cogent education

TEACHER EDUCATION & DEVELOPMENT | RESEARCH ARTICLE

Measuring classroom management expertise (CME) of teachers: A video-based assessment approach and statistical results

Johannes König

Cogent Education (2015), 2: 991178

König, *Cogent Education* (2015), 2: 991178
http://dx.doi.org/10.1080/2331186X.2014.991178

cogent education

TEACHER EDUCATION & DEVELOPMENT | RESEARCH ARTICLE

Measuring classroom management expertise (CME) of teachers: A video-based assessment approach and statistical results

Johannes König[1]*

Received: 11 April 2014
Accepted: 19 November 2014
Published: 08 January 2015

*Corresponding author: Johannes König, Empirical School Research, University of Cologne, Gronewaldstr. 2, 50931 Cologne, Germany
E-mail: johannes.koenig@uni-koeln.de

Reviewing editor:
John Lee, Hong Kong Institute of Education, Hong Kong

Additional information is available at the end of the article

Abstract: The study aims at developing and exploring a novel video-based assessment that captures classroom management expertise (CME) of teachers and for which statistical results are provided. CME measurement is conceptualized by using four video clips that refer to typical classroom management situations in which teachers are heavily challenged (involving the challenges to manage transitions, instructional time, student behavior, and instructional feedback) and by applying three cognitive demands posed on respondents when responding to test items related to the video clips (accuracy of perception, holistic perception, and justification of action). Research questions are raised regarding reliability, testlet effects (related to the four video clips applied for measurement), intercorrelations of cognitive demands, and criterion-related validity of the instrument. Evidence is provided that (1) using a video-based assessment CME can be measured in a reliable way, (2) the CME total score represents a general ability that is only slightly influenced by testlet effects related to the four video clips, (3) the three cognitive demands conceptualized for the measurement of CME are highly intercorrelated, and (4) the CME measure is positively correlated with declarative-conceptual general pedagogical knowledge (medium effect size), whereas it shows only small size correlations with non-cognitive teacher variables.

ABOUT THE AUTHOR

Johannes König is a full professor of Empirical School Research, Quantitative Methods at the University of Cologne in Germany. He received the First State Examination for Teachers at Humboldt University of Berlin in 2003, a D. Phil. at Freie Universität Berlin in 2006, and Habilitation in 2011. Since 2014, he is the director of the Interdisciplinary Center for Empirical Research on Teachers and Teaching at the University of Cologne. His current research fields are teacher education research, teacher competence, and teacher knowledge (with a special focus on general pedagogical knowledge), and international comparisons. In various projects such as TEDS-M, he has worked extensively on assessing teacher knowledge and teacher education quality for the purpose of international comparisons. The novel video-based assessment presented in this paper is part of his specialization in the research on teacher knowledge and the assessment of teacher competence.

PUBLIC INTEREST STATEMENT

Classroom management is one of the most important tasks teachers have to master. Adequate mastery is clearly related to student achievement, whereas insufficient mastery may lead to teacher stress and burnout. From previous research on teacher expertise, we know specific teacher knowledge of the learning environment and their knowledge of procedures for adequate classroom management are highly relevant. This study aims at developing and exploring a novel video-based assessment that captures teachers' classroom management expertise (CME) and for which statistical results are provided. Thus, the article and its research orientation as well as empirical findings point toward a new form of assessment regarding a central issue of teachers' daily work. This form of discourse can serve as a powerful context and as an important channel for improving challenges of teaching. It supports the debates and efforts about the clarification of what teachers should learn, know, and be able to do.

Subjects: Classroom Management & Organisation; Education; Educational Research; Teaching & Learning

Keywords: classroom management; general pedagogical knowledge; teacher expertise; video clips; test; assessment

