



## Evidence for measuring teachers' core practices

M. Van Der Schaaf, B. Slof, L. Boven & A. De Jong

To cite this article: M. Van Der Schaaf, B. Slof, L. Boven & A. De Jong (2019) Evidence for measuring teachers' core practices, *European Journal of Teacher Education*, 42:5, 675-694, DOI: [10.1080/02619768.2019.1652903](https://doi.org/10.1080/02619768.2019.1652903)

To link to this article: <https://doi.org/10.1080/02619768.2019.1652903>



© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 20 Aug 2019.



Submit your article to this journal [↗](#)



Article views: 759



View related articles [↗](#)



View Crossmark data [↗](#)

## Evidence for measuring teachers' core practices

M. Van Der Schaaf <sup>a,b</sup>, B. Slof<sup>b</sup>, L. Boven<sup>b</sup> and A. De Jong<sup>b</sup>

<sup>a</sup>University Medical Center Utrecht, Utrecht, The Netherlands; <sup>b</sup>Department of Education, Utrecht University, Utrecht, The Netherlands

### ABSTRACT

Teaching is a complex profession and feedback on teacher practices is needed for teachers' development. Many instruments are available to measure teacher practices, but little is known about their quality. This systematic review aimed to gain insight into the quality of instruments available to measure teacher practices. A systematic review based on ERIC, PsychINFO, and Web of Science databases (2000–2016) was conducted. In total 96 journal articles were included, describing 127 measurement instruments. The instruments were mainly self-evaluation questionnaires, focussing on activities during teaching. Most evidence was provided for the validity and impact of the instruments. Evidence for utility was generally low. Questionnaire data gathered from students seems to best meet the quality requirements. It is discussed to evaluate teachers with different measurement instruments to provide a rich perspective of their practices.

### ARTICLE HISTORY

Received 26 June 2018

Accepted 25 July 2019

### KEYWORDS

Teacher development;  
professional development;  
student teacher evaluation

## 1. Introduction

To reach a high level of education that positively affects students' learning, providing high-quality feedback on teachers' practices is crucial (Darling-Hammond 2012). This demands insight into teachers' core practices, i.e. the main professional activities teachers have to carry out in the workplace, as part of their teaching profession (Grossman, Hammerness, and McDonald 2009; Reynolds 1992). Core practices underline the relevance of teacher's knowledge in action (Zeichner 2012). In contrast to the predominant focus on knowledge for teaching or on competencies in the past, core practices aim to enact with teachers' daily practice and to support high-quality content-rich, meaningful teaching (McDonald, Kazemi, and Kavanagh 2013). Teachers' core practices are often categorised into: (1) pre-lesson activities, i.e. goal setting, developing learning tasks and lessons; (2) lesson activities, i.e. instructing, guiding, and assessing and (3) post-lesson activities, i.e. reflecting on one's own teaching. Since teaching involves interacting with students in different contexts, no single measurement tool is likely to fully capture teachers' core practices. Instead, multiple types of measurement instruments and assessors are needed to foster a teachers' professional development (Grossman et al. 2014; Maulana and Helms-Lorenz 2016). It is known that measurements can have limitations as well as potential to stimulate teachers' development (Bryk et al. 2015). So far, little is

**CONTACT** M. Van Der Schaaf  [m.f.vanderschaaf-5@umcutrecht.nl](mailto:m.f.vanderschaaf-5@umcutrecht.nl)  Center for Research and Development of Education, University Medical Center Utrecht, POBox 85500, Utrecht 3508 GA, The Netherlands

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

known about the evidence that is provided for the use of measurement instruments in terms of validity, reliability, utility as well as the impact they might have on teachers' development. Insight in this topic is urgent to better ground decisions regarding the measurement of teachers' practices and associated consequences such as receiving support or getting promoted (Baartman et al. 2007; Author and Stokking 2008; Schoenherr and Hamstra 2016).

A review study is carried out to provide an overview of the evidence revealed by studies regarding the measurement of teachers' core practices. Based on Kane (2004) and Mislavy (2011), we advocate that high-quality measurements should be supported by solid and proficient theoretical and empirical rationales for their quality. For example, by a theoretically underpinning of how the domain of teacher practices is covered, and by empirical evidence based on factor analyses (e.g. confirmative or explorative). Evidence for the quality of instruments may vary regarding the amount of evidence (no evidence at all, to multiple pieces of evidence) and should at least meet the quality requirements below (Baartman et al. 2007; Clark and Sampson 2007; Author, Baartman, and Prins 2012):

### **1.1. Validity**

Evidence for validity entails proof for the relationship between the teachers' core practices, the goals and the consequences of a measurement (Messick 1989). Evidence for validity is required to reduce systematic measurement errors (e.g. construct irrelevance) and often implies: (1) content validity, i.e. the instrument covers the intended constructs to be measured, in our case teachers' core practices; (2) construct validity, i.e. the constructs are accurately operationalized in terms of items; (3) criterion validity, i.e. the instrument shows high correlations between scores on external measurements of the same constructs (convergent) or unrelated constructs (discriminant).

### **1.2. Reliability**

Evidence for reliability concerns consistency of the measurement to reduce random measurement errors (e.g. differences between tests and assessors) and often implies: (1) internal consistency: do the items measure the constructs in a consistent manner? (2) stability: does the measurement estimate the constructs consistently over time? (3) interrater agreement: do assessors reach agreement on scores given to a teacher based on the measurement?

Furthermore, evidence for *utility* and *impact* should be gathered. Utility means that the instrument is transparent, feasible and efficient to use in the workplace. A measurement's impact refers to the interpretations made, and the consequences or effects of the feedback based on the measurement and is related to the educational worthiness of the measurement (Baartman et al. 2007; Birenbaum 2007; Linn, Baker, and Dunbar 1991; Poldner, Simons, and Wijngaards 2012; Stokking, Jaspers, and Erkens 2004).

In this review study empirical studies describing the quality of measurements of teachers' core practices, used in secondary and tertiary education, were gathered and analysed to answer the main research question: 'What evidence is provided for the quality of measurements in teachers' core practices?'

## 2. Method

The review followed a four-step procedure (Hammick, Dornan, and Steinert 2010): (1) searching databases and downloading relevant articles, (2) selecting suitable articles based on inclusion and exclusion criteria, including interrater reliability, (3) coding the selected articles, including interrater reliability, and (4) reporting the findings.

### 2.1 Searching databases

A Boolean search query was conducted to search the commonly used databases – ERIC, PsycINFO, and Web of Sciences – for relevant articles (Moher et al. 2009). The query comprised keywords aimed at selecting articles including instruments that measure pre-service and qualified teachers' activities, i.e. teachers' attitude, behaviour, competence, skills, and performance. The review focuses on instruments that can be used in secondary and higher education. The query also specified that articles should *not* measure teacher's knowledge or relate to other types of education than specified, e.g. not primary or special education. We selected peer reviewed English written articles from 2000 onwards. The search query resulted in an initial set of 1,453 articles.

### 2.2 Article selection

The procedure to select suitable articles for further analysis was threefold. First, 156 articles were excluded based on duplicity ( $k = 65$ ) or unavailability of an English abstract ( $k = 91$ ). Second, two researchers (3<sup>rd</sup> and 4<sup>th</sup> author) independently scored all 1,297 remaining abstracts based on the inclusion and exclusion criteria (population, topic and target group). Inclusion criteria were: (1) (Student)teachers in secondary and tertiary education, who teach a general (not special) population of students; (2) The measurement instrument addresses teachers' *core practices* (teacher behaviour rather than knowledge) to some extent (teaching quality in general or specific skills); (3) The measurement instrument is applicable in a range of educational domains.

Exclusion criteria were: (1) Irrelevant type of education or population, for example: pre-school, toddlers, adult education, online education, medical teachers, patient education, distance education, special education; (2) The topic of the study is not teaching practice, for example: learning from feedback in general, quality of assessment in schools, evaluating a teacher education programme, intervention programmes, and quality management; (3) The study focuses on teachers' knowledge or beliefs, instead of teaching practice (skills). Examples are: subject matter *knowledge*, pedagogical content knowledge, knowledge about inclusive education or disabilities, *attitudes* towards race, learning theories, educational philosophy, teaching context, self-efficacy, and reasons to enter or leave the job; (4) Instruments that are only applicable in specific domains e.g. in teaching science, teaching language, online learning, and distance education, medical education, patient education, pre-school, toddlers, adult education, online education, distance education, special education.

Independent selection of abstracts by two researchers based on inclusion and exclusion criteria led to a percentage agreement of 88.63% and a Cohen's Kappa of .68. The abstracts that showed variance in scoring were discussed by the two researchers to

reach consensus. The procedure resulted in the exclusion of 1,074 articles. The main reason for exclusion was that an article did not focus on the combination of teachers in secondary or tertiary education and teaching activities.

