frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 24 July 2023
DOI 10.3389/feduc.2023.1068938

Check for updates

OPEN ACCESS

EDITED BY
Ariel Mariah Lindorff,
University of Oxford, United Kingdom

REVIEWED BY
Grzegorz Szumski,
University of Warsaw, Poland
Kizito Ndihokubwayo,
Iowa State University, United States

*CORRESPONDENCE
Ridwan Maulana
✉ r.maulana@rug.nl

RECEIVED 20 October 2022
ACCEPTED 03 July 2023
PUBLISHED 24 July 2023

CITATION
Maulana R, Kington A, Ko J, Feng X, Helms-Lorenz M, Looker B, Hibbert-Mayne K and Blackmore K (2023) Observing secondary school teachers' effective teaching behavior in the Netherlands, England, and the United States using the ICALT observation instrument. *Front. Educ.* 8:1068938. doi: 10.3389/feduc.2023.1068938

Observing secondary school teachers' effective teaching behavior in the Netherlands, England, and the United States using the ICALT observation instrument

Ridwan Maulana[1]*, Alison Kington[2], James Ko[3], Xiangyuan Feng[1], Michelle Helms-Lorenz[1], Benjamin Looker[2], Kimberley Hibbert-Mayne[2] and Karen Blackmore[2]

[1]Department of Teacher Education, University of Groningen, Groningen, Netherlands, [2]School of Education, University of Worcester, Worcester, United Kingdom, [3]Department of Education Policy and Leadership, The Education University of Hong Kong, Hong Kong, Hong Kong SAR, China

Introduction: The purpose of this study was to examine measurement invariance in observer scoring of effective teaching behavior in three secondary education contexts–the Netherlands, England, and the United States. It also aimed to describe what effective teaching behavior looks like in secondary education across the three education contexts.

Methods: A uniform observation measure called International Comparative Analysis of Learning and Teaching (ICALT) was used to observe teachers.

Results: Results revealed that the hypothesized factor structure of effective teaching behavior was confirmed for the Dutch and English data, but not for the US data. Teachers in the Netherlands showed higher levels of more basic teaching behaviors, but lower levels of more complex teaching behaviors, compared to teachers in England.

Discussion: Implications of the findings are discussed.

KEYWORDS

classroom observation, measurement invariance, effective teaching behavior, secondary education, cross-country comparison

1. Introduction

Research shows that effective teaching behavior plays a central role in student learning and outcomes (e.g., Chapman et al., 2012; Hattie, 2012). International large-scale studies such as the Program for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS) have provided useful insights regarding general trends in educational outcomes in secondary schools around the world (Martin et al., 2016c; Mullis et al., 2016). The PISA and TIMSS studies revealed variations in educational outcomes across education contexts (OECD, 2019). Explanations of differences in educational outcomes across education contexts and countries can be explained by various factors across a number of levels including students', classrooms', schools' and regional/national characteristics. In general, educational effectiveness research reveals that the classroom level matters the most. Particularly, about 15–25% of differences in student achievement can be explained by the work of teachers (van de Grift et al., 2017). Investigating teaching practices across education contexts may contribute to advance our understanding of variations in educational outcomes. However, little is known regarding differences in effective teaching behavior across education contexts because teaching is inadequately studied in a comparative fashion (Suter, 2019). Comparing teaching practices across contexts offers insights to stimulate cooperation across settings regarding best practices in effective teaching behavior for improved education quality globally (Adamson, 2012; Maulana et al., 2020), and for informing the continuous discourse about teaching and learning across various education contexts (Klette, 2022; Luoto, 2023).

A lack of knowledge regarding differences in effective teaching behavior across education contexts can be explained in several ways. First, the current literature on effective teaching behavior is still rather fragmented and most of the research was conducted in a single education context/country. Second, evaluation of teaching is typically executed using student reports, commonly called Student Evaluation of Teaching (SET, van der Lans et al., 2021), mainly due to the low cost and ease of administration (Maulana and Helms-Lorenz, 2016). Using an observation instrument to measure teaching behavior is unusual (Stroet et al., 2013), partly because it is viewed as costly and highly laborious (Maulana and Helms-Lorenz, 2016), but also because observing classroom teaching behavior is not a common practice around the world due ethics as well as culturally related values. In some contexts, classroom observation is highly valued and widely accepted by schools and teachers, while in others this can be viewed as intrusive and not culturally accepted (Maulana et al., 2023). Nevertheless, observation is seen as a more objective method to describe what is actually happening in the classroom compared to the more commonly used survey approach (Maulana and Helms-Lorenz, 2016).

Various classroom observation instruments exist, but little is known regarding their psychometric properties (i.e., measurement invariance) and potential for international comparisons. To date, there are at least two large-scale classroom observation studies aimed at investigating variations in teaching quality across education contexts. The Teaching and Learning International Survey (TALIS) Video Study, currently known as Global Teaching Insights Video Study, used the Global Teaching Insight (GTI) observation instrument to study teaching in mathematics classrooms across eight education contexts including Chile, Colombia, England (UK), Germany, Japan, Spain, Mexico, and China (OECD, 2020). A further study investigated effective teaching behavior using the Comparative Analysis of Learning and Teaching (ICALT; van de Grift, 2007) in natural classroom settings across school subjects and multiple contexts, including the Netherlands, Indonesia, Republic of Korea, Pakistan, South Africa, and Hong Kong SAR, China (Maulana et al., 2021). These two relatively recent studies offer a promising avenue for cross-country comparison in teaching behavior to stimulate knowledge exchange globally. The present study complements the previous work on effective teaching behavior by including three education contexts, including the Netherlands, England, and the US, and by using a uniform ICALT observation instrument. The aims of the study are twofold: (1) to examine whether the observation measure of teaching behavior can be meaningfully compared (i.e., measurement invariance) in secondary education in the three education contexts; and (2) to investigate differences in effective teaching behavior across the three education contexts.

To our knowledge, this is the first comparative study of teaching behavior using the same observation instrument conducted in these three, specific education contexts. It is our intention that the study will contribute to the discourse regarding the complex nature of teaching by investigating the efficacy of a uniform observation instrument (i.e., ICALT) applied to measure teaching behavior across these contexts.

2. Literature review

In general, there are three common methods for studying classroom teaching, including student surveys, teacher self-reports, and classroom observations (Maulana and Helms-Lorenz, 2016). The three methods have different underlying assumptions and considerations. The current study focuses on classroom observations using a standardized observation instrument.

Comparisons of teaching have been undertaken using a variety of approaches (see Adamson, 2012; Bray et al., 2014). Most straightforward, originating from the early 19th century is the direct comparison of two systems or country specific approaches to determine similarities or differences. An example of this was a comparison of differentiation as a pedagogical strategy to support learners studying for an International Baccalaureate (IB) in a two-center study in Hong Kong SAR, China and Australian schools (Dulfer and Akhlaghi Koopaei, 2021). While this classical approach is still used in educational research (Jortveit et al., 2020; Moberg et al., 2020; Goodwin and Low, 2021), other approaches have since developed. Another is to focus on a particular country as a point of reference, before then comparing other countries to the benchmarks created (Adamson, 2012; Powell, 2020). A third approach sees researchers comparing several countries equally and undertaking complex statistical analysis as a means of examining teaching effectiveness in the different contexts (Adamson, 2012; Powell, 2020). Measurement invariance (e.g., Millsap and Yun-Tein, 2004) is an essential aspect in this approach. With all three approaches it is important to note that research across different countries is only effective if the comparisons add to the field of research; there must be commonality between the compared countries and any international appropriation should be applied sensitively (Adamson, 2012). The current study follows a recent approach by Maulana et al. (2020), by taking the third approach with the ICALT observation instrument.

2.1. Observing effective teaching behavior

Research into effective education has been varied in its approach and areas covered, most notably curricula, teacher behaviors, policy making, leadership and self-efficacy. If effective education can be measured by academic gains and pupil achievement (Coe et al., 2014), then teaching behavior, which has been well documented as playing a significant role in student learning and outcomes (e.g., Scheerens and Bosker, 1997; Creemers and Kyriakides, 2008; Hattie and Clinton, 2008; Hattie, 2012; Chapman et al., 2016), is a central concept to effective education and teacher effectiveness. Particularly, classroom observations are highly valued in the teacher effectiveness research strand (Muijs et al., 2014; Bell et al., 2019).

It is only since the development of a variety of established observational instruments that the nuances of teacher behavior and quality have been examined. The first classroom observational instruments were developed in the early 1960s, when observational studies examining teacher quality first became popular (van de Grift, 2014). Despite the early instruments lacking validity and reliability, their use and exposure paved a route for the development of more refined and robust methods (Maulana et al., 2020). Since the development of the initial observation measures of teaching, many validated instruments have emerged, situated in strong theoretical and standardized frameworks. These frameworks are generally grounded in the positivist paradigm relying on quantitative approaches. Examples of these include the Protocol for Language Arts Teaching Observation (PLATO; Grossman et al., 2013), the Classroom Assessment Scoring System (CLASS; Pianta et al., 2010), the Framework for Teaching (FfT; Danielson, 2007), the International System for Teacher Observation and Feedback (ISTOF; Muijs et al., 2018), the Global Teaching Insight (GTI Observation System (OECD, 2020)), the Teach Observation System (World Bank, 2022), and the International Comparative Analysis of Learning and Teaching instrument (ICALT; van de Grift, 2014).

Whilst all of these tools differ in structure, theoretical underpinning and implementation techniques, they do share some concepts and characteristics; that is that they are based in the tradition of teaching and teacher effectiveness research and recognized as being measures of effective teaching behavior (Maulana et al., 2014, 2021; van de Grift et al., 2017; OECD, 2020; World Bank, 2022). In addition to more quantitative classroom observation tools, classroom observations based on qualitatively driven frameworks also exist. Examples of these include the Joint Action Framework in Didactics (JAD; Sensevy, 2014), the Cambridge Dialog Analysis Scheme (CDAS; Hennessy et al., 2020), and the ethnographic Documentary Method (DM; Martens and Asbrand, 2022). In the current study, we focus our investigation on the ICALT observation instrument. This framework will be elaborated further below.

2.2. International classroom observation instruments

Observation instruments refer to systematized and standardized tools consisting of a set of predetermined criteria and metric rules for measuring subject-specific and/or generic aspects of teaching skills based on certain views and frameworks of teaching quality. Existing international classroom observation instruments are typically quantitative in nature. There are at least five (high-inference) classroom observation instruments that were designed generically for measuring teaching behavior internationally. These comparative instruments include Virgilio Teacher Behavior Inventory (VITB), International System for Teacher Observation and Feedback (ISTOF), Global Teaching Insights (GTI), Teach, and International Comparative Analysis of Learning and Teaching (ICALT).

