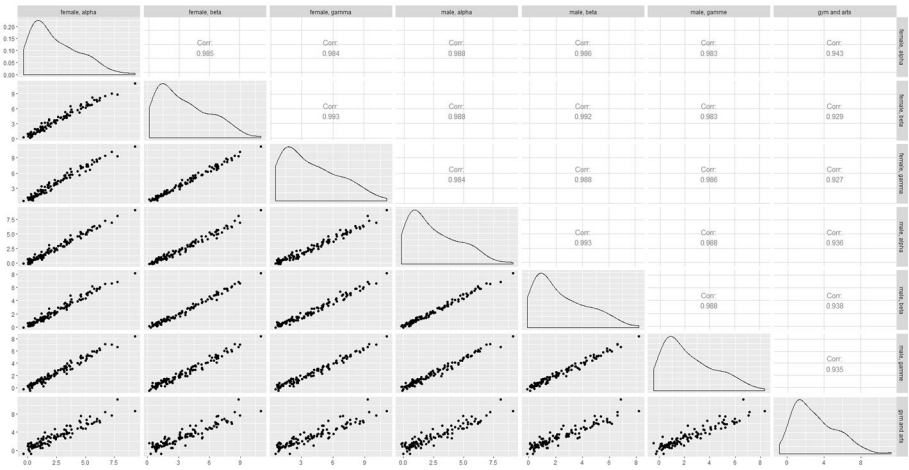


Fig. 1 Observed raters' response patterns across the classified groups, gender versus school subjects ($N_{student} = 12{,}036$, MTQ 41-iems)

rating in category k or higher on Item i has the same relative ordering across all values of the latent variable θ) using MS polytomous DMM. Under the DMM, IRFs can take on a variety of shapes as long as they do not intersect.

Regarding (2) *Scaling* during the first stage, the data were scanned for missing scores, inadmissible scores and outliers for MS. The number of Guttman errors showed that the Guttman pattern was consistent (Meijer et al. 2016). Missing values varied between 0.5 and 2.0% at the item level. The missing values were less than 5% and within acceptable range to be considered as missing at random (Tabachnick et al. 2013) and less than the figure of 10% that is unproblematic for MS (Sijtsma and van der Ark 2017). MS properties (i.e. element accuracy, scalability coefficients, and confidence intervals around scalability coefficients) have been shown to be sensitive to sample size. The applied sample (11,230 students, 6.7% missing) is sufficient to perform MS polytomous DMM with the real data (Crișan et al. 2020; Watson et al. 2018).

During the second stage, items were examined for scalability (H coefficient) and dimensionality using both the Automated Item Selection Procedure (AISP) and Generic Algorithm (g.a) in R 'package Mokken' because these different searching algorithms can provide different results (Meijer et al. 2015). The Loevinger's H coefficient indicated an unscalable scale if $H<0.3$, a weak scale if $0.3\leq H<0.4$, a medium scale if $0.4\leq H<0.5$, and a strong scale if $0.5\leq H\leq 1.0$ (Mokken 1971; Sijtsma and Molenaar 2002). Higher H values imply a more reliable ordering of both items and persons (Hemker and Sijtsma 1993). Items were selected stepwisely, consistent with the procedure proposed and taken by earlier initiatives for MS item reduction (Molenaar and Sijtsma 2000). H was used to select scales with both *Type* = *Search normal* and *Test*. The default settings were used in each algorithm. The procedure was run for positive constant c[1] initially set at $I_{LowH}=0.00$ as a control condition (Crișan et al. 2020; Meijer et al. 2015; Sijtsma et al. 2011) and then re-run with c increased by increments of 0.05 up to 0.80 (the upper bound 1). Meanwhile, some items were separated into more subscales (Hemker et al. 1995; Moorer et al. 2001). The H value at each c value and the number of suggested scales were examined to confirm and test the unidimensional structure (Sijtsma and Molenaar 2002).

During the first round of item refining, this procedure was used to remove unscalable items [e.g. My teacher talks interestingly, *ACT,* H < 0.3, item-pair scalability (H_{ij}) and item scalability ($H_j=0.08$) were positive, $M_{Item}=2.65$] and the rest of 40 items were scaled into one dimension $H\leq 0.3$ (Scale: $H=0.52$, $\rho=0.97$, $H_j=0.42$–0.60). At c 0.40, two items were removed because of the lower bound (e.g. My teacher makes sure that I treat others with respect., LC, $H_j=0.37$) and the rest of the 38 items were scaled on one dimension (Scale: $H=0.54$, $\rho=0.97$, $Hj=0.42$–0.60). With c set at 0.50, seven items were removed because $H_j<0.50$ (e.g. My teacher makes sure that others treat me with respect, *LC,* $H_j=0.49$). The rest of the 31 items were classified into one dimension (Scale: $H=0.57$, $\rho=0.97$) while two items formed a second scale (Scale: $H=0.65$, $\rho=0.80$). These two items (My teacher lets me summarise the content of the lesson, TLS, and My teacher lets me explain the content of the lesson to other students, TLS) were removed stepwisely. Afterwards, 31 items at c 0.50 fitted the unidimensional measure (Scale: $H=0.56$, $\rho=0.97$) with H_j varying between 0.51 and 0.61 (strong scale). During the second round, the item's factor loadings on the scale were calculated. The items with the lowest factor loading were deleted stepwisely when the scale internal consistency (Molenaar Sijtsma rho-ρ) was lower than 0.70 and $H\leq 0.50$ at c 0.50 (e.g. My teacher makes sure that I use my time effectively, ACT). During the third round, content-based and item correlations were examined to identify redundant items. If items were similar in content or in the same domain (van de Grift 2007, 2014a), the item with the lowest H score was deleted (e.g. My teacher answers my

questions, LC, $H_j = 0.51$). After removing 12 items stepwisely, 19 items remained for further evaluation.

During the third round, monotonicity and local independence assumptions were examined. The last assumption, ISRFs, was checked based on PMatrix information. Nine items with *Crit* > 80 (for *Crit* see foodnote 2) showed a strong violation and were discarded in succession (e.g. My teacher motivates me, ACL, *Crit* 116) (Molenaar and Sijtsma 2000). After these rounds of item reduction, 10 items (MTQ10) remained and fitted the unidimensional structure and satisfied all assumptions for the MS polytomous DMM (Table 1).