Third, the remaining set of 223 articles were scored on the inclusion and exclusion criteria as mentioned above. The two researchers independently scored about 10% ( $k = 20$ ) randomly selected articles (out of the set of  $k = 223$ ). They reached a percentage agreement of 100%. This led to the exclusion of 130 articles. The main reasons for exclusion were that articles did not measure activities that teachers carry out before, during or after teaching their students, or did not provide original empirical evidence. As a result, 96 articles were included in the study, see [Table 2](#) and [Appendix 1](#) for an overview.

### **2.3 Coding of the selected articles**

Two researchers (3<sup>rd</sup> and 4<sup>th</sup> author) each coded an equal subset of the 96 selected articles according to the coding scheme. That is, all articles received a code for the topics of interest, namely 1) teacher's core practices, 2) type of instrument, 3) type of assessor, and 4) quality. When multiple types of measurement instruments were described in an article, each instrument was coded separately. This resulted in a total of  $k = 127$  instruments. Thereafter, the core practices measured by each instrument were fully described in a document by the two researchers and categorised in collaboration with the first and second author. Next, a random sample of 25 percent of all instruments was independently coded by two researchers regarding pre-lesson activities, lesson activities, and post-lesson activities. This resulted in a percentage agreement of 96,8% and a Cohen's Kappa of .59.

Finally, the quality of each measurement instrument was coded according to the degree of the provided theoretical and empirical support. First, each quality aspect was rated with 0 to 2 points (see [Table 1](#)). When no information about a quality aspect was provided in an article, it was coded as 'no evidence' (0 points). For the existing evidence we discerned between 'medium' (1 point) and 'strong' evidence (2 points) based on statistical and conceptual grounds. That is, higher scores were given when an article provided a more solid and proficient theoretical and empirical rationale for the quality of the instrument at hand. For example, one point was given for criterion validity when there was one argument reported for criterion validity, or when a correlation between .70 and .80 was shown for two criterion related measurements in the article. Two points was given for criterion validity when multiple arguments or correlations above .80 were shown. This is in line with common benchmarks for validity and reliability (DeVellis 2003; Field 2013). Similarly, higher scores were given if multiple arguments for quality are reported in an article, such as outcomes of previous studies, as this substantiates the claim for quality of the instrument. High scores on common measures as well as strong qualitative arguments both provide stronger evidence for conceptual conclusions about teacher performance in practice and were thus coded as 'strong' evidence. Subsequently, the sum scores for aspects of validity (content, construct and criterion validity) and reliability (internal consistency, stability and interrater agreement) ranged from 0 to 6 points, i.e. 0 to 2 points for each quality aspect. Consequently, the sum scores for aspects of validity and reliability were divided into three categories: 'no evidence for quality' (average sum score 0), 'medium evidence for quality' (average sum score 1–3), and 'strong evidence quality' (average sum score 4–6).

**Table 1.** Coding of quality requirements.

| Requirement | Aspect               | Description   | Score  |
|-------------|----------------------|---|--|
| Validity    | Content              | The constructs, i.e. teachers' practices, are described in previous studies and pilots.   | 0 = No evidence<br>1 = Previous study or pilot<br>2 = Previous study and pilot   |
|             | Construct            | The operationalisation of the constructs are described in terms of Factor-, Rasch- and/or Item Characteristic Curves (ICC)-analyses.              | 0 = No evidence<br>1 = One analysis<br>2 = Two or more analyses  |
|             | Criterion            | The outcome of the measurement is related to the outcomes of other measurements that intend to measure the same construct.                        | 0 = No evidence or low correlation<br>1 = Correlation between .70 – .80 with measurement of related construct or a theoretical argument(s) that the outcome differs from that of the measurement of an unrelated construct<br>2 = Correlation > .80 with measurements of related construct and multiple theoretical arguments for criterion validity |
| Reliability | Internal consistency | The items measure the construct consistently in terms of Cronbach's Alpha or Kuder-Richardson analyses  | 0 = No evidence or low consistency score<br>1 = Consistency score between .70 – .80<br>2 = Consistency score > .80 and multiple analyses   |
|             | Stability            | The instrument measures the construct consistently over time in terms of test-retest and split-half analyses.                                     | 0 = No evidence or low consistency score<br>1 = Consistency score between .70 – .80.<br>2 = Consistency score > .80 and multiple analyses  |
|             | Interrater agreement | They assessors reach sufficient agreement in terms of percentage agreement and Cohen's Kappa.   | 0 = No evidence or low percentage of agreement<br>1 = Percentage agreement and Cohens' Kappa .60 – .80.<br>2 = Percentage agreement and Cohens' Kappa > .80 and multiple analyses.   |
| Utility     |                      | The measurement procedure is transparent and efficient in terms of required time, experienced difficulties, and suitability for the target group. | 0 = No evidence.<br>1 = One type or multiple types of evidence.  |
| Impact      |                      | The consequences of the measurement are described in terms of feedback and the relation with the obtained score.                                  | 0 = No evidence.<br>1 = One type or multiple types of evidence   |

In utility and impact, we scored the number of pieces of evidence that indicate the transparency and efficiency of the use of the instrument, as well as evidence for the way the use of the instrument can contribute to teachers' professional development, e.g. by means of feedback, reflection or otherwise. The sum scores for evidence of utility and impact were, thereafter, divided into two categories: 'no evidence for quality' (average sum score 0) and 'one or multiple pieces of evidence for quality' (average sum score 1–2). Two researchers (3<sup>rd</sup> and 4<sup>th</sup> author) independently coded the same 15% of the instruments, randomly chosen from the set of  $k = 127$  instruments. This resulted in a percentage agreement and weighted Cohen's Kappa for validity (72.2%, .58), reliability (88.9%, .69), utility (72.2%, .40), and impact (88.9%, .77).

After coding the instruments, we ran descriptive statistics and frequencies to create an overview on what claims for quality can be found in literature for the different assessment instruments. Based on the coding, we also examined whether triangulation procedures (e.g. multiple assessment moments, different instruments, different assessors) were applied to measure teachers' core practices. This was done by scoring how many assessment moments and what kind of instruments and assessors were considered when assessing teachers' core practices.

### 3. Results

#### 3.1 Measurement instruments and assessors

The 96 selected articles described 127 instruments for measurement of teacher practices (see [Tables 2](#) and [3](#) for an overview). When looking at the *type of teachers' core practices* the obtained results indicate that most instruments (72%) are aimed at measuring lesson-related activities ( $k = 92$ ). Within this category there is a focus on measuring content-related activities ( $k = 51$ ), such as the Learning Object Evaluation Scale (Kay and Knaack 2008), and aspects of classroom climate, including interaction ( $k = 49$ ), such as the Questionnaire on Teacher Interaction (Wubbels et al. 2012). There seems to be less interest in measuring post-lesson activities ( $k = 24$ ), such as keeping reflective journals or measuring teacher collaboration on schoolwide intervention programs (e.g. Martinez et al. 2016). There were barely instruments aimed at measuring the pre-lesson activities ( $k = 1$ ) or on a combination of aforementioned activities ( $k = 10$ ). A scarce example of such an instrument is the Classroom Assessment Scoring System (CLASS), measuring content and climate aspects (Allen et al. 2013).

When distinguishing between the *different types of instruments*, the obtained results indicate that especially questionnaires (72%) were used to measure teachers' core practices. In addition, observation instruments (14%), interviews (6%), and performance and logbooks (8%) were used. Further, the obtained results indicate that assessment of teachers' core practices is based on data gathered from *different kinds of assessors*. That is, the measurement is based on: (1) teachers' self-reports (44% of the instruments), (2) students' assessment of their teachers (34%) and (3) perceptions of colleagues and supervisors (21%). Mainly questionnaires were used to gather data from the teachers and their students. Observations (67%) and interviews (22%) were mostly used to gather data from the colleagues and supervisors. Only in one case, different kind of instruments