An earlier major, large-scale study examining school and teacher effectiveness in Europe, Pacific countries, US, Canada, Australia, and others (Reynolds et al., 2002), suggested that important factors identified through national effectiveness projects could also be found in other national contexts. The study used VITB consisting of three domains: Classroom Management, Instruction, and Classroom Climate (Teddlie et al., 1990). The International School Effectiveness Research Project (ISERP), however, revealed that the VITB had limited external validity when used in only nine countries, each with established school and teacher effectiveness traditions (Reynolds et al., 2002). This finding directed researchers toward further, more detailed examinations of teaching behavior which utilized an observation instrument which was validated across many national contexts. This prompted the subsequent development of ISTOF, used in a study of 20 countries (Reynolds et al., 2002). The instrument has also been adopted in single-country projects. For example, more recently Muijs et al. (2018) reported a satisfactory level of the ISTOF factor structure across a single national site, although they were unable to find ISTOF's 7-factor structure on a consistent basis. Muijs et al. (2018) also highlighted the fact that factor invariance of the ISTOF instrument across countries remains unknown, making it challenging to achieve a cross-country comparison.

Global Teaching Insight was developed by the OECD. The instrument was used to capture an overall picture of teaching quality across eight education contexts including Chile, Colombia, England (UK), Germany, Japan, Spain, Mexico, and China. GTI captures effective teaching behavior in terms of three generic domains including: Classroom Management, Social-emotional Support, and Instruction. The Instruction domain is further divided into three subdomains including Discourse, Quality of Subject Matter, and Student Cognitive Engagement (OECD, 2020). This large-scale study offers general insights into differences in mathematics classroom practices across the eight education contexts. Although construct validity and measurement invariance of the instrument is assumed, we found no specific information regarding the psychometric quality of GTI, particularly on measurement invariance, in more details. Rater quality checks, however, were reported to be conducted systematically (OECD, 2021b).

Teach is a generic classroom observation instrument developed by the World Bank. The instrument was designed to help countries collect data on teaching practices with the aim to improve teaching quality (World Bank, 2022). Teach measures teaching quality in terms of two domains: Time on Task, and Quality of Teaching Practices. The domain Quality of Teaching practices consists of three dimensions: Classroom Culture, Instruction, and Socioemotional Skills. The instrument has mainly been used in low- and middle-income countries, but it was claimed that the instrument can be contextualized for different settings (World Bank, 2022). We also found no specific information on the psychometric quality of the instrument for comparing teaching across education contexts.

International Comparative Analysis of Learning and Teaching is a generic instrument initially developed by four European inspectorates of education including the Netherlands, England,

TABLE 1 Summary of the six ICALT domains.

Domain	Description
Learning climate: safe and stimulating learning climate	Providing a positive classroom environment for learning. This can include promoting respect between all members of the class and the development of students' self-confidence (Cornelius-White, 2007; Hattie and Clinton, 2008; Smith et al., 2008; Ginner Hau et al., 2021)
Classroom management: providing efficient classroom management	The planning and organization of the lesson. This can include quality planning and preparation of lessons, effective time management and pace of lessons, smooth transitions between activities and the swift management of classroom disruption (Wang et al., 1995; Yair, 2000; Marzano, 2003; Opdenakker and Minnaert, 2011; Simonsen et al., 2020)
Clarity of instruction: clear and structured instruction	This relates to the quality of instruction given by the teacher. It can include a clear and organized structure to the lesson, clear explanations of subject content and the effective communication of group and individual classroom tasks (Kindsvatter et al., 1988; Mortimore et al., 1988; Rosenshine, 2010, 2012) alongside the regular assessment of student understanding (Hattie and Clinton, 2008; Smith et al., 2008)
Activating teaching: intensive and activating teaching	This includes behavior for learning principles (Maulana et al., 2017) which ensure the learning of students is maximized, cognitive load is minimized (Bolkan, 2016) and prior knowledge is used effectively to develop schemas for learning (Paas and van Merriënboer, 2020)
Differentiated instruction: adjusting instructions and student processing to inter-student differences	This is an inclusive aspect of teaching behavior, where the teacher involves students from a diverse range of personalities and backgrounds in the lesson. It involves adapting their teaching approaches and learning episodes to suit students' diverse range of learning needs in order to facilitate knowledge acquisition and improved student outcomes (Tomlinson et al., 2003; Reis et al., 2011; Ismajli and Imami-Morina, 2018)
Teaching learning strategies: teaching students learning strategies	This domain focuses on metacognitive learning and teaching strategies employed by the teacher. By employing these strategies and providing students with learning opportunities, it has been shown to provide students with the scaffolded support they need to succeed academically (Rosenshine and Stevens, 1986; Houtveen and van de Grift, 2007; Bae and Kwon, 2021)

Flanders, Lower Saxony (Germany) (van de Grift, 2007). The first ICALT version was developed to capture mathematics teaching quality in primary schools. The instrument was developed further for use in secondary schools across subjects (van de Grift et al., 2014; Maulana et al., 2017). The ICALT observation instrument has been found to be sufficiently invariant across several countries in both primary and secondary education contexts (van de Grift, 2014; van de Grift et al., 2017). Maulana et al. (2020) has since used the ICALT observation instrument with notable success and limitations. They discovered that observers rated Republic of Korea the highest of four countries (compared with Netherlands, South Africa, and Indonesia) in terms of effective teaching behavior. They also found that Differentiated Instruction was observed as being low in both Republic of Korea and in the Netherlands, though observers rated Republic of Korea higher than the Netherlands within this domain. While measurement invariance for two further countries involved in the study (Hong Kong SAR, China and Pakistan) was not found when retaining the full sets of ICALT items, the measurement invariance found for four of the six partaking countries provides further interest in using the ICALT observation instrument across multiple national contexts.

2.3. Effective teaching behavior: the ICALT framework

The ICALT framework is grounded in the evidence-based teacher effectiveness research (TER). Based on reviews of TER, six observable effective teaching behavior domains were synthesized (van de Grift, 2007). These six domains, discussed in depth by Maulana et al. (2021), are included in the ICALT observation tool. The six ICALT domains resonate with findings from other studies of effective teaching behavior (e.g., Klieme et al., 2009; Pianta and Hamre, 2009; Danielson, 2013; Ko and Sammons, 2013; Kington et al., 2014; Muijs et al., 2018). A summary of each of the six domains is given in **Table 1** below.

van de Grift et al. (2014) linked the theory of teachers' concerns (Fuller, 1970) with the ICALT theoretical framework. Fuller's theory posits that teachers' concerns may develop from self-related (focused on self), to task-related (focused on task), and then to impact-related (focused on impact for students) concerns. By applying the Rasch modeling approach, it has been shown that the six domains of effective teaching behaviors can be separated into two different levels of difficulty (van de Grift et al., 2014). The Rasch modeling offers a unique possibility for arranging teaching skill scores on a single dimension and for estimating individuals' skill levels on a latent variable. Hence, this modeling approach makes it possible to link the teacher's concern theory with teaching behavior domains. The first three (Learning Climate, Classroom Management, Clarity of Instruction) have been identified as lower levels of teaching behavior difficulty. This contrasts with the final three domains (Activating Teaching, Differentiated Instruction, Teaching Learning Strategies) providing greater levels of difficulty for teachers (van de Grift et al., 2014; van der Lans et al., 2018).

The ICALT observation instrument has been used to measure teaching behavior for nearly 20 years. In the two decades since its conception, the tool has been developed and used to compare teaching behavior across primary schools in Europe (van de Grift, 2007) and subsequently validated using confirmatory factor analysis (CFA) (van de Grift, 2014). The tool has also been validated for secondary education in the Netherlands (Maulana et al., 2017), Republic of Korea (van de Grift et al., 2017) and Indonesia (Irnidayanti et al., 2020), and for university education in the Netherlands (Noben et al., 2021). In a similar study to the one reported here, the ICALT observation measure was used to measure invariance of observed secondary teaching behavior across the Netherlands, Republic of Korea, South Africa, Indonesia, Hong Kong SAR, China, and Pakistan (Maulana et al., 2021). Four of the six countries showed measurement invariance, suggesting that the ICALT instrument might be an effective tool for cross-national comparisons of teaching behavior more globally.

3. The current study

3.1. Research questions

Given the gap in comparative educational research for a cross-nationally validated observation instrument to measure teaching behavior, this paper reports on results from an international study that employed the ICALT instrument three across different national contexts. The research questions which guided the analysis were:

1. To what extent is there evidence of an invariant internal structure regarding effective teaching behavior in the Netherlands, England, and the United States?
2. What are similarities and differences regarding the six domains of effective teaching behavior across the three education contexts?

3.2. Contexts of the study

3.2.1. The Netherlands

International comparisons in secondary education indicated that students attending Dutch schools perform above average, comparable to other high performing European and Asian educational systems (Martin et al., 2016a,b; Mullis et al., 2016; OECD, 2016). The majority of students in their teenage years achieve the basic skills in reading, mathematics and science (scores 485, 519, and 503, respectively) (OECD, 2019). The Dutch educational system is highly tracked (i.e., students are split by ability in a large number of different educational tracks from the age of twelve), does not apply a national curriculum, shares national educational standards and gives extensive autonomy to schools and teachers (OECD, 2014, 2016). The high level of decentralization is balanced by a strong school inspection mechanism and a national examination system. The teaching profession does not have a high status in the general public opinion (Brouwer et al., 2016). Nevertheless, the quality of teachers is generally high with the large majority mastering the basic teaching skills to a good standard (OECD, 2016; Inspectie van het Onderwijs, 2018). For secondary school practitioners, there are two types of teaching qualification; first degree qualification (highest degree, license for teaching in all grades), and second degree qualification (license for teaching in lower grades) (Brouwer et al., 2016). First degree qualified teachers generally show higher levels of effective teaching behavior compared to second degree teachers (Helms-Lorenz et al., 2020). It is also possible to teach in secondary education without teaching qualification as a lateral second-career entrant (in Dutch: zij-instromer) or a guest teacher (Rijksoverheid, 2023) with the requirement to qualify within 2 years.