During the third stage, scale properties were investigated. MS provides the scale reliability statistic, Molenaar Sijtsma rho-(ρ), which is comparable to Cronbach's α (Molenaar and Sijtsma 1984). A value of $\rho > 0.7$ is considered acceptable (Kline 2000; Nunnally and Bernstein 1994). Items generally scored higher ($M = 2.98$, $SD = 0.0072$, skewness = −0.488 $SD = 0.023$, kurtosis = −0.559 $SD = 0.046$). Cronbach's α and rho-(ρ) were 0.93. MTQ10 properties are presented in Table 2. Meeting these four assumptions provides evidence that the MTQ10 is sufficiently unidimensional, represents the teaching behaviour (construct) more concisely and is reliable (Wind 2019).

Furthermore, scale equivalance was examined (ITC 2018, p. 116). MTQ10 satisfied all MS polytomous DMM assumptions, which indicates that the 10-item set does not exhibit Differential Item Functioning (DIF) (Moorer et al. 2001). Thus, the analysis was extended with Differential Scale Functioning (DSF). The scale was tested with three randomly-formed subsamples ($N1_{student} = 3{,}744$; $N2_{student} = 3{,}743$; $N3_{student} = 3{,}743$) according to grade level, school subject and gender for equal functioning to determine whether the scale composition and properties are generalisable. There was no indication of DSF across these groups. Results for the school subjects are given in Fig. 2.

Eventually, (3) predictive validity, the MTQ10 was validated by *consensus ad idem* (Downing 2003; Nunnally and Bernstein 1994). Initially, face and content validity were examined by an expert group. Next, MTQ10 met the four assumptions of MS polytomous DMM's and shows the unidimensional structure with high reliability (α and $\rho = 0.93$) (Wind 2019). Irrespective of these results, researchers agreed to cross-validate the MS results as *sine qua non* of assessment (Crișan et al. 2020). Thus, the predictive validity was determined between the measures, MTQ10 and Student Engagement (criterion measure).

For Student engagement, first, PCA was performed on the 11388X10 matrix (5.4% missing). The Bartlett-test (χ^2 (45) = 48,738,334 $p < 0.000$) and the Kaiser–Meyer–Olkin measure (KMO = 0.854) were suitable for PCA (Field 2009). The scales correlated with each other ($r = 0.607$, $p < 0.001$) and explained 59.02% of the variance in Model 1 (10 items, see Table 3). All the items fell into their respective factors with two exceptions. The Behavioral Engagement BEHE item "In this class, I participate in class discussion" loaded on Emotional Engagement-EMEN (0.43) and marginally (0.32) on the BEHE. The exact opposite pattern was found for the EMEN item "In this class, when we work on something, I feel interested." which was loaded on the BEHE (0.44) and loaded marginally on the EMEN (0.34).

It is possible that many students interpreted this item as a mixture of behavioural and emotional engagement as the word 'interested' in Turkish language and culture also implies an affective state. Second, these two items were excluded from the analysis and PCA was performed on the 11513X8 matrix (4.35% missing). The Bartlett-test (χ^2 (28) = 39,043.277 $p < 0.000$) and the Kaiser–Meyer–Olkin measure (KMO = 0.816) and BEHE $\alpha = 0.84$, EMEN $\alpha = 0.76$ were satisfactory. The scales correlated with each other ($r = 0.50$, $p < 0.001$) and explained 64.43% of the variance in Model 2 (after removing two cross loaded items, 8 items, see Table 3). Considering the discussion about the Kaiser

Table 1 Summary of items and monotonicity [a] checks for 10-item of My Teacher Questionnarie (MTQ10) and its item distribution over the six domains of original MTQ

Domains (Inda-Caro et al. 2019)	Item	Item scalability Hj	Restcores
Clarity of Instruction (CLR)	My teacher makes sure that I keep on working	0.57	67
Teaching Learning Strategy (TLS)	My teacher teaches me to check my solutions	0.58	65
Teaching Learning Strategy (TLS)	My teacher tells how I should learn something	0.59	[b]
Learning Climate (CLM)	My teacher makes me feel self-confident with difficult tasks	0.60	[b]
Clarity of Instruction (CLR)	My teacher states the lesson objectives	0.62	43
Differentiation (DIF)	My teacher checks whether I have understood the content of the lesson	0.64	46
Activating Teaching (ACT)	My teacher motivates me	0.66	68
Differentiation (DIF)	My teacher knows what I have difficulty with	0.65	44
Activating Teaching (ACT)	My teacher makes sure that I do my best	0.62	27
Classroom Management (ORG)	My teacher involves me in the lesson	0.62	34

[a]There was no violation for Monocity and P-Matrix

[b]No violations

Table 2 Scale properties, univariate frequencies and item means and reliability for MTQ10 [$N_{student}$ = 11,230 ($N_{student}$ = 12,036, missing 806, 6.7%)]

Item	*H*	*M*	*SD*	Frequency								Corrected item-total correlation	Cronbach's alpha if item deleted
				Never		Seldom		Frequently		Often			
				n	%	*n*	%	*n*	%	*n*	%		
My teacher makes sure that I keep on working	0.58	3.10	0.92	775	6.9	1938	17.3	3856	34.3	4661	41.5	0.69	0.92
My teacher teaches me to check my solutions	0.57	2.75	1.05	1739	15.5	2663	23.7	3441	30.6	3387	30.2	0.66	0.93
My teacher tells how I should learn something	0.62	3.03	0.95	937	8.3	2124	18.9	3848	34.3	4321	38.5	0.74	0.92
My teacher makes me feel self-confident with difficult tasks	0.62	2.90	1.07	1579	14.1	2284	20.3	3064	27.3	4303	38.3	0.74	0.92
My teacher states the lesson objectives	0.60	3.22	0.89	625	5.6	1639	14.6	3656	32.6	5310	47.3	0.69	0.92
My teacher checks whether I have understood the content of the lesson	0.62	3.07	0.94	831	7.4	2117	18.9	3673	32.7	4609	41.0	0.74	0.92
My teacher motivates me	0.66	3.05	0.96	945	8.4	2084	18.6	3634	32.4	4567	40.7	0.79	0.92
My teacher knows what I have difficulty with	0.64	2.70	1.04	1756	15.6	3011	26.8	3324	29.6	3139	28.0	0.74	0.92
My teacher makes sure that I do my best	0.65	2.86	1.04	1460	13.0	2556	22.8	3326	29.6	3888	34.6	0.78	0.92
My teacher involves me in the lesson	0.59	3.14	0.92	738	6.6	1870	16.7	3650	32.5	4972	44.3	0.69	0.92