**Table 2.** Overview of studies and teacher practices.

| Study  | Teacher Practice  |
|--|---|
| Ang (2005), Azigwe et al. (2016), Bear et al. (2011), Bernardo et al. (2008), Bonney et al. (2015), Casas et al. (2015), Castillo et al. (2013), den Brok et al. (2002), Fan (2012), Fenzel et al. (2009) <sup>1</sup> , Hill (2004), Huang et al. (2006), Jacobs et al. (2013), Kiany et al. (2011), Kosir et al. (2014), Kunter et al. (2007), Kyriakides (2005), Lang et al. (2005), Martin et al. (2010) <sup>1</sup> , Maulana et al. (2015), Maulana et al. (2012), Murray et al. (2011a), Passini et al. (2015), Ryan et al. (2011), Sakiz (2012), Scrimin et al. (2014), She et al. (2000), Siddall, et al. (2013), Skinner et al. (2008), Thijs et al. (2012), Veldman et al. (2013), Wallace et al. (2012), Walsh et al. (2010), Zullig et al. (2014)  | Classroom Climate<br>Examples: interpersonal behaviour, school climate, communication, rule clarity, encouragement.       |
| Al-Shabatat (2014), Azigwe et al. (2016) <sup>1</sup> , Barton et al. (2006) <sup>2*</sup> , Beran et al. (2009), Beswick (2005), Boardman et al. (2004) <sup>123</sup> , Brown, G.T.L. et al. (2015) <sup>4</sup> , Driscoll et al. (2010), Emesini (2015), Flowers et al. (2000), Goh et al. (2010), Good et al. (2015) <sup>1</sup> , Hailaya et al. (2014) <sup>3</sup> , Hargreaves (2014) <sup>1</sup> , Haydn et al. (2007), Kay et al. (2008) <sup>3*</sup> , Khourey-Bowers et al. (2005) <sup>1</sup> , Klug et al. (2014) <sup>2</sup> , Kyriakides et al. (2009), Kyriakides et al. (2014), Larose et al. (2009), Martinez et al. (2016), Mehta et al. (2013), Meintjes et al. (2010) <sup>3</sup> , Murray et al. (2011), Nelson et al. (2014), Nunnery et al. (2008) <sup>1</sup> , Opendakker et al. (2011), Panayiotou et al. (2014), Park et al. (2014) <sup>1</sup> , Peters et al. (2010) <sup>1</sup> , Pontefract et al. (2005) <sup>1</sup> , Reddy et al. (2015) <sup>1*</sup> , Reddy et al. (2013) <sup>1*</sup> , Robertson et al. (2013) <sup>2</sup> , Sach (2012), Schroeder et al. (2011), Schumacher et al. (2011) <sup>2</sup> , Watzke (2007), Williams et al. (2007) <sup>2*</sup> | Didactics<br>Examples: assessments, designing learning tasks, instruction, feedback, research-based practice.             |
| Castillo et al. (2013), Cengiz et al. (2015) <sup>3</sup> , Daley et al. (2015) <sup>3</sup> , Delvaux et al. (2013), Elstad et al. (2012), Huang, et al. (2006), Huang et al. (2009), Kalk et al. (2014), Landmann (2013), Martinez et al. (2016), Neves et al. (2014), Ordu (2016), Schoeman et al. (2012), Sirin, S. R. et al. (2010) <sup>3</sup> , Veldman et al. (2013) <sup>2</sup> , Vanlommel et al. (2016), Yates (2007), Zurlo et al. (2013)  | Professional Development<br>Examples: professionalisation, role in the school, communication with colleagues and parents. |
| Allen et al. (2013) <sup>1</sup> , Brown, E.L. et al. (2015) <sup>3</sup> , Gitomer et al. (2014) <sup>1*</sup> , Reese et al. (2014) <sup>1</sup> , Strong et al. (2011),   | Combination of aforementioned practices   |

All studies refer to questionnaires, except when one of the following codes is given: 1 = observation; 2 = interview; 3 = performance measurements and logbooks; An asterisks indicates that two instruments are included, of which at least one is a questionnaire another is indicated by the predisplayed code, that is: 1\* = observation and questionnaire; 2\* = interview and questionnaire; 3\* = perf. ass. and logs, and questionnaire

**Table 3.** Overview of teacher practices measured in assessment instruments.

| Teacher practice       | Type of assessment instrument |             |           |                     | Total    |
|------------------------|-------------------------------|-------------|-----------|---------------------|----------|
|                        | Questionnaire                 | Observation | Interview | Perf. ass. and logs |          |
| Pre-lesson activities  | 0                             | 0           | 0         | 1                   | 1 (1%)   |
| Lesson activities      | 69                            | 14          | 4         | 5                   | 92 (72%) |
| Post-lesson activities | 18                            | 0           | 3         | 3                   | 24 (19%) |
| Combined activities    | 4                             | 4           | 1         | 1                   | 10 (8%)  |
| Total                  | 91 (72%)                      | 18 (14%)    | 8 (6%)    | 10 (8%)             | 127      |

and assessors (data triangulation) were used to measure teachers' core practices. This indicates that teachers' core practices are commonly measured at one specific moment in time with a one specific instrument or assessor.



### 3.2 Evidence for instrument quality

The obtained results for each quality aspect are reported per type of instrument and assessor (see Tables 4 and 5). Analysis of the *validity quality aspect* revealed that most instruments were rated as ‘medium’ evidence (76%). This indicates that for the majority of the instruments a theoretical or empirical argument regarding the measurement, for instance a factor analysis, was provided. For respectively 12% and 13% of the instruments ‘strong’ evidence or ‘no’ evidence was provided. Instruments that gathered data from students about their teachers, received, relatively more ‘strong evidence’ scores (21% of the instruments) compared to teachers’ self-report data and data from colleagues and supervisors (7% of the instruments).

Analysis of the *reliability quality aspect* indicated that, again, for the majority of the instruments (66%) at least some evidence (e.g. internal consistency) was provided. In respectively 29% and 5% of the cases ‘no’ evidence or ‘strong’ evidence was provided. Instruments in which data was gathered from colleagues and supervisors received, relatively speaking, more scores in the category ‘no evidence’ (44% of the instruments) compared to instruments based on teachers’ self-reports or student data (26%).

Analysis of the *utility quality aspect* revealed that the majority of the instruments provided no evidence (70%). In other words, no information was given regarding the transparency, feasibility and efficiency of the instrument. In 30% of the cases, at least, information on one of these topics was provided.

**Table 4.** Overview of quality of evidence per type of assessment instruments.

| Quality     | Degree of Evidence     | Type of assessment instrument |             |           |                      | Total      |
|-------------|------------------------|-------------------------------|-------------|-----------|----------------------|------------|
|             |                        | Questionnaire                 | Observation | Interview | Perf. meas. logbooks |            |
| Validity    | No evidence            | 10                            | 4           | 2         | 0                    | 16 (12,5%) |
|             | Medium                 | 69                            | 13          | 6         | 8                    | 96 (75,5%) |
|             | Strong                 | 12                            | 1           | 0         | 2                    | 15 (12,0%) |
| Reliability | No evidence            | 23                            | 7           | 6         | 1                    | 37 (29%)   |
|             | Medium                 | 65                            | 9           | 2         | 8                    | 84 (66%)   |
|             | Strong                 | 3                             | 2           | 0         | 1                    | 6 (5%)     |
| Utility     | No                     | 63                            | 12          | 5         | 8                    | 88 (69%)   |
|             | One or multiple pieces | 28                            | 6           | 3         | 2                    | 39 (31%)   |
| Impact      | No                     | 34                            | 7           | 4         | 6                    | 51 (40%)   |
|             | One or multiple pieces | 57                            | 11          | 4         | 4                    | 76 (60%)   |

**Table 5.** Overview quality of evidence per type of assessor.

| Quality     | Degree of evidence     | Type of assessor |        |        |          | Total    |
|-------------|------------------------|------------------|--------|--------|----------|----------|
|             |                        | Self-assessment  | Pupils | Others | Multiple |          |
| Validity    | No evidence            | 7                | 4      | 5      | 0        | 16 (13%) |
|             | Medium                 | 45               | 30     | 20     | 1        | 96 (76%) |
|             | Strong                 | 4                | 9      | 2      | 0        | 15 (12%) |
| Reliability | No evidence            | 14               | 11     | 12     | 0        | 37 (29%) |
|             | Medium                 | 40               | 31     | 12     | 1        | 84 (66%) |
|             | Strong                 | 2                | 1      | 3      | 0        | 6 (5%)   |
| Utility     | No                     | 41               | 27     | 19     | 1        | 88 (69%) |
|             | One or multiple pieces | 15               | 16     | 8      | 0        | 39 (31%) |
| Impact      | No                     | 26               | 14     | 11     | 0        | 51 (40%) |
|             | One or multiple pieces | 30               | 29     | 16     | 1        | 76 (60%) |

Analysis of the *impact quality aspect* revealed that 60% of the instruments included a theoretical rationale for the formative or summative consequences. Another 40% of the instruments did not provide any information about this quality aspect. Furthermore, specific evidence about how an instrument could enhance teachers' professional development was lacking.

#### 4. Discussion

The aim of this review study was to gain more insight into the quality of instruments that measure teachers' core practices. Quality was coined as the kinds of measurement instruments and assessors involved as well as the provided evidence for their quality, i.e. validity, reliability, utility, and impact. The findings indicated that the instruments mainly focused on measuring teachers' lesson related core practices, especially classroom climate and didactic activities. Although these may be regarded as central categories of teaching activities, it is remarkable that almost no attention was paid to the measurement of teachers' professionalisation activities. After all, the measurement of teachers' core practices can be very helpful to stimulate further professional development, as the measurements themselves can be seen as 'interventions' in teachers' practices and often demand reflection, feedback and collaboration to be carried out (Butler 2006; Rezgui, Mhiri, and Ghédira 2014). So, from a meta perspective, the measurements of teachers' core practices could stimulate teachers' development within the social cultural work context (Butler 2006) and for that reason it is relevant to include teachers' professionalisation as an aim in itself more prominently. In other words, since teachers develop as part of their daily work (Billet 1998), and learning is mainly informally based on learning from experiences at the workplace (Eraut 1994), measuring a teacher's core practices is an intervention with potential impact on professional development.