3.2.2. England

In England, secondary school education (pupils aged 11–18 years) takes place in either a state-funded or independent school, with the majority of pupils (93.6%) being educated in state schools (Gov.uk, 2021).[1] Schools maintained by a local authority follow the national curriculum which is set by the Department for Education (DfE). Within the state school system, over half (57.7%) have been designated with "free school" or "academy school" status (see text footnote 1 Gov.uk, 2021) and have more flexibility to deviate from the National Curriculum. Examination boards must set their learning specifications against the DfE's subject content, which means that students must cover the same subject content for the General Certificates of Secondary Education (GCSE) qualifications. Furthermore, the Office for Standards in Education (Ofsted) carry out a rigorous inspection program of all state-funded schools, ensuring the curriculum is broad and that teachers have "good knowledge of the subject" (Ofsted, 2021, p. 40) they teach. Academic outcomes of pupils in the UK have seen an improvement, with reading ability (score 504) rising from 25th to 14th place amongst OECD countries (OECD, 2021a), and scores for pupils in England being the highest of the UK nations (math and science scores 502 and 505, respectively) (OECD, 2019). In mathematics and science, the UK is above the average for OECD countries, and is showing a continuing upward trend (OECD, 2021a). All teachers in England need to have an undergraduate degree to teach. Although it is possible to be employed as an unqualified teacher, this is not common practice as most teachers must undertake formal training and be awarded Qualified Teacher Status (QTS). There is strict performativity and accountability agenda in schools, which is overseen by the Ofsted inspectorate. This is paired with an increasing rate of attrition; for example, in 1996, 9% of teachers left the profession after 1 year. This has increased to 16% in 2019 (Gov.uk, 2021).

3.2.3. The United States

Education is more decentralized in the United States than in most European countries. State-funded public schools represent a high percentage, about 87%, of K-12 education, while tuition and foundation-funded private schools account for about 10%, and home-schooling is roughly 3% (U.S. Department of Education, 2013). Federal spending on education is relatively modest at about one-sixth of the state spending. Therefore, it is not surprising that each state sets its own State Compulsory School Attendance Laws (U.S. Department of Education, 2013). Compulsory education starts between five and eight and ends between sixteen and eighteen. Each state can set its curriculum and staffing policies, except private schools are free to determine theirs, which can also obtain accreditation through independent regional accreditation authorities. While the U. S. spent more per student on education than any other country, 15-year-old American students ranked the 31st in the world in reading literacy, mathematics, and science in The Program for International Student Assessment (OECD, 2018). The average American students scored 487.70, compared with the OECD average of 493 in overall knowledge and skills. Teacher quality is found to significantly improve student achievement (Rockoff, 2004; Hanushek and Rivkin, 2010; Chetty et al., 2014;

[1] www.Gov.uk

TABLE 2 Demographic information of participating teachers.

Countries	N_{school}	School denom		N_{teache}	Teacher gender		Teacher subject		Teacher experience		Class size
		Public	Private		Male	Female	Science[a]	Non-science[b]	Inexperienced	Experienced	
Netherlands	99	99	0	200	96 (48.0%)	104 (52.0%)	74 (37.0%)	126 (63.0%)	100 (50.0%)	100 (50.0%)	22.09 (5.30)
England	15	15	0	115	50 (43.5%)	65 (56.5%)	61 (53.0%)	54 (47.0%)	6 (5.2%)	109 (94.8%)	23.18 (3.05)
United States	64	*NA*	*NA*	105	*NA*	*NA*	50 (47.6%)	55 (52.4%)	*NA*	*NA*	*NA*
Total	178			420							

[a]Natural science subjects normally go to the category of *beta* in the Dutch context.
[b]Non-science subjects can be further categorized into alpha subjects (i.e., humanities) and gamma subjects (social sciences) in the Dutch context.

Bruns and Luque, 2015). While some states have improved their teacher preparation programs (e.g., Commission of Teacher Credentialing of California, 2021), an earlier study by Harris and Sass (2011) failed to establish a consistent relationship between formal professional development training and teacher productivity. Although teachers reported being largely satisfied with their jobs and career, only a minority believed that teaching is valued by U.S society (OECD, 2013). The current state of the teaching profession in the U.S. is near its lowest levels in over a half century (Kraft and Lyon, 2022).

4. Materials and methods

The current study is part of a larger longitudinal project on effective teaching behavior across countries called ICALT3/Differentiation.[2] In this project, effective teaching behavior was studied from researcher observations, teacher perceptions, and student perceptions. The current study focuses on reporting the observation data of one particular measurement moment (cross-sectional) from three educational contexts: the Netherlands, England and the US. Particularly, data collected during the second year of fieldwork was used in all three countries.

4.1. Sample and procedure

The current study included 420 teachers from 178 secondary schools in the Netherlands (N_{teacher} = 200), England (N_{teacher} = 115), and the United States (N_{teacher} = 105). In the Netherlands, the cross-sectional data were collected from different cohorts between 2015 and 2018 school years. A random sample of 200 observed teachers from a total of 2157 teachers was selected to compensate for sample imbalance with the other two datasets. The observed teachers taught in schools across 12 provinces in the country. In England, the data were also collected from different cohorts between 2017 and 2019 school years. The observed teachers were from schools across the West Midlands region. Hence, the data is not a nationally representative sample as the other eight regions (including London) were not included. In the US, the video data were collected in 2011 as part of the Measurement of Teaching (MET) project.[3] A random selection of 105 teachers from 12 US regions was done from a pool of available video data.

Table 2 illustrates the denomination of schools, the gender, subject areas and teaching experience of teachers, as well as the average size of their classes. The dividing line between inexperienced and experienced teachers is set at 5 years of teaching experience. The background information of schools and teachers, except teachers' subject expertise, was missing in the US dataset. All participating schools in the Netherlands and England were publically funded schools. The distribution of teachers is relatively balanced across gender, subjects and experience levels, except for teacher subject in the Netherlands and teacher experience in the US. Specifically, non-science teachers are overrepresented in the Dutch

[2] https://www.rug.nl/gmw/lerarenopleiding/onderzoek/psychometrisch/

[3] https://usprogram.gatesfoundation.org/news-and-insights/usp-resource-center/resources/met-project-data

data, while experienced teachers are overrepresented in English data.

In the Netherlands and England, teachers were invited to participate in the study in accordance with each institution's approved ethical procedures. This included sending information to teachers and having them sign a consent form that served as an agreement between the researchers, the teacher and the teacher's school. Consent was given voluntarily and with full knowledge. In the Netherlands and England, lessons were observed in real-time in their natural classroom settings.

For the US, we cooperated with the MET project coordinator and researcher. The MET project collected video data on classroom teaching. The trained observers conducted a secondary analysis and coded the selected lesson videos/teachers of the 10000+ lessons originally rated with CLASS (Pianta et al., 2010).

4.2. Measures

Effective teaching behaviors were observed using the International Comparative Analysis of Learning and Teaching (ICALT) observation instrument (van de Grift et al., 2014). The instrument includes 120 low inferential observable teaching indicators and 32 high inferential observable teaching behaviors. These high inference items were rated on a four-point scale ranging from 1 ("mostly weak") to 4 ("mostly strong,") and represent the six aforementioned domains of teaching behavior: Learning Climate (four items), Classroom Management (four items), Clarity of Instruction (seven items), Activating Teaching (seven items), Differentiated Instruction (four items), and Teaching Learning Strategies (six items). The six-factor structure of observed teaching behavior has been validated by prior research in other national contexts (Maulana et al., 2020).

The ICALT instrument was used for the current study for several reasons (see Maulana et al., 2021). In summary, this observation tool is, relatively speaking, simple to use in the classroom context, has been translated into many languages, and has already been validated through its use in a number of previous international studies (Maulana et al., 2021, 2022). The instrument was developed with a strong grounding in evidence-based teacher effectiveness research, and its validity has been demonstrated in both primary and secondary education settings, adding to its external validity (van de Grift, 2014; Maulana et al., 2017, 2021). The ICALT tool has also been shown to be appropriate for use in both comparative educational cultures—for example, in several countries in Europe (see van de Grift et al., 2014)–and in contrasting cultures, such as Indonesia, Republic of Korea and South Africa, amongst others (Maulana et al., 2021). Furthermore, the instrument has been established as valuable tool for use in both research and practice in the Netherlands and in Republic of Korea (Maulana et al., 2020). It is frequently used as a diagnostic measure for the professional development of teachers and pre-service teachers (Maulana et al., 2017; Helms-Lorenz et al., 2019), highlighting its versatility and the relative ease with which observers can be trained.

4.3. Observer training

Before using the ICALT observation instrument, each observer underwent onsite training led by two expert trainers. The training was conducted in accordance with the same standards, structure and procedure in all three countries. In order to be qualified for training, observers had to meet two criteria. First, they needed to demonstrate sufficient knowledge of effective teaching. Second, the observers needed at least 3 years of teaching experience, preferably in secondary education. Compared to the Netherlands and England, observers in the US were less experienced teachers but highly knowledgeable with regard to teaching behavior research. In England, all observations were conducted by the research team; no observations were carried out by teachers.

Training consisted of three phases: planning, execution, and evaluation. Phase 1 was dedicated to exploration of the theoretical basis and context of the instrument to deepen trainees' understanding of the theory underlying the ICALT instrument. Phase 2 was devoted to administration of the instrument, including how to rate indicators of teaching behavior using the applied scoring rules. More specifically, trainees coded the teacher behavior in two videotaped lessons using the observation instrument. When analyzing the observation scores of trainees, a consensus

TABLE 3 Categorical confirmatory factor analysis for three countries.

	Model	*N*	χ^2 (df)	RMSEA with 90% CI	SRMR	CFI	TLI
Netherlands	All items	200	911.259* (449)	0.072 [0.065, 0.078]	0.085	0.925	0.917
England	All items	115	588.954* (449)	0.052 [0.040, 0.063]	0.098	0.962	0.958
	11 correlations set to 1	115	599.547* (460)	0.051 [0.039, 0.063]	0.098	0.962	0.959
United States	All items	105	1303.847* (449)	*0.135 [0.126, 0.143]*	*0.232*	*0.594*	*0.552*
	Items 1, 14, 18 removed	105	821.182* (362)	*0.110 [0.100, 0.120]*	*0.188*	*0.705*	*0.669*
	Items 1. 14. 18, 20, 25, 32 removed, CLAR with ORG set to 1	105	569.205* (284)	*0.098 [0.086, 0.109]*	*0.171*	*0.778*	*0.746*
	Items 1, 14, 18, 20, 25, 32 removed, v17 with v27, v24 with v29, CLAR with ORG set to 1	105	507.825* (283)	*0.087 [0.075, 0.099]*	*0.157*	*0.825*	*0.799*
	Items 1, 14, 16, 18, 19, 20, 24, 25, 32 removed	105	412.907* (215)	*0.094 [0.080, 0.107]*	*0.158*	*0.801*	*0.766*

Unacceptable fit indices are in *italics*. *$p < 0.05$.