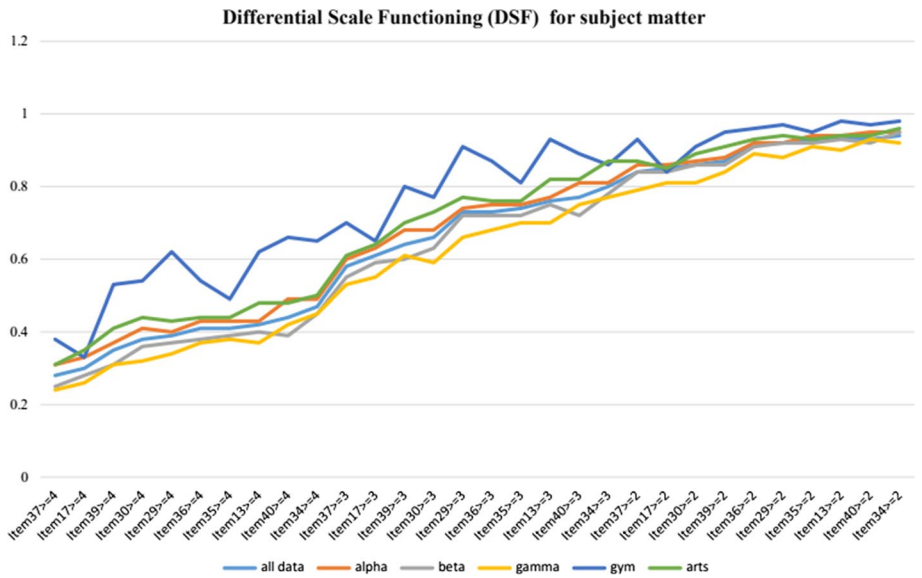


Fig. 2 MTQ10 Differential Scale Functioning (DSF) across school subjects

criterion (Fabrigar et al. 1999; O'connor 2000) and comments about checking the unidimensional structure in multiple ways (Ziegler and Hagemann 2015), we conducted Horn Parallel Analysis (Horn 1965) that is considered among the most accurate methods (Dinno 2009; Glorfeld 1995).

Parallel analysis involves extracting eigenvalues from random data (Horn 1965) and Glorfeld's (1995) extension. For this study, the Horn Parallel Analysis (Horn 1965) was performed using R package 'paran' which showed that the two-factor structure was retained for Model 1 and Model 2 (Fig. 3).

Third, CFA was performed (R package Lavaan). The fit indicated slightly lower values for CFI and TLI, but a high value for RMSEA (Schreiber et al. 2006). Results for both models (see Skinner et al. 2009 for details) are presented in Table 3. The person correlation coefficients (varies between 0.41 and 0.47) and the Corrected Attenuation-CA[3] (varies between 0.52 and 0.64) (Spearman 1904; Nunnally and Bernstein 1994, p. 240–241) were calculated between the MTQ10 and Student Engagement (two models) for predictive validity (see Table 4).

Discussion

This study's aim was to shorten the MTQ (Inda-Caro et al. 2019; Maulana and Helms-Lorenz 2016a) to assess perceived teaching behaviours in Turkey. When the MS polytomous DMM was applied, the resulting MTQ10 showed strong psychometric characteristics, internal consistency and construct validity. Its unidimensional structure is consistent with previous findings and the original version of MTQ (Maulana et al. 2015a, 2019; van der Lans et al. 2019) and is consistent across groups (random samples, gender, subject, grade level). MTQ10 met all the MS polytomous DMM assumptions. The observed violations of monotonicity were minor [Table 1, *Crit* (see footnote 2) less than 80], which could be

Table 3 Descriptive information for student engagement, Cronbach α and CFA goodness-of-fit indices

Model	Engagement type	Item	Component		*N*	*M*	*SD*	α	R*	*N*	χ^2	*df*	CFI	TLI	RMSEA	SRMR
		In this class……	1	2												
Model 1-10 items	BEHE	I pay attention	0.85		11,690	3.07	0.64	0.82	0.61	11,388	35624.975	45	0.834	0.780	0.145	0.086
		I listen very carefully	0.84													
		I try hard to do well	0.82													
		I work as hard as I can	0.79													
		I participate in class discussions		0.43												
	EMEN	it's fun		0.90	11,631	2.93	0.70	0.78								
		I feel good		0.78												
		When we work on something, I get involved		0.68												
		I enjoy learning new things		0.57												
		When we work on something, I feel interested	0.44													
Model 2-8 items	BEHE				11,751	3.14	0.66	0.84	0.50	11,513	27739.242	28	0.880	0.823	0.147	0.074
	EMEN				11,713	2.86	0.76	0.76								

*Correlation is significant at the 0.01 level (2-tailed). All chi-square values significant at the $p < .001$

BEHE Behavioural Engagement

EMEN Emotional Engagement

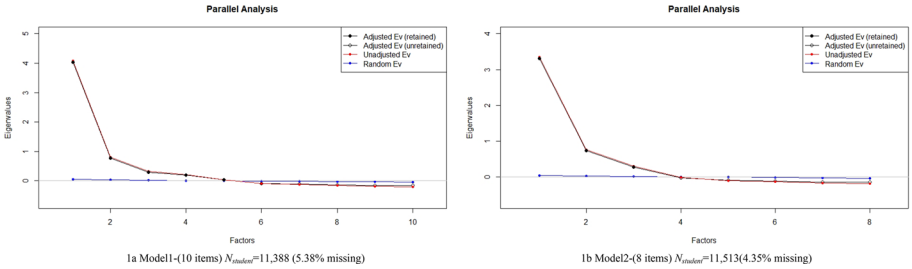


Fig. 3 Plot for Horn's Parallel Analysis for Model 1 (1a) and Model 2(1b)