Findings also indicated that mainly questionnaires (72%) were administered to measure teachers' core practices. This aligns with today's practice in which teachers often work with instruments for self-evaluation purposes (Darling-Hammond et al. 2012). Results regarding the quality of the questionnaires indicated that in most cases (70%) a 'medium' level of evidence was provided for the validity and reliability of the instrument. In 60% of the cases a 'medium' level of evidence was provided for the impact of the instrument. Only in 30% of the cases evidence for the utility of the questionnaire was provided. Most observation and interview instruments also did not fully meet the quality criteria to measure teachers' core practices. Alternative forms of measurements, based on performances measures, such as simulations or portfolios and logbooks provided more evidence for their validity (100% of the cases provide at least a 'medium' level of evidence) and reliability (90% of the cases provide at least a 'medium' level of evidence). As the complex profession of teaching is not easily reconciled with traditional measurements, more research is needed into evidence for the quality of such alternative forms of measuring teacher practices.

The overarching interpretation of the obtained findings is that for each type of instrument, regardless of the assessor, sound theoretical and empirical evidence for its quality is lacking. This aligns with findings of several authors who reported that measuring teaching quality is a complex endeavour, which is based on a diversity of theoretical constructs and methodologies with their specific validity and reliability issues (Feistauer

and Richter 2017; Goos and Salomons 2017; Gunn 2018; Spooren, Brock, and Mortelman 2013). In addition to these studies, the current literature review also took edumetric quality aspects, such as utility and impact, into consideration. As indicated before, these quality aspects are also subject for improvement in most measurement instruments. Without a sound description of the quality of instruments to measure teachers' core practices, users should be cautious when interpreting the obtained scores. It is for instance well known that instruments that assume to measure students' perceptions of how they were taught, i.e. student evaluations of teaching, also include other constructs such as personal traits and attractiveness of the teacher, which might hamper their validity (Clayson and Sheffet 2006; Hornstein 2017; Sax, Gilmartin, and Bryant 2003).

When interpreting the findings of this review study, one should consider some limitations. First, despite the high agreement percentages between the coders, the Cohen's Kappas for the category's validity, utility and impact could be improved. This implies that the findings of this study should be taken with caution. Further, the analyses were based on information available in selected studies. It is possible that authors of the selected studies did not publish all information that we are rating the quality of instruments on regarding developed and validated instruments in full-text articles.

To conclude, three suggestions for advancing the development and application of instruments aimed at measuring teachers' core practice are provided. First, since all measurement instruments have their benefits and pitfalls it seems feasible to shift to a different, more constructivist, assessment approach, based on triangulation. We advocate that assessment data from multiple sources (type of instrument and assessor) and different moments in time should be used to gain a better understanding of a teachers' professional development (cf. Catrysse et al. 2016). Besides teachers' self-assessments and student evaluations one might also want to collect behavioural data, such as lesson observations (van de Grift 2007), and information from colleagues and supervisors. By doing so, more insight into teacher's core practices can be provided and contrasting findings can be placed into perspective. This is important since the quality of measurements is also affected by many context related factors, including how well the instrument fits with a teacher's work context, the support and time a teacher receives to prepare for and to reflect on the measurements and the training and support of assessors. This is in line with a programmatic assessment approach (Van der Vleuten et al. 2015) in which the use of multiple measurements, preferably with different kinds of assessors, and discussion of the obtained scores is advocated.

Secondly, it seems feasible to devote more attention to the utility and the impact quality aspects. When selecting a specific instrument or a combination of instruments, assessors need a proper understanding of how they should use the instrument(s). To this end, information about the transparency, feasibility and efficiency is required. In this respect, for example, Feistauer and Richter (2017) indicated that questionnaire data from at least 25 students should be collected in order to obtain meaningful findings from this type of instrument. If this sample size is not available, such an instrument should not be used to assess teachers' core practices. Further, we advocate that instruments more clearly state for which purpose (e.g. the purpose of furthering professional development) the measurement will be used and what the associated consequences are for the assessed teacher. This aligns with Duckworth and Yeager's (2015) recommendations that practitioners should seek for the most valid measure given their intended aims.

When a sound theoretical and empirical underpinning of the instrument quality is lacking, we recommend to solely use the instruments as diagnostic tools.

Third, considering the measurement of teachers' core practices as a source for further professional development by means of diagnostic tools, stresses an explicit and strong relation with teachers' learning processes and work context. The development of teachers requires measurement instruments that can be used to improve their daily practice (Bryck et al. 2015). This demands that they are feasible, context-rich and directly related to how students learn. The focus should not lie on the quality of the instruments only, but rather on the question how to make measurements work effectively for individual teachers. This implies that instruments should be embedded in school organisations and teacher education pedagogies for becoming teachers. The enactment and embeddedness in practice is necessary to contribute to teachers' professional development.

Fourth, in this regard measurement instruments should be related to teachers' learning processes. For instance, giving room for teachers' agency in measuring core practices is relevant, e.g. in terms of formulating own learning goals that meet their experiences in the context at hand. We hardly found indications in our study of how the instruments to measure teachers' core practices align with or feed into teachers' learning processes. This implies a need for narrative and interpretative approaches that are able to describe processes and development in the rich context of teaching (Eraut 1994; Sandberg 1994). It also includes paying attention to the purposes and intentions underneath teachers' actions (Kennedy 2015). In our study we only found some instruments that paid attention to narratives or qualitative data regarding teacher development at the work place. It is recommended to invest more in teachers' agency and the use of narrative data to provide meaning for teachers, when measuring their core practices (Bouwen 1998).

On the whole, the most important goal of measuring teachers' core practices is to provide teachers with feedback that can stimulate professional development. In this regard, impact is one of the most important quality criteria. In the end, teachers' development can only be stimulated when the measurements are used as means for dialogues with students, peers and supervisors about teaching, as well as instruments for further reflection. Therefore, it is valuable to integrate information of how students learn to how teachers teach (Darling-Hammond 2015). Useful examples are: multisource feedback, possibilities for peer review and peer feedback from and by teachers, involving students (formatively) in the evaluation of education, the role of teamwork in teaching. This study shows that instruments to measure teachers' core practices can improve in theoretical and empirical argumentation and that evidence for utility of the measurements is in general weak. Consequently, there is a need for another, more constructivist view on what is crucial for quality measurements, as well as a stronger focus on the teacher as learner in the context of the workplace.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Notes on contributors

**M. van der Schaaf**, PhD, is a professor of Research and Development of Health Professions Education at University Medical Center Utrecht and a researcher at the department of Education at Utrecht University in the Netherlands. Her research focuses on feedback that stimulates professional development, learning analytics, and expertise of health professionals.

**B. Slof**, PhD, is an assistant professor at Utrecht University, Department of Education, at the Faculty of Social and Behavioral Sciences, the Netherlands. His research focuses on studying the effects of feedback and visualizations, fostering collaborative problem solving, learning analytics and the professionalization of (student) teachers.

**L. Boven**, MSc, is an educational researcher at the institute for applied natural sciences, TNO, at the department of Training and Performance innovations, the Netherlands. Her research focuses on how to improve human performance of individuals and teams in complex and demanding environments.

**A. de Jong**, MSc, is a doctoral candidate at Utrecht University, Department of Education, at the Faculty of Social and Behavioral Sciences and researcher at Oberon, a research and consultancy institute in the Netherlands. Her doctoral research focuses on formal and shared leadership in schools in the context of collaborative innovation. Her research at Oberon mainly focuses on professionalization of teachers and school principals, learning culture and social networks in schools.

## ORCID

M. Van Der Schaaf  <http://orcid.org/0000-0002-4810-5464>

## References

- Allen, J., A. Gregory, A. Mikami, J. Lun, B. Hamre, and R. Pianta. 2013. "Observations of Effective Teacher-Student Interactions in Secondary School Classrooms: Predicting Student Achievement with the Classroom Assessment Scoring System – Secondary." *Grantee Submission* 42 (1): 76.
- Author, L., Baartman, A. L., and F. Prins. 2012. "Exploring the Role of Assessment Criteria during Teachers' Collaborative Judgement Processes of Students' Portfolios." *Assessment & Evaluation in Higher Education* 37 (7): 847–860. doi:10.1080/02602938.2011.576312.
- Author and Stokking, K. M. 2008. "Developing and Validating a Design for Teacher Portfolio Assessment." *Assessment & Evaluation in Higher Education* 33 (3): 245–262. doi:10.1080/02602930701292522.
- Baartman, L. K. J., T. J. Bastiaens, P. A. Kirschner, and C. P. M. van der Vleuten. 2007. "Evaluating Assessment Quality in Competence-based Education: A Qualitative Comparison of Two Frameworks." *Educational Research Review* 2 (2): 114–129. doi:10.1016/j.edurev.2007.06.001.
- Billet, S. 1998. "Understanding Workplace Learning: Cognitive and Sociocultural Perspectives." In *Current Issues and New Agendas in Workplace Learning*, edited by D. Boud, 47–68. Springfield, Australia: National Centre for Vocational Education Research.
- Birenbaum, M. 2007. "Assessment and Instruction Preferences and Their Relationship with Test Anxiety and Learning Strategies." *Higher Education* 53 (6): 749–768. doi:10.1007/s10734-005-4843-4.
- Bouwen, R. 1998. "Relational Construction of Meaning in Emerging Organization Contexts." *European Journal of Work and Organizational Psychology* 7 (3): 299–319.
- Bryk, A. S., L. M. Gomez, A. Grunow, and P. G. LeMahieu. 2015. *Learning to Improve: How America's Schools Can Get Better at Getting Better*. Cambridge: Harvard Education Press.
- Butler, P. 2006. *A Review of the Literature on Portfolios and Electronic Portfolios*. Palmerston North, New Zealand: Massey University College of Education.