TABLE 4 Standardized factor loadings of separate CFA for the Netherlands and England.

Country and domains	Standardized factor loadings							Domain correlations					Variance explained
								1	2	3	4	5	
Netherlands (N_{total} = 200)													
1. Learning climate (four items)	0.707	0.816	0.759	0.799									59.5%
2. Classroom management (four items)	0.774	0.843	0.683	0.756				0.814					58.7%
3. Clarity of instruction (seven items)	0.649	0.745	0.657	0.802	0.732	0.793	0.699	0.850	0.887				52.9%
4. Activating teaching (seven items)	0.694	0.688	0.810	0.804	0.769	0.727	0.542	0.797	0.705	0.874			52.4%
5. Differentiated instruction (four items)	0.736	0.862	0.847	0.799				0.371	0.399	0.426	0.610		66.0%
6. Teaching learning strategies (six items)	0.805	0.853	0.752	0.817	0.734	0.847		0.444	0.368	0.467	0.735	0.611	64.4%
England (N_{total} = 115)													
1. Learning climate (four items)	0.622	0.775	0.829	0.594									50.7%
2. Classroom management (four items)	0.694	0.511	0.751	0.637				*1.000*					42.8%
3. Clarity of instruction (seven items)	0.649	0.590	0.692	0.648	0.651	0.723	0.648	*1.000*	*1.000*				43.4%
4. Activating teaching (seven items)	0.704	0.661	0.560	0.612	0.812	0.831	0.621	0.953	*1.000*	*1.000*			47.9%
5. Differentiated instruction (four items)	0.600	0.793	0.590	0.630				0.959	*1.000*	0.980	*1.000*		43.3%
6. Teaching learning strategies (six items)	0.651	0.768	0.557	0.769	0.637	0.525		*1.000*	*1.000*	*1.000*	*1.000*	0.977	43.3%

All factor loadings and factor correlations reported in the table are significant (*p* <0.001) except the ones in *italics*.

level of 70% within the group and between the group and the expert norm was set as a sufficient cut-off. Discussions to resolve significant differences and improve the consensus level were conducted subsequently. Finally, the evaluation phase involved the investigation of rating patterns and significant deviations from the average pattern. A small number of observers who deviated from the average were followed up and extra guidance was given to this group prior to conducting the observation in natural classroom settings. Observers failing to meet the minimum consensus of 70% were not invited to conduct observations. The consensus levels between the trainees and the expert norm are 86% in the Netherlands, 71% in England, and 75% in the US, respectively.

4.4. Analytic approach

The measurement model of the ICALT instrument was first subjected to cross-country validation using categorical confirmatory factor analysis (CFA). After confirming the validity of the measurement model in each country data, this study tested the measurement invariance by performing multi-group confirmatory factor analysis (MGCFA) on the combined country data using MPlus version 8 (Muthén and Muthén, 2017). We followed the works of Millsap and Yun-Tein (2004) and Xing and Hall (2015) as references when estimating the models. Three levels of measurement invariance (configural, metric, and scalar) were tested successively. The test of configural equivalence examined whether the same factor structure was applied across countries (i.e., same factorial structure). Metric invariance test verified whether teaching behavior from different countries was rated by the items in identical ways (i.e., same factor loadings on measured items). Scalar invariance examined whether teachers with the same score of a latent construct (i.e., 6 behavioral domains) would be rated with the same observed scores (i.e., 32 high-inference items) when measured, irrespective of their country of origin (i.e., equal intercepts on measured item). Scalar invariance permits trustworthy comparisons of factor means across groups (Byrne, 2012).

Root-mean-square error of approximation (RMSEA), comparative fit index (CFI) and Tucker-Lewis index (TLI) are frequently used goodness of fit indices for categorical CFA and MGCFA models, and they all adhered to the same guidelines for a good model fit (i.e., CFI ≥ 0.90, TLI ≥ 0.90, SRMR ≤ 0.10, and RMSEA ≤ 0.08 are considered acceptable) (Hu and Bentler, 1999; Vandenberg and Lance, 2000). Additionally, the deterioration of model fit between successively constraint invariance models was examined by referring to the changes in CFI (ΔCFI), RMSEA (ΔRMSEA), and SRMR (ΔSRMR), with changes above 0.01

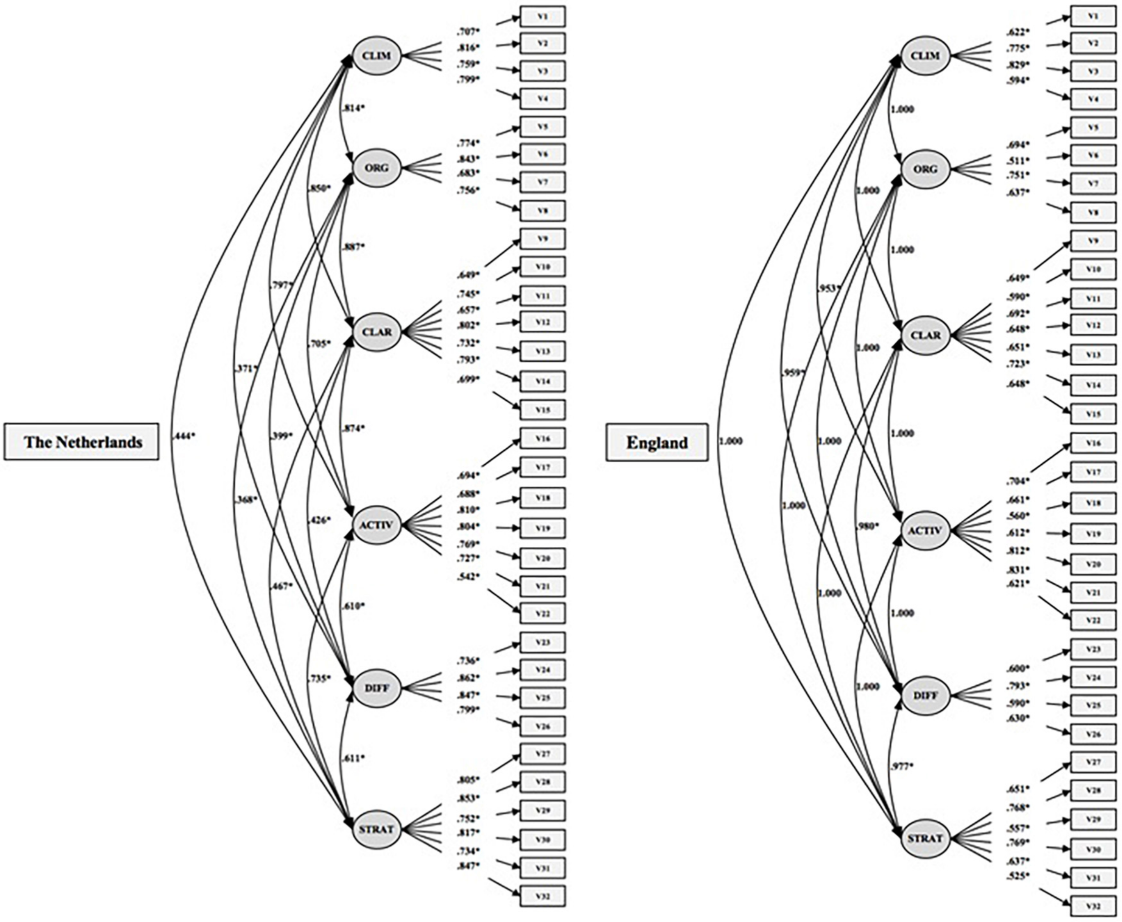

FIGURE 1
Factorial structures of effective teaching behavior for the Netherlands and England.

TABLE 5 Categorical multi-group confirmatory factor analysis for the Netherlands and England.

	Chi-square (df)	CFI	RMSEA with 90% CI	SRMR	Model comp	ΔCFI	ΔRMSEA	ΔSRMR	Decision
M1: Configural invariance	1918.894* (936)	0.887	0.082 [0.076, 0.087]	0.104					
M2: Metric invariance	1865.558*(968)	0.897	0.077 [0.071, 0.082]	0.111	M1	0.010	−0.005	0.007	Accepted
M3: Scalar invariance	2206.200* (994)	0.861	0.088 [0.083, 0.093]	0.115	M2	*−0.036*	*0.011*	0.004	Rejected
M3a: Partial scalar invariance[a]	1965.743* (986)	0.888	0.079 [0.074, 0.085]	0.111	M2	−0.009	0.002	0.000	Accepted
M4: Comparing latent means	1601.528* (942)	0.924	0.067 [0.061, 0.072]	0.097					

[a]Set free the thresholds of Item 2 in learning climate, Item 26 in differentiated Instruction, and Item 29 and 32 in teaching learning strategies. *$p < 0.05$.

indicating significant differences (Cheung and Rensvold, 2002; Chen, 2007).

5. Results

5.1. Preliminary analysis

Before investigating measurement invariance, the factor structure of effective teaching behavior in the three education contexts was investigated. If the hypothesized factor structure was confirmed using the empirical data, the investigation of measurement invariance was possible. The full measurement model of CFA displayed acceptable fit levels in the Netherlands and England, but not in the US (see **Table 3**). In the English model, 11 correlations between the latent constructs (i.e., 6 behavioral domains) were found above 1.00. After these correlation coefficients were adjusted to 1.00, the England model consistently showed acceptable fit indices. Comparing the adjusted model with the prior model revealed no significant changes in model fit. For the US data, the desired acceptable full measurement model could not be reached. One of the major issues was the cross-loadings of Item 1 ("The teacher shows respect for learners in his/her behavior and language"), Item 18 ("The teacher stimulates learners to think about solutions") and Item 4 ("The teacher fosters mutual respect"). In contrast to the positive domain correlations in the other two countries' models, the correlations between Learning Climate and Differentiated Instruction ($r = -0.355$, $p < 0.01$) and Teaching Learning Strategies ($r = -0.655$, $p < 0.001$) were moderately negative. Minor (e.g., error-term covariance) to major modifications (removal of 28% of the items) to the US data, which were made based on the modification indices, progressively improved the model-data fit, but not to the acceptable level. Therefore, the subsequent MGCFA only includes the full measurement models of the Netherlands and England.

All items sufficiently loaded on their corresponding domains, as shown by the factor loadings of all behavioral domains in the Netherlands and England being above the standard criterion of 0.40 (see **Table 4** and **Figure 1**). Comparisons of the loadings across nations are not advised at this time since we have not yet established cross-contexts invariance.