Table 4 Pearson correlations and the Corrected Attenuation (CA) between MTQ10 and student engagement (two models)

Model engagement type	MTQ10	
	Correlation*	Corrected Attenuation
Model 1		
BEHE	0.44	0.57
EMEN	0.47	0.64
Model 2		
BEHE	0.41	0.52
EMEN	0.45	0.63

*Correlation is significant at the 0.01 level (2-tailed)

BEHE Behavioural Engagement, EMEN Emotional Engagement

because of sampling fluctuations (Molenaar and Sijtsma 2000), and the Guttman Error in Response Pattern was consistent (Meijer et al. 2016, see Appendix). MTQ10 had adequate validity and strong reliability (Cronbach's α and Molenaar Sijtsma rho-ρ are 0.93).

The applied methodology, MS polytomous DMM, confirmed the pyschometric quality of the MTQ10. Firstly, parametric IRT, item factor analysis was used to test the assumption of unidimensionality (Reise and Waller 2009). MS selects items that circumvent the assumption by upper and lower asymptotes, because the H coefficient is used as a criterion for including items in a scale. Items with asymptotes substantially different from 0 and/or 1 were rejected stepwisely (for not being discriminating enough). This means that the ceiling-floor effects are eliminated. Therefore, in the NIRT literature, it is suggested that nonparametric approaches for assessing unidimensionality are preferred over parametric ones (Meijer and Baneke 2004; Sijtsma and Molenaar 2002). Second, students' and teachers' preferences for short questionnaires are well recognised (Maulana and Helms-Lorenz 2016a; Maulana et al. 2019). However, reliability increases with the test length and the shorter tests often consist of items with relatively low inter-item correlations. It could be difficult to optimise both reliability and predictive validity at the same time (Magnusson 1967; Nunnally and Bernstein 1994).

In the context of MS, H_i values or discrimination parameters optimise both predictive validity (through content heterogeneity) and reliability (through test length) (Crișan et al. 2020). In this respect, the present study shows that the psychometric quality of MTQ10 is sufficiently strong to give prompt feedbacks to teachers. Third, MS, similar to other NIRT methods, measures constructs at the ordinal level (categorical variable). Hovewer,

the distinction between continuous and categorical variables is not always clear-cut (Tabachnick et al. 2013, p. 7, 204). Data collected with the MTQ10, which satisfies the MS assumptions, can be treated directly as a continuous distribution which is straightforward and easy to apply in practice. Last, educational settings are intertwined dynamic systems and difficult to disentangle. The necessity of *system thinking*[4] is evident for understanding the basic level of the setting. Research has shown the theoretical and empirical links in this structure (i.e. student engagement relates to teaching quality and ultimately learning outcomes) (Maulana and Helms-Lorenz 2016a; Pianta et al. 2012). Thus, in addition to construct validity, evidence of the predictive validity of MTQ10 for student engagement (Skinner et al. 2009) also was established. The results (Fig. 3) confirm the theoretical structure, with all the items falling into their respective factors with at least three loadings (Zwick and Velicer 1986), with only two exceptions in the PCA results. Table 3 provides descriptive statistics, correlations and the fit parameters for Model 1 and Model 2 (Skinner et al. 2009).

The concatenated psychosocial components in the educational settings could be reliable at a certain measurement time, but they might fluctuate over short periods of time. This cross-sectional survey study did not include this possible fluctuation over time (Akkerman and Bakker 2019; Christenson and Reschly 2012; Downings 2003; Reeve and Lee 2014). Student engagement measures might suffer from the chosen research design by revealing lower Pearson's correlation coefficient, while the Corrected Attenuation (CA) (Nunnally and Bernstein 1994) generally showed adequate convergent validity ($r \geq 0.60$, Hinkle 2003, see Table 4).

Additionally, as noted by Skinner et al. (2009), more multidimensionality could be present than was identifed in their study. These are the most probable reason for the slightly-lower values of fit indices. Considering that validation is a continuing process (Messick 1995), further studies with MTQ10 would benefit from more powerful approaches with instrument batteries, such as the Multitrait-MultiMethod Matrix-MTMM (Campbell and Fiske 1959) and the clinical study design (i.e. retrospective analysis and time series to understand the puzzle in perceived teaching behaviours more comprehensively.)

In summary, understanding student perceptions of teaching behaviour requires advanced knowledge and statistics to analyse large data sets with fine measures. Future research should investigate the response process and how participants benefit from the interaction over time to undertand the obstacles for teaching quality. Such research designs should consider how to tackle the procedures that can fluctuate and develop over time while controlling or eliminating the error sources associated with the test takers (Lüdtke et al. 2009; Mainhard et al. 2019). Students' perceptions of teaching behaviour provide new opportunities for providing real-time feedback to teachers to boost their own teaching behaviours in learning environments or in co-operation with coaches. In this manner, teaching practice, especially on the complex level of teaching behaviours, can benefit from tailored interventions.

The MTQ10 is also subject to limitations. First, perhaps the most important limitation is that the MS procedure for item selection is sample dependent. Although multiple methods were employed in the item-refining process (see Results), the sample was very large and representative of the secondary-school context[5] in Turkey, and the MTQ10 performed well in terms of DSF, results still could be sensitive to the population and learning environment (i.e. laboratory, classroom, outdoor). Moreover, methodologically, Meijer and Egberink (2012) strongly advised that care be taken to investigate the inclusion and exclusion of outliers in their sample because H is sensitive to outliers. This means that researchers should carefully examine the data before performing any

analyses. Second, although this study assumed that general teaching behaviours apply to all subjects, the MTQ10 does not cover any subject-specific teaching factors. Hence, investigations of subject-specific didactics could require the inclusion of subject-specific measures. Third, this study ignored possible sources of bias in sampling fluctuation, test takers' response behaviours, and their perceptions in responses (ITC 2018; Mokken 1971).