- Catrysse, L., D. Gijbels, V. Dochy, S. De Maeyer, P. Van Den Bossche, and L. Gommers. 2016. "Mapping Processing Strategies in Learning from Expository Text: an Exploratory Eye Tracking Study Followed by a Cued Recall." *Frontline Learning Research* 4: 1. doi:10.14786/flr.v4i1.192.
- Clark, D. B., and V. D. Sampson. 2007. "Personally Seeded Discussions to Scaffold Online Argumentation." *International Journal of Science Education* 29 (3): 253–277. doi:10.1080/09500690600560944.
- Clayson, D. E., and M. J. Sheffet. 2006. "Personality and the Student Evaluation of Teaching." *Journal of Marketing Education* 28: 149–160.
- Darling-Hammond, L. 2012. *Creating a Comprehensive System for Evaluating and Supporting Effective Teaching*. Stanford, CA: Stanford Center for Opportunity Policy in Education.
- Darling-Hammond, L. 2015. "Can Value Added Add Value to Teacher Evaluation?" *Educational Researcher* 44 (2): 12–137. doi:10.3102/0013189X15575346.
- DeVellis, R. F. 2003. *Scale Development: Theory and Applications*. Thousand Oaks, California: Sage Publications, Inc..
- Duckworth, A. L., and D. S. Yeager. 2015. "Measurement Matters: Assessing Personal Qualities Other than Cognitive Ability for Educational Purposes." *Educational Researcher* 44 (4): 237–251. doi:10.3102/0013189X15584327.
- Eraut, M. E. 1994. *Developing Professional Knowledge and Competence*. London: Falmer Press.
- Feistauer, D., and T. Richter. 2017. "How Reliable are Students' Evaluations of Teaching Quality? A Variance Components Approach." *Assessment & Evaluation in Higher Education* 42 (8): 1263–1279. doi:10.1080/02602938.2016.1261083.
- Field, A. 2013. *Discovering Statistics Using SPSS*. Fourth Ed. London: Sage.
- Goos, M., and A. Salomons. 2017. "Measuring Teaching Quality in Higher Education: Assessing Selection Bias in Course Evaluations." *Research in Higher Education* 58: 341–364. doi:10.1007/s11162-016-9429-8.
- Grossman, P., J. Cohen, M. Ronfeldt, and L. Brown. 2014. "The Test Matters: the Relationship between Classroom Observation Scores and Teacher Value Added on Multiple Types of Assessment." *Educational Researcher* 43 (6): 293–303. doi:10.3102/0013189X14544542.
- Grossman, P., K. Hammerness, and M. McDonald. 2009. "Redefining Teaching, Re-imagining Teacher Education." *Teachers and Teaching: Theory and Practice* 15 (2): 273–289. doi:10.1080/13540600902875340.
- Gunn, A. 2018. "Metrics and Methodologies for Measuring Teaching Quality in Higher Education: Developing the Teaching Excellence Framework (TEF)." *Educational Review* 70 (2): 129–148. doi:10.1080/00131911.2017.1410106.
- Hammick, M., T. Dornan, and Y. Steinert. 2010. "Conducting a Best Evidence Systematic Review. Part 1: from Idea to Data Coding. BEME Guide No. 13." *Medical Teacher* 32 (1): 3–15. doi:10.3109/01421590903414245.
- Hornstein, H. A. 2017. "Student Evaluations of Teaching are an Inadequate Assessment Tool for Evaluating Faculty Performance." *Cogent Education* 4: 1304016. doi:10.1080/2331186X.2017.1304016.
- Kane, M. 2004. "Certification Testing as an Illustration of Argument-based Validation." *Measurement: Interdisciplinary Research & Perspective* 2 (3): 135–170. doi:10.1207/s15366359mea0203\_1.
- Kay, R. H., and L. Knaack. 2008. "A Multi-component Model for Assessing Learning Objects: the Learning Object Evaluation Metric (LOEM)." *Australasian Journal of Educational Technology* 24 (5): 574–591. doi:10.14742/ajet.1192.
- Kennedy, A. 2015. "What Do Professional Learning Policies Say about Purposes of Teacher Education?" *Asia-Pacific Journal of Teacher Education* 43 (3): 183–194. doi:10.1080/1359866X.2014.940279.
- Linn, R. L., E. L. Baker, and S. B. Dunbar. 1991. "Complex, Performance-based Assessment: Expectations and Validation Criteria." *Educational Researcher* 20 (8): 15–21. doi:10.3102/0013189X020008015.
- Martinez, A., S. D. McMahon, C. Coker, C., and C. B. Keys. 2016. "Teacher Behavioral Practices: Relations to Student Risk Behaviors, Learning Barriers, and School Climate." *Psychology in the Schools* 53 (8): 817–830. doi:10.1002/pits.21946.
- Maulana, R., and M. Helms-Lorenz. 2016. "Observations and Student Perceptions of the Quality of Preservice Teachers' Teaching Behaviour: Construct Representation and Predictive Quality." *Learning Environments Research* 19 (3): 335–357. doi:10.1007/s10984-016-9215-8.

- McDonald, M., E. Kazemi, and S. S. Kavanagh. 2013. "Core Practices and Pedagogies of Teacher Education: A Call for A Common Language and Collective Activity." *Journal of Teacher Education* 64 (5): 378–386. doi:10.1177/0022487113493807.
- Messick, S. 1989. "Validity." In *Educational Measurement*, edited by R. L. Linn, 13–103. 3rd ed. New York: MacMillan.
- Mislevy, R. J. 2011. "Evidence-centered Design for Simulation-based Assessment." *Cresst report 800*. Retrieved from: <http://www.cse.ucla.edu/products/reports/R800.pdf>
- Moher, D., A. Liberati, J. Tetzlaff, D. G. Altman, and D. G. The PRISMA Group. 2009. "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA Statement." *PLoS Med* 6 (6): e1000097. doi:10.1371/journal.pmed1000097.
- Poldner, E., P. R. J. Simons, and G. Wijngaards; Author. 2012. "Quantitative Content Analysis Procedures to Analyse Students' Reflective Essays: A Methodological Review of Psychometric and Edumetric Aspects." *Educational Research Review* 7 (1): 19–37. doi:10.1016/j.edurev.2011.11.002.
- Reynolds, A. 1992. "Getting to the Core of the Apple: A Theoretical View of the Knowledge Base of Teaching." *Journal of Personnel Evaluation in Education* 6 (1): 41–55. doi:10.1007/BF00126919.
- Rezgui, K., H. Mhiri, and K. Ghédira. 2014. "Ontology-based e-Portfolio Modeling for Supporting Lifelong Competency Assessment and Development." *Procedia Computer Science* 112: 397–406. doi:10.1016/j.procs.2017.08.04.
- Sandberg, J. 1994. "Human competence at work – an interpretative approach." Doctoral dissertation. Gotenberg, Sweden.
- Sax, L. J., S. K. Gilmartin, and A. N. Bryant. 2003. "Assessing Response Rates and Non-Response Bias in Web and Paper Surveys." *Research in Higher Education*, 44: 409–432.
- Schoenherr, J. R., and S. Hamstra. 2016. "Psychometrics and Its Discontents: an Historical Perspective on the Discourse of the Measurement Tradition." *Advances in Health Sciences Education* 21 (3): 719–729. doi:10.1007/s10459-015-9623-z.
- Spooren, P., B. Brockx, and B. A. D. Mortelmans. 2013. "On the Validity of Student Evaluation of Teaching: the State of the Art." *Review of Educational Research* 83 (4): 598–642.
- Stokking, K., A. J. Jaspers, and G. Erkens. 2004. "Teachers' Assessment of Students' Research Skills." *British Educational Research Journal* 30 (1): 93–116. doi:10.1080.01411920310001629983.
- 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: 127–152.
- Van der Vleuten, C. P. M., L. W. T. Schuwirth, E. W. Driessen, M. J. B. Govaerts, and S. Heeneman. 2015. "12 Tips for Programmatic Assessment." *Medical Teacher* 37: 641–646. doi:10.3109/0142159X.2014.973388.
- Wubbels, T., P. Den Brok, P. J. van Tartwijk, and J. Levy. 2012. *Interpersonal relationships in education: An overview of contemporary research* Vol. 3. Rotterdam: Springer Science & Business Media.
- Zeichner, K. 2012. "The Turn once Again toward Practice-based Teacher Education." *Journal of Teacher Education* 63 (5): 376–382. doi:10.1177/0022487112445789.