5.2. Measurement invariance of teaching behavior

Although the six-factor structure was validated in each of the two education contexts separately, the results of categorical MGCFA using the two country data showed no convergence. A careful examination of the data detected instances of no celling filling in various item response categories of both countries (Netherlands: Item 1 with categories 1 and 2 unfilled, Item 2 and 3 with the category 1 unfilled; England: 20 items with the category 1 unfilled). Hence, we decided to re-categorize the value from 4 to 3 categories by collapsing the category 1 to 2.

Model fit indicators of MGCFA on the collapsed merged data are reported in **Table 5**. The configural model falls slightly under the minimum required standard. When the loadings were constraint to be identical (i.e., metric invariance), the change in fit was fairly minimal (ΔCFI = 0.010, ΔRMSEA = −0.005, and ΔSRMR = 0.007), showing an acceptable model-data fit relative to the configural model. For the scalar model, both CFI and RMSEA suggested a significant fit reduction (ΔCFI = −0.036, ΔRMSEA = 0.011, and ΔSRMR = 0.004). To obtain a better model fit (ΔCFI = −0.009, ΔRMSEA = 0.002, and ΔSRMR = 0.000), eight intercepts were released. This indicates that partial scalar invariance was supported for the two country's data. Partial scalar invariance is a sufficient criterion for comparing mean scores. No further modification or additional invariance models were included. A total of 28 items (87.5%) had invariant intercepts. The parameter estimates for this model are shown in **Table 6**.

5.3. Teaching behavior across the three education contexts

The comparison of latent means between the Netherlands and England was supported after achieving partial scalar invariance. Results revealed that teachers in England displayed lower levels of relatively basic behavioral domains (Stimulating Learning Climate and Classroom Management), but performed better on all relatively advanced domains (Activating Teaching, Differentiated Instruction, and Teaching Learning Strategies) ($p < 0.001$, see

TABLE 6 Standardized factor loadings for two countries in the partial scalar model.

Country and subscales	Standardized factor loadings							Domain correlations					Variance explained
								1	2	3	4	5	
Netherlands													
1. Stimulating teaching (four items)	0.687	0.802	0.779	0.728									56.3%
2. Classroom management (four items)	0.764	0.741	0.733	0.716				0.884					54.6%
3. Clarity of instruction (seven items)	0.661	0.703	0.688	0.758	0.740	0.771	0.695	0.905	0.958				51.5%
4. Activating teaching (seven items)	0.738	0.729	0.790	0.758	0.762	0.758	0.595	0.826	0.735	0.892			54.1%
5. Differentiated teaching (four items)	0.768	0.857	0.785	0.761				0.443	0.510	0.494	0.645		63.0%
6. Teaching learning strategies (six items)	0.797	0.857	0.715	0.806	0.740	0.748		0.563	0.447	0.551	0.815	0.596	60.6%
England													
1. Stimulating teaching (four items)	0.687	0.802	0.779	0.728									56.3%
2. Classroom management (four items)	0.764	0.741	0.733	0.716				0.890					54.6%
3. Clarity of instruction (seven items)	0.661	0.703	0.688	0.758	0.740	0.771	0.695	0.879	0.813				51.5%
4. Activating teaching (seven items)	0.738	0.729	0.790	0.758	0.762	0.758	0.595	0.887	0.865	0.925			54.1%
5. Differentiated teaching (four items)	0.768	0.857	0.785	0.761				0.740	0.697	0.762	0.831		63.0%
6. Teaching learning strategies (six items)	0.797	0.857	0.715	0.806	0.740	0.748		0.797	0.804	0.767	0.789	0.665	60.6%

All factor loadings and factor correlations reported in the table are significant ($p < 0.001$).

Table 7 and Figure 2) compared to teachers in the Netherlands. There was no significant difference between the two education contexts reported for Clarity of Instruction. Among these behavioral domains, Differentiated Instruction showed the largest cross-context variation.

TABLE 7 Latent means of the partially scalar equivalent MGCFA model for the Netherlands and England.

	Netherlands (N =200)	England (N =115)
Learning climate	0.000	−0.603**
Classroom management	0.000	−0.742**
Clarity of instruction	0.000	−0.214
Activating teaching	0.000	0.759**
Differentiated instruction	0.000	2.559**
Teaching learning strategies	0.000	1.484**

**p <0.001. Netherlands is the reference category.

Examination of the mean scores revealed similar patterns with the MGCFA results, showing that teachers in the Netherlands showed higher levels of Learning Climate and Classroom Management, but lower levels of Activating Teaching, Differentiated Instruction and Teaching Learning Strategies compared to teachers in England (see Figure 3). Most notably, the largest difference was observed for Differentiated Instruction. Although it is tempting to say something about observed effective teaching behavior of teachers in the US, which based on mean scores seem to be generally lower compared that of teachers in the Netherlands and England, we have refrained from doing so due to the construct validity issue with the current US data and no evidence of measurement invariance.

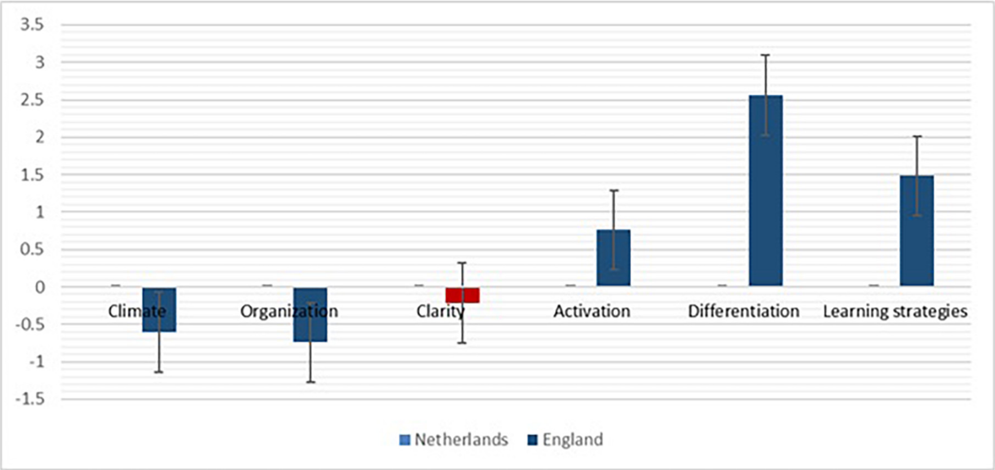

FIGURE 2
Latent means on the six domains between the Netherlands and England. The difference in clarity instruction in not significant. The Netherlands is reference category.

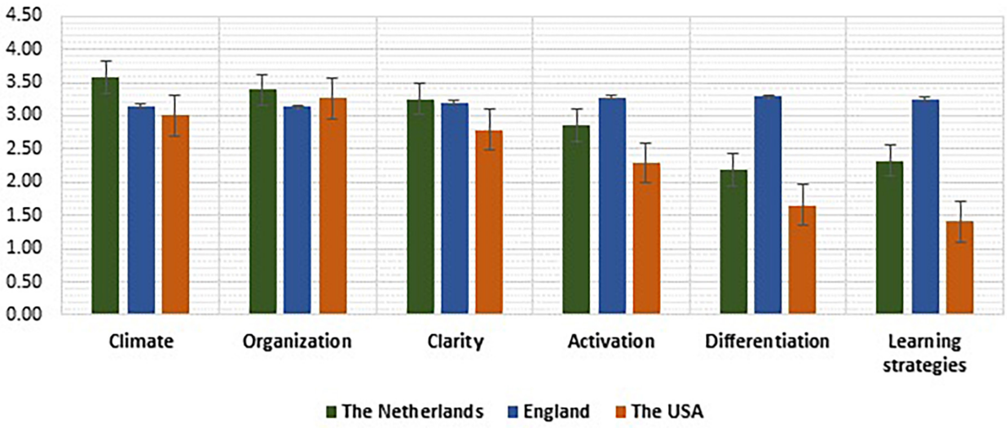

FIGURE 3
Raw scores means of the six domains. The six-factor structure is not supported in the USA data. Hence, the US result are not compared subsequently. Display is for general indication only.

6. Conclusion and discussion

The current study aimed to investigate measurement invariance of observer scoring of effective teaching behavior of secondary school teachers in three education contexts, including the Netherlands, England and the US, using a uniform ICALT observation instrument. It also aimed to identify differences in effective teaching behavior in the three education contexts. The work extends the previous work using a similar observation instrument (i.e., van de Grift et al., 2014; Maulana et al., 2021) by studying effective teaching behavior in the Netherlands, England, and the US.

The first guiding question was: *To what extent is there evidence of an invariant internal structure regarding effective teaching behavior in the Netherlands, England, and the United States?* We found that the six-factor structure of effective teaching behavior, as measured by the ICALT observation instrument, was confirmed in the Netherlands and England, but not in the US. Subsequently, the US data could not be included in the investigation of measurement invariance for comparing effective teaching behavior with the Netherlands and England. Prior to this study, face validity of the ICALT observation instrument was conducted including the experts from the three education contexts which confirmed the relevance of the six-factor structure of effective teaching behavior in their contexts. The current study provides further empirical evidence supporting the relevance of effective teaching behaviors and the usefulness of the ICALT instrument for measuring teaching behaviors in secondary education in the Netherlands and England.

Despite heavy modifications, the six-factor structure of teaching behavior was not confirmed in the current US data. Poor model-data fit indicates the data set did not match with the hypothesized model either because of the poor item quality or because of the poor understanding of the observers regarding the ICALT items. Although reasons for the poor model-data fit in the US data were unclear, further examinations of potential sources of bias may shed some light, including observation objects (real classroom vs. videoed classroom) and observer characteristics. In the Netherlands and England, teachers were observed and rated by trained observers in their classroom directly and in real-time, while in the US teachers were videotaped first and then rated by trained observers at a different time. The difference in the way teachers were observed (real-time actual classroom setting vs. videotapes) may explain differences in scoring teaching behavior. Furthermore, observers who observed and rated teachers in Netherlands and England were highly experienced (mostly senior) teachers or researchers, while observers who observed and rated teachers in the US, although highly knowledgeable in teaching behavior research, were relatively inexperienced teachers. In addition, the observers scoring teachers in the US are not from American backgrounds, while the observers scoring teachers in the Netherlands and England are from their original background. There is an indication that observers with different background interpret and answer the same items differently (Jansen et al., 2013). These differences may pose bias and affect the quality of the data (van de Vijver and Tanzer, 2004), and should be investigated in further research.