In conclusion, particularly because of its strong content validity (Downings 2003; Nunnally and Bernstein 1994), the MTQ10 is a robust and practical measure for assessing perceived teaching behaviours in secondary schools in Turkey. Overall, the teaching profession is traditionally a highly-respected profession in Turkish society (Dolton et al. 2018). The profession faces universal transformations (Dijkema et al. 2019; Papanastasiou and Karagiorgi 2019). Todays' teachers need viable collaboration and professional feedback to educate not only future citizens, but also their future colleagues. Also more practice-oriented training is anticipated. The MTQ10 has the potential to deepen our understanding of students' perceptions of teaching behaviour. It can be used to assess, formulate and set tailored interventions. The results of the present study are anticipated to support the teaching profession and contribute to understanding of teaching behaviours as perceived by students.

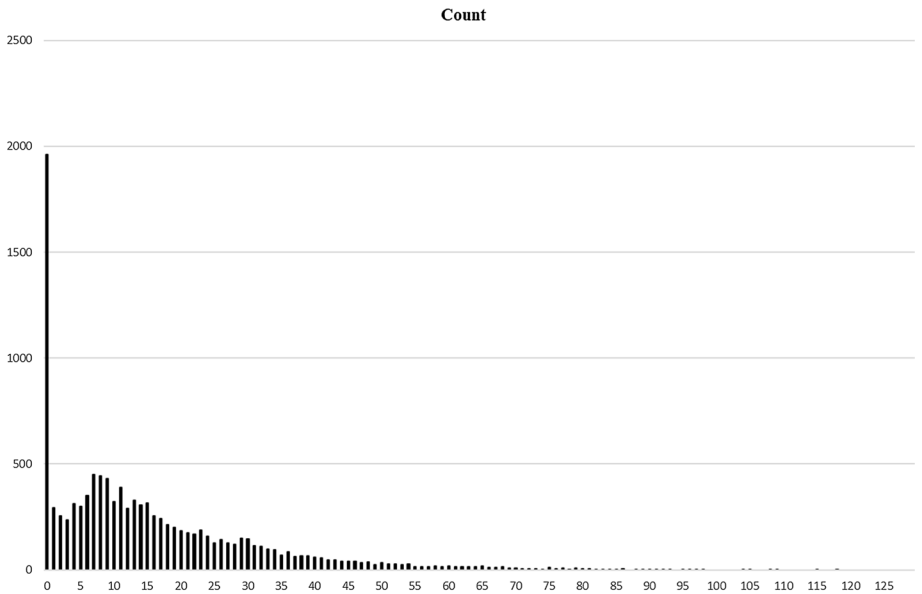
Notes

1. Sijtsma and Molenaar (2002, p. 68) define a Mokken scale as a set of items for measuring a common trait that is determined by reasonable discriminative power c that is a user-specified value. Reasonable discriminative power is defined by a lower bound $c = 0.3$ that is not strictly necessary. $c = 0$ provides interesting information about which items comply to the minimum requirements of the Monotone Homogeneity Model (MHM). Intermediate values of c are between 0.40 and 0.60 (Meijer et al. 2014; Sijtsma and Molenaar 2002, p. 68).
2. *Crit* is an effect size measure, a critical value, calculated by summing the coefficient values of ItemH, #ac, #vi, #vi/#ac, maxvi, sum, sum/#ac, zmax and #zsig into a single statistic (Molenaar and Sijtsma 2000, p. 74). If $Crit > 80$, there is serious doubt about the validity of the model for this item. If $Crit < 40$, the violations reported could be ascribed to sampling variation. If $Crit \leq 40$ and ≥ 80, a decision can depend on further consideration of the pros and cons. *Crit* values provide an idea about the seriousness of model violations in the data analyses (Meijer et al 2014).
3. Corrected Attenuation (CA) (Disattenuation) is a statistical procedure developed by Charles Spearman in 1904 to allow researchers to estimate the relationship between two constructs as if they were measured perfectly reliably and free from random errors that occur in all observed measures (Nunnally and Bernstein 1994).
4. System thinking is the ability to understand how an entire system works; how an action, change or malfunction in one part of the system affects the rest of the system; and adopting a 'big picture' perspective on work. It includes judgement and decision-making, system analysis and system evaluation, as well as abstract reasoning about how the different elements of a work process interact (NRC 2010, p. 63–64).
5. Schools were invited to participate in the study by providing at least 20 teachers with their one class. Schools with less than 20 classes were also invited but, in those cases, all teachers were asked to participate voluntarily.

Acknowledgements The authors wish to acknowledge and thank Peter Moorer for his assistance and guidance in data analysis, as well as the students, teachers and schools for their voluntary participation in this study.

Funding This paper is a part of the project International Comparative Analysis of Learning and Teaching (ICALT3)/Differentiation on Teaching Quality from an International Perspective initiated by researcher from the Teacher Education Department of the University of Groningen (the Netherlands). Data (with multiple measures) were collected. The work is partially supported by the Netherlands Initiative for Education Research (NRO) under Grant [NRO, Grant Number: 405-15-732] by the second and third authors and partially self-funded by the first author.

Appendix



Guttman Error in Response Pattern for MTQ10 ($N_{student}$=11230, 6.7% missing)

References

Akkerman, S. F., & Bakker, A. (2019). Persons pursuing multiple objects of interest in multiple contexts. *European Journal of Psychology of Education, 34*(1), 1–24.

Bozkuş, K., & Taştan, M. (2016). Teacher opinions about qualities of effective teaching. *Pegem Journal of Education and Instruction, 6*(4), 469–490.

Brekelmans, M., Wubbels, T., & van Tartwijk, J. (2005). Teacher–student relationships across the teaching career. *International Journal of Educational Research, 43*(1–2), 55–71.

Brick, J. M. (2018). Sampling to minimize nonresponse bias. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave handbook of survey research* (pp. 23–28). New York: Palgrave Macmillan. https://doi.org/10.1007/978-3-319-54395-6_5.