## Appendix 1. Selected articles including instruments for teacher assessment ordered by pre-lesson activities, lesson activities or post-lesson activities or a combination

### *Pre-lesson activities*

- (1) Meintjes, H., and M. Grosser. 2010. "Creative thinking in prospective teachers: the status quo and the impact of contextual factors." *South African Journal of Education*, 30 (3): 361–386.

### *Combination Pre-lesson activities and Lesson activities*

- (2) Allen, J., A. Gregory, A. Mikami, J. Lun, B. Hamre, and R. Pianta. 2013. "Observations of effective teacher-student interactions in secondary school classrooms: Predicting student

- achievement with the classroom assessment scoring system–secondary." *Grantee Submission*, 42 (1): 76.
- (3) Kourey-Bowers, C., R. Dinko, and R. Hart. 2005. "Influence of a shared leadership model in creating a school culture of inquiry and collegiality." *Journal of Research in Science Teaching*, 42 (1): 3-24. doi:10.1002/tea.20038.
  - (4) Murray, D. W., D.L. Rabiner, and K.K. Hardy. 2011b. "Teacher management practices for first graders with attention problems." *Journal of Attention Disorders*, 15 (8): 638-645. doi:10.1177/10870547110378234.
  - (5) Opdenakker, M., and A. Minnaert. 2011. "Relationship between learning environment characteristics and academic engagement." *Psychological Reports*, 109 (1): 259-284. doi:10.2466/09.10.11.PR0.109.4.259-284.
  - (6) Panayiotou, A., L., Kyriakides, B.P.M. Creemers, L. McMahon, G. Vanlaar, M. Pfeifer, G. Rekalidou, and M. Bren. 2014. "Teacher behavior and student outcomes: Results of a European study." *Educational Assessment and Accountability*, 26 (1): 73-93. doi:10.1007/s11092-013-9182-x.
  - (7) Park, J., Y. Park, Y. Kim, J. Park, and J. Jeong. 2014. "The Development of the Korean Teaching Observation Protocol (Ktop) for Improving Science Teaching and Learning." *Journal of Baltic Science Education*, 13 (2): 259-275.
  - (8) Schumacher, G., B. Grigsby, and W. Vesey. 2011. „Development of research-based protocol aligned to predict high levels of teaching quality." *International Journal of Educational Leadership Preparation*, 6 (4).

#### *Lesson activities*

- (9) Al-Shabatat, A. 2014. "Gifted teachers' stages of concerns for integrating e-learning in the gifted schools in Jordan." *Turkish Online Journal of Educational Technology*, 13 (2): 79.
- (10) Azigwe, J. B., L. Kyriakides, A. Panayiotou, and B.P.M. Creemers. 2016. "The impact of effective teaching characteristics in promoting student achievement in Ghana." *International Journal of Educational Development*, 51: 51-61. doi:10.1016/j.ijedudev.2016.07.004.
- (11) Barton, R., and T. Haydn. 2006. "Trainee teachers' views on what helps them to use information and communication technology effectively in their subject teaching." *Journal of Computer Assisted Learning*, 22 (4): 257-272. doi:10.1111/j.1365-2729.2006.00175.x
- (12) Bear, G. G., C. Gaskins, J. Blank, and F.F. Chen. 2011. "Delaware School Climate Survey-Student: Its factor structure, concurrent validity, and reliability." *Journal of School Psychology*, 49 (2): 157-174. doi:10.1016/j.jsp.2011.01.001.
- (13) Beran, T., and C. Violato. 2009. "Student Ratings of Teaching Effectiveness: Student Engagement and Course Characteristics." *Canadian Journal of Higher Education*, 39 (1): 1.
- (14) Bernardo, A. B. I., A.A. Limjap, M.S. Prudente, and L.S. Roleda. 2008. "Students' perceptions of science classes in the Philippines." *Asia Pacific Education Review*, 9 (3), 285. doi:10.1007%202FBF03026717.
- (15) Beswick, K. 2005. "The Beliefs/Practice Connection in Broadly Defined Contexts." *Mathematics Education Research Journal*, 17 (2): 39. doi:10.1007%202FBF03217415.
- (16) Boardman, A., and A. Woodruff. 2004. "Teacher change and "high-stakes" assessment: what happens to professional development?" *Teaching and Teacher Education*, 20 (6): 545-557. doi:10.1016/j.tate.2004.06.001.
- (17) Bonney, E. A., D.F. Amoah, S.A. Micah, C. Ahiameny, and M.B. Lemaire. 2015. "The relationship between the quality of teachers' and students' academic performance in the STMA Junior High Schools of the Western Region of Ghana." *Journal of Education and Practice*, 6 (24): 139-150.
- (18) Brown, E. L., J. Suh, S.A. Parsons, A.K. Parker, and E.M. Ramirez. 2015. "Documenting teacher candidates' professional growth through performance evaluation." *Journal of Research in Education*, 25 (1): 35-47.
- (19) Brown, G. T. L., H. Chaudhry, and R. Dhamija. 2015. "The impact of an assessment policy upon teachers' self-reported assessment beliefs and practices: A quasi- experimental study of



- Indian teachers in private schools." *International Journal of Educational Research*, 71: 50-64. doi:10.1016/j.ijer.2015.03.001.
- (20) Byrnes, D., G. Kiger, and Z. Shechtman. 2003. "Evaluating the use of group interviews to select students into teacher-education programs." *Journal of Teacher Education*, 54 (2): 163-172. doi:10.1177/0022487102250310.
- (21) Casas, J. A., R. Ortega-Ruiz and R. Del Rey. 2015. "Bullying: The impact of teacher management and trait emotional intelligence." *British Journal of Educational Psychology*, 85 (3): 407-423. doi:10.1111/bjep.12082.
- (22) Den Brok, P., J. Levy, R. Rodriguez, and T. Wubbels. 2002. "Perceptions of Asian-American and Hispanic-American teachers and their students on teacher interpersonal communication style." *Teaching and Teacher Education*, 18 (4): 447-467. doi:10.1016/S0742-051X(02)00009-4.
- (23) Driscoll, J., and D. Cadden. 2010. "Student evaluation instruments: The interactive impact of course requirement, student level, department and anticipated grade." *American Journal of Business Education*, 3 (5): 21. doi:10.19030/ajbe.v3i5.424.
- (24) Emesini, N. O. 2015. "Pattern of acquisition of ICT-based skills by student-teachers: Implications for teacher education in Nigeria in this era of digitalization." *Journal of Education and Practice*, 6 (33): 81-88.
- (25) Fan, F. A. 2012. "Teacher students' interpersonal relationships and students' academic achievements in social studies." *Teachers and Teaching*, 18 (4): 483-490. doi:10.1080/13540602.2012.696048.
- (26) Fenzel, L. M., and J. Domingues. 2009. "Educating urban African American children placed at risk: A comparison of two types of Catholic middle schools." *Catholic Education: A Journal of Inquiry and Practice*, 13 (1): 30.
- (27) Flowers, C., and R. Algozzine. 2000. "Development and validation of scores on the basic technology competencies for educators inventory." *Educational and Psychological Measurement*, 60 (3): 411-418. doi:10.1177/00131640021970628.
- (28) Gitomer, D., C. Bell, Y. Qi, D. Mccaffrey, B.K. Hamre, and R.C. Pianta. 2014. "The instructional challenge in improving teaching quality: lessons from a classroom observation protocol." *Teachers College Record*, 116 (6): 060304.
- (29) Goh, J. W. P., O.K. Lee, and H. Salleh. 2010. "Self-rating and respondent anonymity." *Educational Research*, 52 (3): 229-245. doi:10.1080/00131881.2010.504060.
- (30) Good, T. L., and A.L. Lavigne. 2015. "Rating teachers cheaper, faster, and better: not so fast." *Journal of Teacher Education*, 66 (3): 288-293. doi:10.1177/0022487115574292.
- (31) Hailaya, W., S. Alagumalai, and F. Ben. 2014. « Examining the utility of Assessment Literacy Inventory and its portability to education systems in the Asia Pacific region." *Australian Journal of Education*, 58 (3): 297-317. doi:10.1177/0004944114542984.
- (32) Hargreaves, E. 2014. "The practice of promoting primary students' autonomy: examples of teacher feedback." *Educational Research*, 56 (3): 295-309. doi:10.1080/00131881.2014.934554.
- (33) Haydn, T. A., and R. Barton. 2007. "Common needs and different agendas: How trainee teachers make progress in their ability to use ICT in subject teaching. Some lessons from the UK." *Computers & Education*, 49 (4): 1018-1036. doi:10.1016/j.compedu.2005.12.006.
- (34) Hill, L. 2004. "Changing minds: Developmental education for conceptual change." *Journal of Adult Development*, 11 (1): 29-40. doi:10.1023/B:JADE.0000012525.69639.5d.
- (35) Huang, F. F., C.C. Wu, C.Y. Hu, and S.S. Yang. 2006. "Teacher over involvement and student depression among junior high school students in Taiwan." *The scientific world journal*, 6: 834-846. doi:10.1100/tsw.2006.152.
- (36) Jacobs, K., E. Struyf, and S. De Maeyer. 2013. "The Socio-Emotional Guidance Questionnaire (SEG-Q): Construct validity and invariance across teacher groups." *Journal of Psychoeducational Assessment*, 31 (6): 538-553. doi:10.1177/0734282913480469.
- (37) Kang, H., and C.W. Anderson. 2015. "Supporting preservice science teachers' ability to attend and respond to student thinking by design." *Science Education*, 99 (5): 863-895. doi:10.1002/sc.21182.