Measurement invariance of effective teaching behavior was established for the Netherlands and England, although full scalar invariance was not obtained. However, partial scalar invariance was a sufficient condition for comparing effective teaching behavior between groups (Steenkamp and Baumgartner, 1998). Valid inferences about the differences between latent factor means in the model can be made as long as there are at least two loadings and intercepts that are constrained equal across groups (Byrne et al., 1989). Based on the obtained partial scalar invariance, the latent factor means of effective teaching behavior in the Netherlands and England can be compared. However, comparing the sum scores or comparable observed means is not deemed acceptable because for this purpose full scalar invariance must be established (Steinmetz, 2013).

The second question was: *What are the differences regarding the six domains of effective teaching behavior across the three education contexts?* Due to the construct validity issue with the US data, and the fact that only partial scalar invariance was achieved for the Netherlands and England, this question can only be answered partially. Based on the latent factor means of scalar invariance, we found that teachers in England displayed lower levels of Learning Climate and Classroom Management behaviors compared to teachers in the Netherlands. On the contrary, English teachers showed higher levels of Activating Teaching, Differentiated Instruction and Teaching Learning Strategies compared to Dutch teachers. Learning Climate and Classroom Management are distinguished as relatively basic teaching skills, while Activating Teaching, Differentiated Instruction and Teaching Learning Strategies as relatively complex (van de Grift et al., 2014; Maulana et al., 2015). This finding is in line with findings of previous studies indicating that the majority of teachers in the Netherlands showed a good mastery of basic teaching skills (OECD, 2016; Inspectie van het Onderwijs, 2019). However, most teachers in the Netherlands still struggle in mastering complex teaching skills, particularly skills related to Differentiated Instruction (Maulana et al., 2020). Our finding seems to be in line with these past studies.

The 2018 TALIS report revealed that teachers in England and the Netherlands widely applied Classroom Management and Clarity of Instruction in their daily instructional practices, with the majority of teachers frequently dealing with disruptive students (OECD, 2019). In both education systems, complex teaching behaviors such as Teaching Learning Strategies were reported as less widespread compared to Classroom Management (OECD, 2019). Based on these TALIS results, there is evidence that teachers in both countries seem to commonly deal with basic teaching behavior in their daily practices, while exercising more complex behaviors seem to be less common. Our findings complement the TALIS findings by showing that although the two education systems share similar trends in teaching behavior practices as reported by teachers in the TALIS study, the current study shows that differences are evident between the two education contexts as reported by observers.

Most notably, the levels of Differentiated Instruction practices were observed to differ the most between the Netherlands and England. This can be related to differences in teacher preparation related to this teaching behavior between the two education contexts. In England, most teachers were trained to teach in mixed-ability settings as part of their formal initial teacher education (OECD, 2019). In the Netherlands, on the other hand, Differentiated Instruction is not widely included as an important part of the initial teacher education curriculum yet, although much discussed in the current education agenda. Previous studies have showed that the quality of complex teaching behaviors, such as Differentiated Instruction and Teaching Learning Strategies were also observed to be low in other education contexts such as Republic of Korea, Indonesia, and South Africa (Maulana et al., 2021). Particularly, Differentiated Instruction was observed to be the lowest (Maulana et al., 2021). This indicates that Differentiated Instruction is not common practice in many countries, and remains a complex skill to master by many teachers (van der Lans et al., 2018).

Differences in effective teaching behavior between England and the Netherlands may also be related to teacher preparation characteristics. As described in the context of the study section, teachers in England must undertake formal training and be awarded Qualified Teacher Status (QTS). There is a strict performativity and accountability agenda in schools, which is overseen by the Ofsted inspectorate. In the Netherlands, there are two teaching qualifications: first degree (academic-focused) and second degree (practice-oriented) qualification. In addition, it is possible to teach in secondary education without teaching qualification as a second-career entrant or a guest teacher (Rijksoverheid, 2023), with the requirement to qualify within 2 years. These background differences between the two education contexts may explain why the Dutch sample shows higher levels of basic teaching behavior, while the English sample shows higher levels of more advance teaching behavior. Effective teaching behavior plays a central role in student learning and outcomes (e.g., Chapman et al., 2012; Hattie, 2012), and teachers' work matters the most for student outcomes (Hattie, 2012; van de Grift et al., 2017). Our findings seem to be in line with this general trend. Higher performances of English teachers in higher levels of teaching behaviors may correspond to higher performances of their students, particularly in reading and science, compared to Dutch teachers and their students (OECD, 2019).

7. Implications

Our study provides evidence that comparing effective teaching behavior using the ICALT observation instrument across different education systems is promising. However, establishing the factor structure of effective teaching behavior across all contexts may remain a first challenge in this endeavor. In our case, the factor structure was established quite well in the two education contexts (the Netherlands and England), but not in the third context (the US). We speculated that differences in rating objects (real classroom vs. video-taped classroom) and rater characteristics may play a role in explaining the failure to confirm the hypothesized factor structure in the US. This suggests that effective teaching behavior may be interpreted more similarly in certain contexts but not in other contexts, which implies that establishing the hypothesized factor structure in certain contexts will require more time and effort to potentially modify poor functioning items due to some cultural and practical differences.

Although the hypothesized factor structure is confirmed in the Netherlands and England data, reaching a full invariance of the measure in the two education contexts was a more difficult challenge. Nevertheless, reaching a partial scalar invariance for a rather complex measure of effective teaching behavior such as the ICALT is quite an accomplishment. Often, other measures of similar constructs hardly fulfill the requirement of scalar invariance (e.g., Muijs et al., 2018; OECD, 2019). This implies that the ICALT measures can be used to compare effective teaching behavior in the Netherlands and England, as long as the latent factor means are used for comparison (Steinmetz, 2013).

Based on similarities and differences in effective teaching behavior found between teachers in the Netherlands and England, implications for research and practices can be drawn. Mutual and reciprocal learning exchanges between the two education contexts are advocated. Teachers in the Netherlands can potentially learn from teachers in England regarding strategies and approaches for mastering higher levels of complex skills, particularly Activating Teaching, Differentiated Instruction and Teaching Learning Strategies. Teachers' Differentiated Instruction practices was particularly observed to differ the most between the two education contexts. This teaching skill has been particularly recommended in contemporary classroom practices, yet quite complex to master for most teachers in many countries (Maulana et al., 2023). Similarly, teachers in the English context can potentially learn from teachers in the Netherlands regarding strategies and approaches to improve basic teaching behavior skills related to Learning Climate and Classroom Management.

The TALIS 2018 study revealed that both teachers in the Netherlands and the UK reported a similar pattern and a comparable degree regarding a high prevalence of practicing more basic teaching behavior such as Classroom Management, but low widespread of practicing more complex behavior such as Teaching learning Strategies (OECD, 2019). In contrast, the current study revealed differences regarding basic and complex teaching behavior between the two education contexts. This implies that modes of collecting information (teacher report vs. observation) about teaching behavior may deliver divergent results, which confirm the necessity for doing triangulation (e.g., questionnaires, observation) to study effective teaching behavior. Observation offers value-added for unraveling differences in actual teaching behavior, which can be contrasted with questionnaire surveys for examining subjective perceptions of participants (Maulana and Helms-Lorenz, 2016).

8. Limitations and future directions

This study is subject to several limitations. First, although a random sampling method was initially planned, this was not realistic to employ. Subsequently, a more convenient sampling method was applied. Second, the sample size per country is relatively small. Given the costly and laborious nature of classroom observations, however, it was not feasible to observe more teachers. Third, observation was only done once for each measurement moment, which creates a snapshot of teaching behavior. Due to these limitations, generalizations of findings at country level is not recommended until more representative and more randomly selected samples are available.

Fourth, although the observer training in the three education contexts was applied using identical procedures and standards, and we made sure that only observers who passed the extensive training were invited to observe classrooms, we have no control over the observation quality of the observers in practice. Cultural influences and day-to-day context-specificity may affect the quality of observation. Future studies should develop a control mechanism during the actual observation to minimize bias and improve data quality, if possible. It may also be worthwhile to employ cross-observer designs across education contexts to minimize cultural bias in observation (Maulana et al., 2021).

Finally, it was not possible to observe teachers in their natural settings directly in the US due to some resource problems. Hence, available video-taped lessons were used as an alternative strategy. It may be possible that observers rate teachers differently when observing actual lessons compared to video-taped lessons using the ICALT instrument. This speculation should be investigated further to establish whether raters and/or the ICALT instrument is sensitive to differences regarding the object of observation. In addition, it is coincident that the US data was rated by less experienced teachers. Failing to confirm the hypothesized factor structure of effective teaching behavior in this education context may partially be explained by this observer characteristics, which should be validated in future research.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Commission of Teacher Education University of Groningen, Worcester, and the Education University of Hong Kong. The patients/participants provided their written informed consent to participate in this study.

Author contributions

RM conceived and designed the study, wrote the manuscript, checked statistical analyses, and coordinated the manuscript. MH-L co-designed the study and contributed to the manuscript. AK, JK, BL, KH-M, and KB wrote sections of the manuscript. XF performed statistical analyses and wrote sections of the manuscript. All authors read and approved the submission of the manuscript.

Funding

This work was supported by the Dutch scientific funding agency (NRO) under Grant number 405-15-732.