For the past decades, the interest in doing research on the measurement of cognitive elements of teacher competence has been growing, as systematic literature reviews demonstrate [e.g. Blömeke and Delaney (2012), for the knowledge of mathematics teachers and König (2014), for teachers' general pedagogical knowledge (GPK)]. One major reason for this is the understanding that knowledge is required for effective teaching, as the research on teacher expertise shows (e.g. Berliner, 1986, 1992; Bromme, 1992, 2001; Carter, Cushing, Sabers, Stein, & Berliner, 1988; Hogan, Rabinotwitz, & Craven, 2003; Klein & Hoffman, 1993; Sabers, Cushing, & Berliner, 1991). For the majority of relevant studies, however, the classical paper-and-pencil assessment represents the dominating paradigm (e.g. Baumert et al., 2010; Hill et al., 2008; König & Seifert, 2012; Tatto et al., 2012) not least because it enables an efficient and reliable way to measure declarative-conceptual knowledge in large samples. For example, in 2008, the *Teacher Education and Development Study – Learning to Teach Mathematics* (TEDS-M) was carried out under the supervision of the *International Association for the Evaluation of Educational Achievement* (IEA). TEDS-M was a comparative study of teacher education and the first IEA study on tertiary education as well as the first international large-scale assessment of future teachers that worked with representative samples (Tatto et al., 2012). The TEDS-M target population were mathematics teachers for elementary and middle schools in their final year of teacher education. More than 20,000 future teachers from 17 countries worldwide were tested using paper–pencil instruments measuring their mathematical content knowledge (MCK), mathematical pedagogical content knowledge (MPCK), and GPK. In TEDS-M, MCK covers the main mathematical areas relevant for future teachers, MPCK refers to curricular knowledge, knowledge of lesson planning, and interactive knowledge applied to teaching situations, and finally GPK is structured in a task-based way, i.e. referring to knowledge teachers need to prepare, structure, and evaluate lessons ("structure"), to motivate and support students as well as manage the classroom ("motivation/classroom management"), to deal with heterogeneous learning groups in the classroom ("adaptivity"), and to assess students ("assessment") (for more details, see König, Blömeke, Paine, Schmidt, & Hsieh, 2011; Tatto et al., 2012).

However, the measurement of context-dependent, procedural teacher knowledge goes beyond the limited scope of classical paper-and-pencil assessments (Blömeke, Gustafsson, & Shavelson, in press; Shavelson, 2010). This is especially true when looking at an action-orientated teacher skill such as effective classroom management. To account for such methodological concerns, a major current focus in the measurement of teacher knowledge and skills is the shift from paper-and-pencil tests to the implementation of instruments using video clips of classroom instruction as item prompts: such studies use videos as a stimulus in the item stem, an assessment format which is frequently referred to as "video-vignette" or "video-cued testing". Video-based assessment instruments are used to address the contextual nature and the complexity of the classroom situation. They are considered to improve the measurement of teacher knowledge when compared with the classical paper-and-pencil test (König, 2014).

Several studies (e.g. Kersting, 2008; König, Blömeke, Klein, Suhl, Busse, & Kaiser, 2014; Seidel, Blomberg, & Stürmer, 2010) have already adopted this approach to provide a more ecologically valid measurement of teacher knowledge and to intend to measure knowledge that is more of a situated nature (Putnam & Borko, 2000). The study by Kersting (2008) examines a video-analysis instrument to measure teacher knowledge of teaching mathematics. Among other findings, evidence is provided the instrument is reliable and a variety of video clips can be used to measure a homogeneous construct. The study by König et al. (2014) examined the relationship of mathematics teachers' general pedagogical skills to notice and interpret typical classroom situations from a pedagogical perspective. As expected, the skill to notice and the skill to interpret are of different quality. This shows that what a teacher knows and is able to do is of multidimensional nature, a finding that

should be accounted for by any assessment instrument measuring teacher knowledge. In the study by Seidel et al. (2010), a video-based assessment instrument was evaluated by student teachers and in-service teachers regarding its usability. Overall acceptance was good. Pre-service teachers as well as in-service teachers evaluated the classrooms situations shown in the video clips as being a helpful tool to analyze lessons, and to discuss teaching and learning. Video clips were regarded as being meaningful authentic, and interesting.

With the growing popularity of video-based measurements in the field of teacher knowledge research, it is essential to establish a convincing empirical rationale for their implementation. However, the theoretical and methodological advantage delivered by using video clips remains to be specified. To expand previous research, our study aims to address the measurement of situational knowledge in teachers by proposing a video-based approach for testing pedagogical knowledge and skill required for successfully meeting the specific requirements involved in effective classroom management.