Burge, B., Lenkeit, J., & Sizmur, J. (2015). *PISA in practice: Cognitive activation in maths*. Slough: NFER.

Çakmak, M. (2009). Prospective teachers' thoughts on characteristics of an "effective teacher". *Education and Science, 34*(153), 76–82.

Çalık, B., & Aksu, M. (2018). A systematic review of teachers' questioning in Turkey between 2000–2018. *İlköğretim Online, 17*(3), 1548–1565.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56*(2), 81–105.

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*(4), 523–562.

Christenson, S., & Reschly, A. L. (2012). Jingle, jangle, and conceptual haziness: Evolution and future directions of the engagement construct. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 3–19). New York: Springer.

Creemers, B. P. M. (1994). *The effective classroom*. London: Cassell.

Crişan, D. R., Tendeiro, J. N., & Meijer, R. R. (2020). On the Practical Consequences of Misfit in Mokken Scaling. *Applied Psychological Measurement*, *44*(6), 482–496.

Darling-Hammond, L. (2000). Teacher quality and student achievement. *Education Policy Analysis Archives, 8*(1), 1–42.

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.

de Jong, R., & Westerhof, K. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research, 4,* 51–85.

Dijkema, S., Doolaard, S., Ritzema, E. S., & Bosker, R. J. (2019). Ready for take-off? The relation between teaching behavior and teaching experience of Dutch beginning primary school teachers with different educational backgrounds. *Teaching and Teacher Education*, *86*, Article 102914.

Dinno, A. (2009). Implementing Horn's parallel analysis for principal component analysis and factor analysis. *The Stata Journal, 9*(2), 291–298.

Dolton, P., Marcenaro, O., Vries, R. D., & She, P. W. (2018). *Global Teacher Status Index 2018*. Retrieved May 3, 2020, from http://repositorio.minedu.gob.pe/handle/MINEDU/6046.

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education, 37*(9), 830–837.

European Commission. (2013). *Supporting teacher competence development for better learning outcomes.* Brussels, Belgium. Retrieved May 3, 2020, from https://ec.europa.eu/assets/eac/education/experts-groups/2011-2013/teacher/teachercomp_en.pdf.

EURYDICE. (2020). *National Education Systems*. Retrieved May 3, 2020, from https://eacea.ec.europa.eu/national-policies/eurydice/content/turkey_en.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272.

Fauth, B., Decristan, J., Decker, A. T., Büttner, G., Hardy, I., Klieme, E., et al. (2019). The effects of teacher competence on student outcomes in elementary science education: The mediating role of teaching quality. *Teaching and Teacher Education*, *86*, Article 102882.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.

Fraser, B. J. (2012). *Classroom environment*. London: Routledge.

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement, 55,* 377–393.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing, 20*(2), 225–240. https://doi.org/10.1191/0265532203lt254oa.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly, 70*(5), 646–675.
Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*(4), 255–282.
Harbour, K. E., Evanovich, L. L., Sweigart, C. A., & Hughes, L. E. (2015). A brief review of effective teaching practices that maximize student engagement. *Preventing School Failure: Alternative Education for Children and Youth, 59*(1), 5–13.
Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.
Hemker, B. T., & Sijtsma, K. (1993). A practical comparison between the weighted and the unweighted *H*-coefficients of the Mokken model. *Kwantitatieve Methoden: Nieuwsbrief voor Toegepaste Statistiek en Operationele Research, 14*(44), 59–73.
Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken I RT model. *Applied Psychological Measurement, 19*(4), 337–352.
Hinkle, D. E. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston, MA: Houghton Mifflin.
Holzberger, D., Praetorius, A. K., Seidel, T., & Kunter, M. (2019). Identifying effective teachers: The relation between teaching profiles and students' development in achievement and enjoyment. *European Journal of Psychology of Education, 34*(4), 801–823.
Horn, J. L. (1965). A rationale for the number of factors in factor analysis. *Psychometrika, 30,* 179–185.
Inda-Caro, M., Maulana, R., Fernández-García, C. M., Peña-Calvo, J. V., del Carmen Rodríguez-Menéndez, M., & Helms-Lorenz, M. (2019). Validating a model of effective teaching behaviour and student engagement: Perspectives from Spanish students. *Learning Environments Research, 22*(2), 229–251.
Ingvarson, L. (2019). Teaching standards and the promotion of quality teaching. *European Journal of Education, 54*(3), 337–355.
International Test Commission (ITC). (2018). ICT guidelines for translating and adapting tests (second edition). *International Journal of Testing, 18*(2), 101–134. https://doi.org/10.1080/15305058.2017.1398166.
Karakelle, S. (2005). Analyzing teachers' definitions of effective teachers according to effective teaching dimensions. *Education and Science, 30*(135), 1–10.
Kim, L. E., Jörg, V., & Klassen, R. M. (2019). A meta-analysis of the effects of teacher personality on teacher effectiveness and burnout. *Educational Psychology Review, 31,* 163–195.
Klassen, R. M., & Tze, V. M. (2014). Teachers' self-efficacy, personality, and teaching effectiveness: A meta-analysis. *Educational Research Review, 12,* 59–76.
Klem, A. M., & Connell, J. P. (2004). Relationships matter: Linking teacher support to student engagement and achievement. *Journal of School Health, 74*(7), 262–273.
Kline, P. (2000). *The handbook of psychological testing*. Hove: Psychology Press.
Kozikoglu, I. (2017). Prospective teachers' cognitive constructs concerning ideal teacher qualifications: A phenomenological analysis based on repertory grid technique. *International Journal of Instruction, 10*(3), 63–78.
Kyriakides, L., Anthimou, M., & Panayiotou, A. (2020). Searching for the impact of teacher behavior on promoting students' cognitive and metacognitive skills. *Studies in Educational Evaluation*, *64*, Article 100810.
Kyriakides, L., Campbell, R. J., & Christofidou, E. (2002). Generating criteria for measuring teacher effectiveness through a self-evaluation approach: A complementary way of measuring teacher effectiveness. *School Effectiveness and School Improvement, 13*(3), 291–325.
Kyriakides, L., Christoforou, C., & Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education, 36,* 143–152.
Kyriakides, L., Creemers, B. P., & Antoniou, P. (2009). Teacher behaviour and student outcomes: Suggestions for research on teacher training and professional development. *Teaching and Teacher Education, 25*(1), 12–23.
Lee, S. W., & Mamerow, G. (2019). Understanding the role cumulative exposure to highly qualified science teachers plays in students' educational pathways. *Journal of Research in Science Teaching, 56,* 1362–1383.
Levy, J., Wubbels, T., den Brok, P., & Brekelmans, M. (2003). Students' perceptions of interpersonal aspects of the learning environment. *Learning Environments Research, 6*(1), 5–36.
Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology, 34*(2), 120–131.
Magnusson, D. (1967). *Test theory*. Reading, MA: Addison-Wesley.