Acknowledgments

We are indebted to Peter Moorer for his assistance with the data management, and to Jaap Buitink, Geke Schuurman, and Carla Griep for their valuable contributions as international trainers. The contribution of Steve Cantrell in assisting with the MET data access is highly acknowledged. We also thank Wang Xiuping for her contribution with the manuscript. The biggest thanks goes to all teachers, schools, and observers participating in this large-scale study in the three countries.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor AL declared a past collaboration with the author AK.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adamson, B. (2012). International comparative studies in teaching and teacher education. *Teach. Teach. Educ.* 28, 641–648. doi: 10.1016/j.tate.2012.02.003

Bae, H., and Kwon, K. (2021). Developing metacognitive skills through class activities: what makes students use metacognitive skills? *Educ. Stud.* 47, 456–471. doi: 10.1080/03055698.2019.1707068

Bell, C. A., Dobbelaer, M. J., Klette, K., and Visscher, A. (2019). Qualities of classroom observation systems. *Sch. Eff. Sch. Improv.* 30, 3–29. doi: 10.1080/09243453.2018.1539014

Bolkan, S. (2016). The Importance of Instructor Clarity and Its Effect on Student Learning: Facilitating Elaboration by Reducing Cognitive Load. *Commun. Rep.* 29, 152–162. doi: 10.1080/08934215.2015.1067708

Bray, M., Adamson, B., and Mason, M. (2014). "Different models, different emphases, different insights," in *Comparative education research*, eds M. Bray, B. Adamson, and M. Mason (Cham: Springer), 417–436. doi: 10.1007/978-3-319-05594-7_15

Brouwer, J., Jansen, E., Flache, A., and Hofman, A. (2016). The impact of social capital on self-efficacy and study success among first-year university students. *Learn. Individ. Differ.* 52, 109–118. doi: 10.1016/j.lindif.2016.09.016

Bruns, B., and Luque, J. (2015). *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean*. Washington, DC: World Bank. doi: 10.1596/978-1-4648-0151-8

Byrne, B. M. (2012). *Structural equation modeling with Mplus: basic concepts, applications, and programming*. New York, NY: Routledge. doi: 10.4324/9780203807644

Byrne, B. M., Shavelson, R. J., and Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychol. Bull.* 105, 456–466. doi: 10.1037/0033-2909.105.3.456

Chapman, C., Harris, A., Armstrong, P., and Chapman, C. (2012). *School effectiveness and improvement research, policy and practice*. London: Taylor & Francis. doi: 10.4324/9780203136553

Chapman, J., Schetzsle, S., and Wahlers, R. (2016). An innovative, experiential-learning project for sales management and professional selling students. *Mark. Educ. Rev.* 26, 45–50. doi: 10.1080/10528008.2015.1091674

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equat/ Model.* 14, 464–504. doi: 10.1080/10705510701301834

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *Am. Econ. Rev.* 104, 2593–2632. doi: 10.1257/aer.104.9.2593

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Struct. Equat. Model.* 9, 233–255. doi: 10.1207/S15328007SEM0902_5

Coe, R., Aloisi, C., Higgins, S., and Major, L. (2014). *What makes great teaching? Review of the underpinning research*. London: Sutton Trust.

Commission of Teacher Credentialing of California (2021). *Preparation Programs*. Sacramento, CA: Commission of Teacher Credentialing of California.

Cornelius-White, J. (2007). Learner-centered teacher-student relationships are effective: A meta-analysis. *Rev. Educ. Res.* 77, 113–143. doi: 10.3102/003465430298563

Creemers, B. P. M., and Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice, and theory in contemporary schools*. London: Routledge. doi: 10.4324/9780203939185

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching*, 2nd Edn. Alexandria, VA: Association for Supervision & Curriculum Development.

Danielson, C. (2013). *The framework for teaching evaluation instrument*. Chicago, IL: The Danielson Group.

Dulfer, N., and Akhlaghi Koopaei, F. (2021). Moving from Feedback to Feedforward in IBDP classrooms. *J. Res. Int. Educ.* 20, 101–113. doi: 10.1177/14752409211032528

Fuller, F. (1970). *Personalized education for teachers: One application of the teachers concerns model*. Austin, TX: R&D Center for Teacher Education.

Ginner Hau, H., Ferrer-Wreder, L., and Allodi, M. W. (2021). "Capitalising on Classroom Climate to Promote Positive Development," in *Handbook of Positive Youth Development. Springer Series on Child and Family Studies*, eds R. Dimitrova and N. Wiium (Cham: Springer). doi: 10.1007/978-3-030-70262-5_25

Goodwin, A. L., and Low, E. L. (2021). Rethinking conceptualisations of teacher quality in Singapore and Hong Kong: A comparative analysis. *Eur. J. Teach. Educ.* 44, 365–382. doi: 10.1080/02619768.2021.1913117

Gov.uk (2021). *Schools, pupils and their characteristics*. London: National Statistics.

Grossman, P., Loeb, S., Cohen, J. J., and Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *Am. J. Educ.* 119, 445–470. doi: 10.1086/669901

Hanushek, E. A., and Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *Am. Econ. Rev.* 100, 267–271. doi: 10.1257/aer.100.2.267

Harris, D. N., and Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *J. Public Econ.* 95, 798–812. doi: 10.1016/j.jpubeco.2010.11.009

Hattie, J. (2012). *Visible learning for teachers: Maximising impact on learning*. London: Routledge. doi: 10.4324/9780203181522

Hattie, J., and Clinton, J. (2008). "Identifying accomplished teachers: A validation study," in *Advances in program evaluation: Vol. 11. Assessing teachers for professional certification: The first decade of the National Board for Professional Teaching Standards*, eds L. Ingvarson and J. Hattie (Bingley: Emerald Group Publishing), 313–344. doi: 10.1016/S1474-7863(07)11011-5

Helms-Lorenz, M., van der Pers, M., Moorer, P., Harmsen, R., and Verkade, A. (2019). *Inductie in het Noorden: Eindrapportage 2014-2019*. Groningen: University of Groningen.

Helms-Lorenz, M., van der Pers, M., Moorer, P., Lugthart, E., van der Lans, R., and Maulana, R. (2020). *Begeleiding Startende Leraren 2014-2019*. Noorden: Eindrapportage.

Hennessy, S., Howe, C., Mercer, N., and Vrikki, M. (2020). Coding classroom dialogue: Methodological considerations for researchers. *Learn. Cult. Soc. Interact.* 25, 100404. doi: 10.1016/j.lcsi.2020.100404

Houtveen, A. A. M., and van de Grift, W. J. C. M. (2007). Effects of metacognitive strategy instruction and instruction time on reading comprehension. *Sch. Eff. Sch. Improv.* 18, 173–190. doi: 10.1080/09243450601058717

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equat. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Inspectie van het Onderwijs (2018). *The state of education in the Netherlands 2016/2017*. Utrecht: Inspectie van het Onderwijs.

Inspectie van het Onderwijs (2019). *The state of education in the Netherlands 2018*. Utrecht: Inspectie van het Onderwijs.

Irnidayanti, Y., Maulana, R., Helms-Lorenz, M., and Fadhilah, N. (2020). Relationship between teaching motivation and teaching behaviour of secondary education teachers in Indonesia. *J. Study Educ. Dev.* 43, 271–308. doi: 10.1080/02103702.2020.1722413

Ismajli, H., and Imami-Morina, I. (2018). 'Differentiated Instruction: Understanding and Applying Interactive Strategies to Meet the Needs of all the Students. *Int. J. Instruct.* 11, 207–218. doi: 10.12973/iji.2018.11315a

Jansen, E., André, S., and Suhre, C. (2013). Readiness and expectations questionnaire: A cross-cultural measurement instrument for first-year university students. *Educ. Assess. Eval. Account.* 25, 115–130. doi: 10.1007/s11092-013-9161-2

Jortveit, M., Tveit, A. D., Cameron, D. L., and Lindqvist, G. (2020). A comparative study of Norwegian and Swedish special educators' beliefs and practices. *Eur. J. Spec. Needs Educ.* 35, 350–365. doi: 10.1080/08856257.2019.1689716

Kindsvatter, R., Wilen, W., and Ishler, M. (1988). *Dynamics of effective teaching*. Harlow: Longman.

Kington, A., Sammons, P., Brown, E., Regan, E., Ko, J., and Buckler, S. (2014). *EBOOK: Effective Classroom Practice*. London: McGraw-Hill Education.

Klette, K. (2022). "The use of Video Capturing in International Large-Scale Assessment Studies: Methodological and Theoretical Considerations," in *International Handbook of Comparative Large Scale Studies in Education: Perspectives, Methods and Findings*, eds T. Nilsen, A. Stancel-Piatak, and J.-E. Gustafsson (Cham: Springer International Publishing), 470–510. doi: 10.1007/978-3-030-88178-8_19

Klieme, E., Pauli, C., and Reusser, K. (2009). "The Pythagoras Study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms," in *The power of video studies in investigating teaching and learning in the classroom*, eds T. Janík and T. Seidel (Halifax: Waxmann), 137–160.

Ko, J., and Sammons, P. (2013). *Effective teaching: A review of research and evidence*. London: CfBT Education Trust.

Kraft, M. A., and Lyon, M. A. (2022). *The Rise and Fall of the Teaching Profession: Prestige, Interest, Preparation, and Satisfaction over the Last Half Century. (Ed Working Paper: 22-679)*. Providence, RI: Annenberg Institute at Brown University.

Luoto, J. M. (2023). Comparative education and comparative classroom observation systems. *Comp. Educ.* [Epub ahead of print]. doi: 10.1080/03050068.2023.2173917

Martens, M., and Asbrand, B. (2022). "Documentary Classroom Research. Theory and Methodology," in *Dokumentarische Unterrichtsforschung in den Fachdidaktiken: Theoretische Grundlagen und Forschungspraxis*, eds M. Martens, B. Asbrand, T. Buchborn, and J. Menthe (Berlin: Springer), 19–38. doi: 10.1007/978-3-658-32566-4_2

Martin, M. O., Mullis, I. V., and Hooper, M. (2016c). *Methods and procedures in TIMSS 2015*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Martin, M. O., Mullis, I. V. S., Foy, P., and Hooper, M. (2016a). *TIMSS 2015 international results in science*. Boston, MA: TIMSS.

Martin, M. O., Mullis, I. V. S., Foy, P., and Hooper, M. (2016b). *TIMSS achievement ethodology. In Methods and procedures in TIMSS 2015*. Boston, MA: TIMSS.

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

Maulana, R., André, S., Helms-Lorenz, M., Ko, J., Chun, S., Shahzad, A., et al. (2021). Observed teaching behaviour in secondary education across six countries: measurement invariance and indication of cross-national variations. *Sch. Effective. Sch. Improve.* 32, 64–95. doi: 10.1080/09243453.2020.1777170

Maulana, R., Feng, X., Helms-Lorenz, M., Ko, J., Chun, S., Abid, S., et al. (2022). *Observing Teaching Behavior Using the International Comparative Analysis of Learning and Teaching Measure Across Countries: Is There Measurement Invariance?*. San Diego, CA: University of Worcester.