- (38) Kay, R. H., and L. Knaack. 2008. "A multi-component model for assessing learning objects: The learning object evaluation metric (LOEM)." *Australasian Journal of Educational Technology*, 24 (5): 574-591. doi:[10.14742/ajet.1192](https://doi.org/10.14742/ajet.1192).
- (39) Kiany, G. R., and P. Shayestefar. 2011. "High school students' perceptions of EFL teacher control orientations and their English academic achievement." *British Journal of Educational Psychology*, 81 (3): 491-508. doi:[10.1348/000709910X522177](https://doi.org/10.1348/000709910X522177).
- (40) Klug, J., N. Krause, B. Schober, M. Finsterwald, and C. Spiel. 2014. "How do teachers promote their students' lifelong learning in class? Development and first application of the LLL Interview." *Teaching and Teacher Education*, 37: 119-129. doi:[10.1016/j.tate.2013.09.004](https://doi.org/10.1016/j.tate.2013.09.004).
- (41) Kosir, K., and S. Tement. 2014. "Teacher-student relationship and academic achievement: a cross-lagged longitudinal study on three different age groups." *European Journal of Psychology of Education*, 29 (3): 409-428. doi:[10.1007/s10212-013-0205-2](https://doi.org/10.1007/s10212-013-0205-2).
- (42) Kunter, M., J. Baumert, and O. Koeller. 2007. "Effective classroom management and the development of subject-related interest." *Learning and Instruction*, 17 (5): 494-509. doi:[10.1016/j.learninstruc.2007.09.002](https://doi.org/10.1016/j.learninstruc.2007.09.002).
- (43) Kyriakides, L. 2005. "Drawing from teacher effectiveness research and research into teacher interpersonal behaviour to establish a teacher evaluation system: A study on the use of student ratings to evaluate teacher behaviour." *Journal of Classroom Interaction*, 40 (2): 44. doi:[10.1016/j.tate.2008.06.001](https://doi.org/10.1016/j.tate.2008.06.001).
- (44) Kyriakides, L., B.P.M. Creemers, B. P. M., and P. Antoniou. 2009. "Teacher behaviour and student outcomes: Suggestions for research on teacher training and professional development." *Teaching and Teacher Education*, 25 (1): 12-23. doi:[10.1080/02619768.2014.882311](https://doi.org/10.1080/02619768.2014.882311).
- (45) Kyriakides, L., B.P.M. Creemers, A. Panayiotou, G. Vanlaar, M. Pfeifer, G. Cankar, et al. 2014. "Using student ratings to measure quality of teaching in six European countries." *European Journal of Teacher Education*, 37 (2): 125-143. doi:[10.1080/02619768.2014.882311](https://doi.org/10.1080/02619768.2014.882311).
- (46) Lang, Q., A. Wong, and B. Fraser. 2005. "Student perceptions of chemistry laboratory learning environments, student-teacher interactions and attitudes in secondary school gifted education classes in Singapore." *Research in Science Education*, 35 (2-3): 299-321. doi:[10.1007/s11165-005-0093-9](https://doi.org/10.1007/s11165-005-0093-9).
- (47) Larose, F., V. Grenon, M. Morin, and A. Hasni. 2009. "The impact of pre-service field training sessions on the probability of future teachers using ICT in school." *European Journal of Teacher Education*, 32 (3): 289-303. doi:[10.1080/02619760903006144](https://doi.org/10.1080/02619760903006144).
- (48) Martin, P. A., D. Daley, J. Hutchings, K. Jones, C. Eames, and C.J. Whitaker. 2010. "The Teacher-Student Observation Tool (T-POT) development and testing of a new classroom observation measure." *School Psychology International*, 31 (3): 229-249. doi:[10.1177/0143034310362040](https://doi.org/10.1177/0143034310362040).
- (49) Martinez, A., S.D. McMahon, C. Coker, and C.B. Keys. 2016. "Teacher behavioral practices: relations to student risk behaviors, learning barriers, and school climate." *Psychology in the Schools*, 53 (8): 817-830. doi:[10.1002/pits.21946](https://doi.org/10.1002/pits.21946).
- (50) Maulana, R., M. Opendakker, P. den Brok and R.J. Bosker. 2012. "Teacher-student interpersonal behavior in secondary mathematics classes in Indonesia." *International Journal of Science and Mathematics Education*, 10 (1): 21-47. doi:[10.1007/s10763-011-9276-1](https://doi.org/10.1007/s10763-011-9276-1).
- (51) Maulana, R., M. Helms-Lorenz, and W. van de Grift. 2015. "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. doi:[10.1080/09243453.2014.939198](https://doi.org/10.1080/09243453.2014.939198).
- (52) Mehta, V., and D.M. Hull. 2013. "Structural validity of the professional development profile of the LoTi Digital-Age Survey." *Journal of Psychoeducational Assessment*, 31 (1): 61-71. doi:[10.1177/0734282912454992](https://doi.org/10.1177/0734282912454992).
- (53) Murray, C., and K. Zvoch. 2011a. "The inventory of teacher-student relationships: factor structure, reliability, and validity among African American youth in low-income urban schools." *Journal of Early Adolescence*, 31 (4): 493-525. doi:[10.1177/0272431610366250](https://doi.org/10.1177/0272431610366250).
- (54) Nelson, P. M., J.A. Demers, and T.J. Christ. 2014. "The Responsive Environmental Assessment for Classroom Teaching (REACT): The dimensionality of student perceptions of the instructional environment." *School Psychology Quarterly*, 29 (2): 182. doi:[10.1037/spq000049](https://doi.org/10.1037/spq000049).