Mainhard, T., Wubbels, T., & den Brok, P. (2019). Doing multilevel statistical modelling with hierarchically nested samples. In M. H. Hoveid, et al. (Eds.), *Doing educational research: Overcoming challenges in practice* (pp. 132–154). London: Sage.

Martin, A. J., & Dowson, M. (2009). Interpersonal relationships, motivation, engagement, and achievement: Yields for theory, current issues, and educational practice. *Review of Educational Research, 79*(1), 327–365.

Maulana, R., & Helms-Lorenz, M. (2016a). Observations and student perceptions of the quality of preservice teachers' teaching behaviour: Construct representation and predictive quality. *Learning Environments Research, 19*(3), 335–357. https://doi.org/10.1007/s10984-016-9215-8.

Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2015a). Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement, 26*(2), 169–194. https://doi.org/10.1080/09243453.2014.939198.

Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2015b). A longitudinal study of induction on the acceleration of growth in teaching quality of beginning teachers through the eyes of their students. *Teaching and Teacher Education, 51,* 225–245.

Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2016c). Validating a model of effective teaching behaviour of pre-service teachers. *Teachers and Teaching: Theory and Practice, 23*(4), 471–493. https://doi.org/10.1080/13540602.2016.1211102.

Maulana, R., Opdenakker, M. C., & Bosker, R. (2016b). Teachers' instructional behaviors as important predictors of academic motivation: Changes and links across the school year. *Learning and Individual Differences, 50,* 147–156.

Maulana, R., Smale-Jacobse, A., Helms-Lorenz, M., Chun, S., & Lee, O. (2019). Measuring differentiated instruction in The Netherlands and South Korea: Factor structure equivalence, correlates, and complexity level. *European Journal of Psychology of Education*, 1–29 **(Online)**.

MEB (Ministry of National Education). (2017). *Teaching strategy paper*. Retrieved May 3, 2020, from http://oygm.meb.gov.tr/meb_iys_dosyalar/2018_05/25170118_Teacher_Strategy_Paper_2017-2023.pdf.

MEB (Ministry of National Education). (2019a). *National Education Statistics Formal Education 2018/'19*. Retrieved May 3 2020. http://sgb.meb.gov.tr/meb_iys_dosyalar/2019_09/30102730_meb_istatistikleri_orgun_egitim_2018_2019.pdf.

MEB (Ministry of National Education). (2019b). *PISA 2018 National Report*. Retrieved May 3, 2020, from http://www.meb.gov.tr/meb_iys_dosyalar/2019_12/03105347_PISA_2018_Turkiye_On_Raporu.pdf

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modelling. *Psychological Methods, 9*(3), 354–368.

Meijer, R. R., & Egberink, I. J. (2012). Investigating invariant item ordering in personality and clinical scales: Some empirical findings and a discussion. *Educational and Psychological Measurement, 72*(4), 589–607.

Meijer, R. R., Niessen, A. S. M., & Tendeiro, J. N. (2016). A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics: Examples and a computer program. *Assessment, 23*(1), 52–62.

Meijer, R. R., Tendeiro, J. N., & Wanders, R. K. (2014). The use of nonparametric item response theory to explore data quality. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (1st ed., pp. 85–110). New York: Routledge.

Messick, S. (1995). Standards of validity and the validity of standards in performance asesssment. *Educational Measurement: Issues and Practice, 14*(4), 5–8.

Mokken, R. J. (1971). *Theory and procedure of scale analysis*. The Hague: Mouton.

Molenaar, I. W. (1982). Mokken scaling revisited. *Kwantitative Methoden, 3*(8), 145–164.

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369–380). New York: Springer.

Molenaar, I. W. (2004). About handy, handmade and handsome models. *Statistica Neerlandica, 58*(1), 1–20.

Molenaar, I. W., & Sijtsma, K. (1984). Internal consistency and reliability in Mokken's nonparametric item response model. *Tijdschrift voor Onderwijsresearch, 9*(5), 257–268.

Molenaar, I. W., & Sijtsma, K. (2000). *User's manual MSP5 for Windows*. Groningen: iecProGAMMA.

Moorer, P., Suurmeijer, T. P., Foets, M., & Molenaar, I. W. (2001). Psychometric properties of the RAND-36 among three chronic disease (multiple sclerosis, rheumatic diseases and COPD) in the Netherlands. *Quality of Life Research, 10*(7), 637.

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art–teacher effectiveness and professional learning. *School Effectiveness and School Improvement, 25*(2), 231–256.

Muijs, D., & Reynolds, D. (2018). *Effective teaching: Evidence and practice*. London: Sage.

National Research Council (NRC). (2010). *Standards for K-12 engineering education?*. Washington, DC: National Academies Press.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments & Computers, 32*(3), 396–402.

Opdenakker, M. C., Maulana, R., & den Brok, P. (2012). Teacher–student interpersonal relationships and academic motivation within one school year: Developmental changes and linkage. *School Effectiveness and School Improvement, 23*(1), 95–119.

Özkan, M., Özkan, Y. Ö., & Güvendir, M. A. (2019). Investigation of Turkey and Singapore schools in terms of teacher professional development and teacher behaviors hindering learning variables. *Education and Science, 44*(198), 309–325.