Maulana, R., and Helms-Lorenz, M. (2016). Observations and student perceptions of the quality of preservice teachers' teaching behaviour: Construct representation and predictive quality. *Learn. Environ. Res.* 19, 335–357. doi: 10.1007/s10984-016-9215-8

Maulana, R., Helms-Lorenz, M., Moorer, P., Smale-Jacobse, A. E., and Feng, X. (2023). *Differentiated Instruction in Teaching from the International Perspective: Methodological and practical insights*. Groningen: University of Groningen Press, doi: 10.21827/62c5541759973

Maulana, R., Helms-Lorenz, M., and van de Grift, W. (2015). A longitudinal study of induction on the acceleration of growth in teaching quality of beginning teachers through the eyes of their students. *Teach. Teach. Educ.* 51, 225–245. doi: 10.1016/j.tate.2015.07.003

Maulana, R., Helms-Lorenz, M., and van de Grift, W. (2017). Validating a model of effective teaching behaviour of pre-service teachers. *Teach. Teach.* 23, 471–493. doi: 10.1080/13540602.2016.1211102

Maulana, R., Opdenakker, M.-C., and Bosker, R. (2014). Teacher–student interpersonal relationships do change and affect academic motivation: A multilevel growth curve modelling. *Br. J. Educ. Psychol.* 84, 459–482. doi: 10.1111/bjep.12031

Maulana, R., Smale-Jacobse, A., Helms-Lorenz, M., Chun, S., and Lee, O. (2020). Measuring differentiated instruction in the Netherlands and South Korea: Factor structure equivalence, correlates, and complexity level. *Eur. J. Psychol. Educ.* 35, 881–909. doi: 10.1007/s10212-019-00446-4

Millsap, R. E., and Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multiv. Behav. Res.* 39, 479–515. doi: 10.1207/S15327906MBR3903_4

Moberg, S., Muta, E., Korenaga, K., Kuorelahti, M., and Savolainen, H. (2020). Struggling for inclusive education in Japan and Finland: teachers' attitudes towards inclusive education. *Eur. J. Spec. Needs Educ.* 35, 100–114. doi: 10.1080/08856257.2019.1615800

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., and Ecob, R. (1988). *School matters*. Sacramento, CA: University of California Press.

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., and Earl, L. (2014). State of the art - teacher effectiveness and professional learning. *Sch. Eff. Sch. Improv*. 25, 231–256. doi: 10.1080/09243453.2014.885451

Muijs, D., Reynolds, D., Sammons, P., Kyriakides, L., Creemers, B. P. M., and Teddlie, C. (2018). Assessing individual lessons using a generic teacher observation instrument: How useful is the International System For Teacher Observation And Feedback (ISTOF)? *ZDM Mathe. Educ.* 50, 395–406. doi: 10.1007/s11858-018-0921-9

Mullis, I. V., Martin, M. O., Foy, P., and Hooper, M. (2016). *TIMSS 2015 international results in science*. Chestnut Hill: TIMSS & PIRLS International Study Centre.

Muthén, B., and Muthén, L. (2017). *Mplus. In Handbook of item response theory*. Boca Raton, FL: Chapman and Hall/CRC, 507–518.

Noben, I., Maulana, R., Deinum, J. F., and Hofman, W. H. (2021). Measuring university teachers' teaching quality: a Rasch modelling approach. *Learn. Environ. Res.* 24, 87–107. doi: 10.1007/s10984-020-09319-w

OECD (2013). *Country note: Results from TALIS 2013: United States of America*. Paris: OECD.

OECD (2014). *Education at a Glance 2014: OECD Indicators*. Paris: OECD.

OECD (2016). *Education at a glance 2016*. Paris: OECD.

OECD (2018). *Education at a Glance 2018: OECD Indicators*. Paris: OECD. doi: 10.1787/eag-2018-en

OECD (2019). *PISA 2018 Results: Combined Executive Summaries Volume I, II, III*. Paris: OECD.

OECD (2020). *Global Teaching InSights: A Video Study of Teaching*. Paris: OECD. doi: 10.1787/20d6f36b-en

OECD (2021b). *Global teaching insights: Technical report*. Paris: OECD.

OECD (2021a). *21st-Century Readers: Developing Literacy Skills in a Digital World*. Paris: OECD.

Ofsted (2021). *Education Inspection Framework*. London: Ofsted.

Opdenakker, M.-C., and Minnaert, A. (2011). Relationship between learning environment characteristics and academic engagement. *Psychol. Rep.* 109, 259–284. doi: 10.2466/09.10.11.PR0.109.4.259-284

Paas, F., and van Merriënboer, J. J. G. (2020). Cognitive-Load Theory: Methods to Manage Working Memory Load in the Learning of Complex Tasks. *Curr. Dir. Psychol. Sci.* 29, 394–398. doi: 10.1177/0963721420922183

Pianta, R. C., and Hamre, B. K. (2009). Conceptualisation, measurement, and improvement of classroom processes: Standardised observation can leverage capacity. *Educ. Res.* 38, 109–119. doi: 10.3102/0013189X09332374

Pianta, R. C., Hamre, B. K., and Mintz, S. L. (2010). *Classroom Assessment Scoring System (CLASS): Upper elementary manual*. Charlottesville, VA: Teachstone.

Powell, J. J. (2020). Comparative education in an age of competition and collaboration. *Comp. Educ.* 56, 57–78. doi: 10.1080/03050068.2019.1701248

Reis, S. M., McCoach, D. B., Little, C. A., Muller, L. M., and Kaniskan, R. B. (2011). The effects of differentiated instruction and enrichment pedagogy on reading achievement in five elementary schools. *Am. Educ. Res. J.* 48, 462–501. doi: 10.3102/0002831210382891

Reynolds, D., Creemers, B., Stringfield, S., Teddlie, C., and Schaffer, G. (2002). *World class schools: International perspectives in school effectiveness*. London: Routledge Falmer. doi: 10.4324/9780203164624

Rijksoverheid. (2023). *Hoe word ik leraar in het voortgezet onderwijs [How can I be a teacher in secondary education?]*. Den Haag: Rijksoverheid.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *Am. Econ. Rev.* 94, 247–252. doi: 10.1257/0002828041302244

Rosenshine, B. (2010). Principles of instruction; Educational practices series. *Int. Acad. Educ.* 21, 376–391.

Rosenshine, B. (2012). Principles of Instruction: Research-Based Strategies That All Teachers Should Know. *Am. Educ.* 36, 12–39.

Rosenshine, B., and Stevens, R. (1986). "Teaching functions," in *Handbook of research on teaching*, 3rd Edn, ed. M. C. Wittrock (Berlin: Macmillan), 376–391.

Scheerens, J., and Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon Press.

Sensevy, G. (2014). Characterizing teaching effectiveness in the Joint Action Theory in Didactics: An exploratory study in primary school. *J. Curricul. Stud.* 46, 577–610. doi: 10.1080/00220272.2014.931466

Simonsen, B., Freeman, J., Myers, D., Dooley, K., Maddock, E., Kern, L., et al. (2020). The Effects of Targeted Professional Development on Teachers' Use of Empirically Supported Classroom Management Practices. *J. Pos. Behav. Interv.* 22, 3–14. doi: 10.1177/1098300719859615

Smith, T., Baker, W., Hattie, J., and Bond, L. (2008). "A validity study of the certification system of the national board for professional teaching standards," in *Assessing teachers for professional certification: The first decade of the national board for professional teaching standards*, Vol. 11, eds L. C. Ingvarson, and J. Hattie (Amsterdam: Elsevier Press), 345–380.

Steenkamp, J. B. E., and Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *J. Consum. Res.* 25, 78–90. doi: 10.1086/209528

Steinmetz, H. (2013). Analyzing observed composite differences across groups: Is partial measurement invariance enough? *Methodology* 9, 1–12. doi: 10.1027/1614-2241/a000049

Stroet, K., Opdenakker, M. C., and Minnaert, A. (2013). Effects of need supportive teaching on early adolescents' motivation and engagement: A review of the literature. *Educ. Res. Rev.* 9, 65–87. doi: 10.1016/j.edurev.2012.11.003

Suter, L. E. (2019). "The status of comparative education research in the 21st century: An Empiricist's views," in *The SAGE Handbook of Comparative Studies in Education*, eds L. E. Suter, E. Smith, and B. D. Denman (London: SAGE), 3–24. doi: 10.4135/9781526470379.n2

Teddlie, C., Virgilio, I., and Oescher, J. (1990). Development and validation of the Virgilio Teacher Behavior instrument. *Educ. Psychol. Meas.* 50, 421–430.

Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K., et al. (2003). Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classrooms: A review of literature. *J. Educ. Gifted* 27, 119–145. doi: 10.1177/016235320302700203

U.S. Department of Education (2013). *The condition of education* 2013. Available online at: https://nces.ed.gov/pubs2013/2013037.pdf

van de Grift, W., Helms-Lorenz, M., and Maulana, R. (2014). Teaching skills of student teachers: Calibration of an evaluation instrument and its value in predicting student academic engagement. *Stud. Educ. Eval.* 43, 150–159. doi: 10.1016/j.stueduc.2014.09.003

van de Grift, W. J. (2007). Quality of teaching in four European countries: A review of the literature and application of an assessment instrument. *Educ. Res.* 49, 127–152. doi: 10.1080/00131880701369651

van de Grift, W. J. (2014). Measuring teaching quality in several European countries. *Sch. Effective. Sch. Improve.* 25, 295–311. doi: 10.1080/09243453.2013.794845

van de Grift, W. J., Chun, S., Maulana, R., Lee, O., and Helms-Lorenz, M. (2017). Measuring teaching quality and student engagement in South Korea and The Netherlands. *Sch. Effective. Sch. Improve.* 28, 337–349. doi: 10.1080/09243453.2016.1263215

van de Vijver, F., and Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Eur. Rev. Appl. Psychol.* 54, 119–135. doi: 10.1016/j.erap.2003.12.004

van der Lans, R., Maulana, R., Helms-Lorenz, M., Fernández-García, C., Chun, S., de Jager, T., et al. (2021). Student perceptions of teaching quality in five countries: A Partial Credit Model approach to assess measurement invariance. *SAGE Open* 11, 21582440211040121. doi: 10.1177/21582440211040121

van der Lans, R. M., van de Grift, W. J. C. M., and van Veen, K. (2018). Developing an instrument for teacher feedback: Using the Rasch model to explore teachers' development of effective teaching strategies and behaviors. *J. Exp. Educ.* 86, 247–264. doi: 10.1080/00220973.2016.1268086

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–70. doi: 10.1177/109442810031002

Wang, M. C., Reynolds, M. C., and Walberg, H. J. (1995). *Handbook of special and remedial education: Research and practice*, 2nd Edn. Oxford: Pergamon Press.

World Bank (2022). *Teach secondary: Helping countries track and improve teaching quality*. Washington, DC: World Bank.

Xing, C., and Hall, J. A. (2015). Confirmatory factor analysis and measurement invariance testing with ordinal data: An application in revising the flirting styles inventory. *Commun. Methods Meas.* 9, 123–151. doi: 10.1080/19312458.2015.1061651

Yair, G. (2000). Not just about time: Instructional practices and productive time in school. *Educ. Adm. Quart.* 36, 485–512. doi: 10.1177/00131610021969083