- (55) Nunnery, J. A., S.M. Ross, and L. Bol. 2008. "The construct validity of teachers' perceptions of change in schools implementing comprehensive school reform models." *Journal of Educational Research & Policy Studies*, 8 (1): 67.
- (56) Passini, S., L. Molinari, and G. Speltini. 2015. « A validation of the Questionnaire on Teacher Interaction in Italian secondary school students: the effect of positive relations on motivation and academic achievement." *Social Psychology of Education*, 18 (3): 547-559. doi:10.1007/s11218-015-9300-3.
- (57) Peters, S. J., and J.C. Gates. 2010. "The Teacher Observation Form: Revisions and Updates." *Gifted Child Quarterly*, 54 (3): 179-188. doi:10.1177/0016986210369258.
- (58) Pontefract, C., and F. Hardman. (2005). "The discourse of classroom interaction in Kenyan primary schools." *Comparative Education*, 41 (1): 87-106. doi:10.1080/03050060500073264.
- (59) Reddy, L. A., C.M. Dudek, G.A. Fabiano, and S. Peters. 2015. "Measuring teacher self-report on classroom practices: Construct validity and reliability of the Classroom Strategies Scale - Teacher Form." *School Psychology Quarterly*, 30 (4): 513-533. doi:10.1037/spq0000043.
- (60) Reese, L., B. Jensen, and D. Ramirez. 2014. "Emotionally Supportive Classroom Contexts for Young Latino Children in Rural California." *Elementary School Journal*, 114 (4): 501-526. doi:10.1086/675636.
- (61) Ryan, R. G., J.H. Wilson, and J.L. Pugh. 2011. "Psychometric characteristics of the professor-student rapport scale." *Teaching of Psychology*, 38 (3): 135-141. doi:10.1177/0098628311411894.
- (62) Sach, E. 2012. "Teachers and testing: an investigation into teachers' perceptions of formative assessment." *Educational Studies*, 38 (3): 261-276. doi:10.1080/03055698.2011.598684.
- (63) Sakiz, G. 2012. "Perceived instructor affective support in relation to academic emotions and motivation in college." *Educational Psychology*, 32 (1): 63-79. doi:10.1080/01443410.2011.625611.
- (64) Schroeder, S., T. Richter, N. McElvany, A. Hachfeld, J. Baumert, W. Schnotz, H. Horz, M. Ullrich. 2011. „Teachers' beliefs, instructional behaviors, and students' engagement in learning from texts with instructional pictures." *Learning and Instruction*, 21 (3): 403-415. doi:10.1016/j.learninstruc.2010.06.001.
- (65) Scrimin, S., L. Mason, and U. Moscardino. 2014. "School-related stress and cognitive performance: A mood-induction study." *Contemporary Educational Psychology*, 39 (4): 359-368. doi:10.1016/j.cedpsych.2014.09.002.
- (66) She, H., and D. Fisher. 2000. "The development of a questionnaire to describe science teacher communication behavior in Taiwan and Australia." *Science Education*, 84 (6): 706-726. doi:10.1002/1098-237X.
- (67) Siddall, J., E.S. Huebner, and X. Jiang. 2013. "A prospective study of differential sources of school-related social support and adolescent global life satisfaction." *American Journal of Orthopsychiatry*, 83 (1): 107-114. doi:10.1111/ajop.12006.
- (68) Skinner, E., C. Furrer, G. Marchand, and T. Kindermann. 2008. "Engagement and disaffection in the classroom: Part of a larger motivational dynamic?" *Journal of Educational Psychology*, 100: 765-781. doi:10.1037/a0012840.
- (69) Strong, M., J. Gargani, and O. Hacifazlioglu. 2011. "Do we know a successful teacher when we see one? Experiments in the identification of effective teachers." *Journal of Teacher Education*, 62 (4): 367-382. doi:10.1177/002248711039022.
- (70) Thijs, J., and L. Eilbracht. 2012. "Teachers' perceptions of parent-teacher alliance and student-teacher relational conflict: Examining the role of ethnic differences and "disruptive" behavior." *Psychology in the Schools*, 49 (8): 794-808. doi:10.1002/pits.21635.
- (71) Veldman, I., J. van Tartwijk, M. Brekelmans, and T. Wubbels. 2013. "Job satisfaction and teacher-student relationships across the teaching career: Four case studies." *Teaching and Teacher Education*, 32: 55-65. doi:10.1016/j.tate.2013.01.005.
- (72) Wallace, T. L., F. Ye, and V. Chhuon. 2012. "Subdimensions of Adolescent Belonging in High School." *Applied Developmental Science*, 16 (3): 122-139. doi:10.1080/10888691.2012.695256.

- (73) Walsh, S. D., Y. Harel-Fisch, and H. Fogel-Grinvald. 2010. "Parents, teachers and peer relations as predictors of risk behaviors and mental well-being among immigrant and Israeli born adolescents." *Social Science & Medicine*, 70 (7): 976-984. doi:10.1016/j.socscimed.2009.12.010.
- (74) Zullig, K. J., R. Collins, N. Ghani, J.M. Patton, E.S. Huebner, and J. Ajamie. 2014. "Psychometric support of the School Climate Measure in a large, diverse sample of adolescents: A replication and extension." *Journal of School Health*, 84 (2): 82-90. doi:10.1111/josh.12124.
- (75) Zurlo, M. C., D. Pes, and R. Capasso. 2013. "Teacher Stress Questionnaire: Validity and reliability study in Italy." *Psychological Reports*, 113 (2): 490-517. doi:10.2466/03.16.PR0.113x23z9.

#### *Combination Lesson activities and Post-lesson activities*

- (76) Castillo, R., P. Fernandez-Berrocal, and M.A. Brackett. 2013. "Enhancing teacher effectiveness in Spain: A pilot study of the ruler approach to social and emotional learning." *Journal of Education and Training Studies*, 1 (2): 263. doi:10.11114/jets.v1i2.203.
- (77) Reddy, L. A., G. Fabiano, C.M. Dudek, and L. Hsu. 2013. "Development and construct validity of the Classroom Strategies Scale-Observer Form." *School Psychology Quarterly*, 28: 317-341. doi:10.1037/spq0000043.

#### *Post-lesson activities*

- (78) Ang, R. 2005. "Development and validation of the teacher-student relationship inventory using exploratory and confirmatory factor analysis." *Journal of Experimental Education*, 74 (1): 55-73. doi:10.3200/JEXE.74.1.55-74.
- (79) Cengiz, C., and F. Karatas. 2015. "Examining the effects of reflective journals on pre-service science teachers' general chemistry laboratory achievement." *Australian Journal of Teacher Education*, 40 (10): 125-146. doi:10.14221/ajte.2015v40n10.8.
- (80) Daley, D., L. Renyard, and E. Sonuga-Barke. 2005. "Teachers' emotional expression about disruptive boys." *British Journal of Educational Psychology*, 75: 25-35. doi:10.1348/000709904x22269.
- (81) Delvaux, E., J. Vanhoof, M. Tuytens, E. Vekeman, G. Devos, and P. Van Petegem. 2013. "How may teacher evaluation have an impact on professional development? A multilevel analysis." *Teaching and Teacher Education*, 36: 1-11. doi:10.1016/j.tate.2013.06.011.
- (82) Dimitrova, R., L. Ferrer-Wreder, and M.R. Galanti. 2016. "Pedagogical and social climate in school questionnaire: factorial validity and reliability of the teacher version." *Journal of Psychoeducational Assessment*, 34 (3): 282-288. doi:10.1177/0734282915595332.
- (83) Elstad, E., K.A. Christophersen, and A. Turmo. 2012. "Exploring antecedents of organizational citizenship behaviour among teachers at Norwegian folk high schools." *Studies in Continuing Education*, 34 (2): 175- 189. doi:10.1080/0158037X.2011.611798.
- (84) Huang, S. 2006. "An assessment of science teachers' perceptions of secondary school environments in Taiwan." *International Journal of Science Education*, 28 (1): 25-44. doi:10.1080/09500690500239862.
- (85) Huang, S. L., and B.J. Fraser. 2009. "Science Teachers' Perceptions of the School Environment: Gender Differences." *Journal of Research in Science Teaching*, 46 (4): 404-420. doi:10.1002/tea.20284.
- (86) Kalk, K., P. Luik, M. Taimalu, and K. Taeht. 2014. "Validity and reliability of two instruments to measure reflection: a confirmatory study." *Trames-Journal of the Humanities and Social Sciences*, 18 (2): 121-134. doi:10.3176/tr.2014.2.02.
- (87) Landmann, M. 2013. "Development of a scale to assess the demand for specific competences in teachers after graduation from university." *European Journal of Teacher Education*, 36 (4): 413-427. doi:10.1080/02619768.2013.837046.
- (88) Neves, P. C., R. Paixao, M. Alarcao, and A. Duarte Gomes. 2014. "Organizational citizenship behavior in schools: Validation of a questionnaire." *Spanish Journal of Psychology*, 17: e17. doi:10.1017/sjp.2014.20.

- (89) Ordu, A. 2016. "The effects of diversity management on job satisfaction and individual performance of teachers." *Educational Research and Reviews*, 11 (3): 105-112. doi:[10.5897/ERR2015.2573](https://doi.org/10.5897/ERR2015.2573).
- (90) Robertson, L., and M.G. Jones, M. G. 2013. "Chinese and US Middle-School science teachers' autonomy, motivation, and instructional practices." *International Journal of Science Education*, 35 (9): 1454-1489. doi:[10.1080/09500693.2013.792439](https://doi.org/10.1080/09500693.2013.792439)
- (91) Schoeman, S., and P.L. Mabunda. 2012. "Teaching practice and the personal and socio-professional development of prospective teachers." *South African Journal of Education*, 32 (3): 240-254.
- (92) Sirin, S. R., L. Rogers-Sirin, and B.A. Collins. 2010. "A measure of cultural competence as an ethical responsibility: Quick-Racial and Ethical Sensitivity Test." *Journal of Moral Education*, 39 (1): 49-64. doi:[10.1080/03057240903528675](https://doi.org/10.1080/03057240903528675).
- (93) Vanlommel, K., J. Vanhoof, and P. Van Petegem. 2016. "Data use by teachers: the impact of motivation, decision-making style, supportive relationships and reflective capacity." *Educational Studies*, 42 (1): 36-53. doi:[10.1080/03055698.2016.1148582](https://doi.org/10.1080/03055698.2016.1148582).
- (94) Watzke, J. L. 2007. "Longitudinal research on beginning teacher development: Complexity as a challenge to concerns-based stage theory." *Teaching and Teacher Education*, 23 (1): 106-122. doi:[10.1016/j.socscimed.2009.12.010](https://doi.org/10.1016/j.socscimed.2009.12.010).
- (95) Williams, D., and L. Coles. 2007. "Evidence-based practice in teaching: an information perspective." *Journal of Documentation*, 63 (6): 812-835. doi:[10.1108/00220410710836376](https://doi.org/10.1108/00220410710836376).
- (96) Yates, S. M. 2007. "Teachers' Perceptions of Their Professional Learning Activities." *International Education Journal*, 8 (2): 213.