Panayiotou, A., Kyriakides, L., Creemers, B. P., McMahon, L., Vanlaar, G., Pfeifer, M., et al. (2014). Teacher behavior and student outcomes: Results of a European study. *Educational Assessment, Evaluation and Accountability, 26*(1), 73–93.

Papanastasiou, E. C., & Karagiorgi, Y. (2019). The involvement of school teachers in research-related activities: Extent, quality and predictors. *European Journal of Education, 54,* 621–634. https://doi.org/10.1111/ejed.12364.

Patrinos, H. A., & Angrist, N. (2018). *Global dataset on education quality: A review and update (2000–2017)* (English) (Policy Research working paper; no. WPS 8592). Washington, DC: World Bank Group. Retrieved May 3, 2020, from http://documents.worldbank.org/curated/en/390321538076747773/Global-Dataset-on-Education-Quality-A-Review-and-Update-2000-2017.

Perner, P., & Imiya, A. (Eds.). (2005). Machine learning and data mining in pattern recognition. In *Proceedings of the 4th international conference, MLDM 2005, Leipzig, Germany*. New York: Springer.

Pianta, R. C., Hamre, B. K., & Allen, J. P. (2012). Teacher–student relationships and engagement: Conceptualizing, measuring, and improving the capacity of classroom interactions. In S. Christenson, A. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement*. Boston, MA: Springer.

Pineda-Báez, C., Manzuoli, C. H., & Sánchez, A. V. (2019). Supporting student cognitive and agentic engagement: Students' voices. *International Journal of Educational Research, 96,* 81–90.

Reeve, J. (2006). Teachers as facilitators: What autonomy-supportive teachers do and why their students benefit. *The Elementary School Journal, 106*(3), 225–236.

Reeve, J., & Lee, W. (2014). Students' classroom engagement produces longitudinal changes in classroom motivation. *Journal of Educational Psychology, 106*(2), 527.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5,* 27–48.

Reynolds, D., Sammons, P., De Fraine, B., van Damme, J., Townsend, T., Teddlie, C., et al. (2014). Educational effectiveness research (EER): A state-of-the-art review. *School Effectiveness and School Improvement, 25*(2), 197–230.

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *ETS Research Bulletin Series*,1968(1), i-169.

Scheerens, J. (2016). *Educational effectiveness and ineffectiveness: A critical review of the knowledge base*. Dordrecht: Springer.

Scheerens, J., & Bosker, R. (1997). *The foundations of educational effectiveness*. London: Pergamon.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research, 99*(6), 323–338.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499.

Sijtsma, K., Meijer, R. R., & van der Ark, L. A. (2011). Mokken scale analysis as time goes by: An update for scaling practitioners. *Personality and Individual Differences, 50*(1), 31–37.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory* (Vol. 5). London: Sage.

Sijtsma, K., & van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology, 70*(1), 137–158.

Skinner, E. A., Kindermann, T. A., & Furrer, C. J. (2009). A motivational perspective on engagement and disaffection conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educational and Psychological Measurement, 69*(3), 493–525.
Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72–101. https://doi.org/10.2307/1412159.
Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.
TALIS (Teaching and Learning International Survey). (2018). *Country Notes, Turkey*. Retrived May 3, 2020, from http://www.oecd.org/education/talis/TALIS2018_CN_TUR.pdf.
Telli, S., den Brok, P., & Cakiroglu, J. (2008). Teachers' and students' perceptions of the ideal teacher. *Egitim ve Bilim, 33*(149), 118-129.
TUK (Turkish Statistical Institute). (2020). *Educational statistics*. Retrieved May 3, 2020, from http://tuik.gov.tr/PreTablo.do?alt_id¼1018.
UN (United Nations). (2019). *Sustainable Development Goals (SDG)*. Retrieved May 3, 2020, from https://www.un.org/sustainabledevelopment/education/.
van de Grift, W. (2007). Quality of teaching in four European countries: A review of the literature and application of an assessment instrument. *Educational Research, 49*(2), 127–152.
van de Grift, W. (2014a). Measuring teaching quality in several European countries. *School Effectiveness and School Improvement, 25*(3), 295–311. https://doi.org/10.1080/09243453.2013.794845.
van de Grift, W. J., 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., Helms-Lorenz, M., & Maulana, R. (2014b). 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 Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology, 54*(2), 119–135.
van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*(11), 1–19.
van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software, 48*(5), 1–27.
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. *Journal of Experimental Education, 86*(2), 247–264.
van der Lans, R. M., van de Grift, W. J., & van Veen, K. (2019). Same, similar or something completely different? Calibrating student surveys and classroom observations of teaching quality onto a common metric. *Educational Measurement: Issues and Practice, 38*(3), 55–56.
van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis, 11*(2), 139–163.
Watson, R., Egberink, I. J. L., Kirke, L., Tendeiro, J. N., & Doyle, F. (2018). What are the minimal sample size requirements for Mokken scaling? An empirical example with the Warwick–Edinburgh Mental Well Being Scale. *Health Psychology and Behavioral Medicine, 6*(1), 203–213.
Wind, S. A. (2017). An instructional module on Mokken scale analysis. *Educational Measurement: Issues and Practice, 36*(2), 50–66.
Wind, S. A. (2019). Nonparametric evidence of validity, reliability, and fairness for rater-mediated assessments: An illustration using Mokken scale analysis. *Journal of Educational Measurement, 56*(3), 478–504. https://doi.org/10.1111/jedm.12222.
World Bank. (2011). *Improving the quality and equity of basic education in Turkey: Challenges and options* (Report No. 54131-TR). Retrieved May 17, 2020, from http://documents.worldbank.org/curated/en/105971468338992381/pdf/541310SR0P107700Quality0Report02011.pdf.
Yıldırım, S., & Yıldırım, H. H. (2019). Conceptions of Turkish mathematics teachers about the effectiveness of classroom teaching. *International Journal of Mathematical Education in Science and Technology, 50*(8), 1152–1165.
Ziegler, M., & Hagemann, D. (2015). Testing the unidimensionality of items pitfalls and loopholes. *European Journal of Psychological Assessment, 31,* 231–237. https://doi.org/10.1027/1015-5759/a000309.
Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*(3), 432–442.