As for example TEDS-M shows teacher knowledge about classroom management can be assigned to the broader understanding of GPK, whereas in turn GPK has been defined as one of the central cognitive components of professional teacher competence (Blömeke et al., in press; Tatto et al., 2012). Since such a conceptualization of teacher competence is based on research on teacher expertise, in the following we consider that classroom management expertise (CME) belongs to the area of GPK thus contributing to an essential component of professional teacher competence. Research on classroom management in general has triggered broad interest, as for example the corresponding handbook by Evertson and Weinstein (2006) demonstrates. Meta-analyses of empirical studies have repeatedly shown adequate mastery of classroom management is clearly related to student achievement (e.g. Hattie, 2012; Wang, Haertel, & Walberg, 1993). By contrast, insufficient mastery may lead to teacher stress and burnout (e.g. López et al., 2008). Successful classroom management depends on the teachers' ability to identify and interpret the critical aspects of the teaching learning process (Kounin, 1970). Knowledge in relation to classroom management refers to an "intellectual framework" (Doyle, 1985, p. 33), consisting of knowledge of the learning environment and procedures for adequate classroom management, which teachers have to acquire rather than an accumulation of isolated scripts and facts such as "don't smile before Christmas". The importance of CME as being part of a teacher's professional competence has been addressed by some studies on teacher competence measurement (König et al., 2011; Voss, Kunter, & Baumert, 2011). Although these studies seize classroom management as an important aspect of teachers' GPK, none of them have started to conceptualize teachers' situational knowledge of classroom management extensively. As a consequence, the measurement of classroom management has not only been limited to the paper-and-pencil approach predominantly, but it has also been kept a subordinated construct of GPK. This becomes critical when a measure of CME conceptualized as a self-contained construct is needed, for instance, for doing specific research on the effectiveness of professional development of teachers in the field of classroom management.

To establish the theoretical rationale for our study, a specification of pedagogical knowledge and skills required for successful classroom management was developed (cf. König & Lebens, 2012). Building on previous research showing expert teachers systematically perceive and interpret classroom events and sequences differently from novices, three cognitive demands that will be outlined in the following were distinguished: "accuracy of perception", "holistic perception", and "justification of action".

First, from the research on teacher expertise which has proven to be valid across different subjects and countries, it is well known that expert teachers outperform novice teachers in recalling meaningful instructional details (Klein & Hoffman, 1993; König & Lebens, 2012). Expert teachers' categorical perception with which phenomena, events, or sequences are cognitively divided into relevant units for perception (e.g. Bromme, 1992) supports them to focus on the *relation* between knowledge elements rather than on discrete elements. Repeated activation of schemata strengthens connections between elements within a schema and support enhanced activation of knowledge for categorizing new information when salient cues are present. Since connectivity and complexity of schemata

required for identifying and categorizing information evolve with practice (e.g. Dehoney, 1995), "accuracy of perception" is an indicator of expertise. Consequently, it can be reasonably assumed that expert teachers identify relevant instructional situations seen in a video-vignette assessment more precisely and correctly than do novices (Sabers et al., 1991).

Second, expert teachers can be characterized by a more "holistic perception" compared to novices (Bromme, 2001; König & Lebens, 2012): they reconstruct and anticipate the context of instruction and engage in reflecting alternative problem-solving strategies. Whereas novice teachers observe classroom situations step by step due to the fragmented structure of their knowledge, experts have an intuitive grasp of the situation since their knowledge is highly interlinked (Bromme, 1992). More specifically, prior knowledge of experts organized in schemata is employed during perception to form a cognitive representation of the situation (Putnam, 1987). By contrast, novices, whose knowledge structures for constructing a mental framework have not yet been developed, are likely to experience difficulties in reconstructing the context of instruction.

The third dimension of cognitive demands ("justification of action") refers to the functional interpretation of instructional events and sequences that depends on reasoning about the instructional intention and rationale amidst the context of classroom teacher–student interaction (Berliner, 1992). Although the functional interpretation of actions is rarely explicated in everyday teaching situations, it can be accessed from long-term memory (Bromme, 1992). In contrast to teachers' holistic perception, the interpretation of events goes beyond generating mental representations, since it strongly depends on reframing and transforming knowledge (König et al., 2014). Whereas the holistic perception can be described as a perceptive-representational process, the interpretation of events refers to transformative processes.

Besides these three cognitive demands relevant for measuring CME, a variety of typical classroom management situations are needed to assure content-related breadth of the assessment. So the video clips used for the assessment in our study refer to typical classroom management situations in which teachers are heavily challenged [following classifications provided by Hawk and Schmidt (1989), Swartz, White, Stuck, and Patterson (1990), and Doyle (2006)], involving to manage transitions, instructional time, student behavior, and instructional feedback. Although each video can be assigned to one of these situations, they also include aspects of the other situations. For example, one video clip focuses on the transition of instructional phases, a central dimension of classroom management supposed to run smoothly and effortless provided that classroom management is effective (Kounin, 1970). It first displays a class working on different group tables. Then, introduced by an acoustic signal, the teacher is displayed. She instructs the students to finish group work, to carry out a relaxation task, and announces the presentation phase of the lesson. Although the clip is very short (about 1 min), besides managing transitions, the teacher also has to manage instructional time and student behavior in that specific situation.

This study aims at developing and exploring a novel video-based assessment that captures teachers' CME conceptualized via a matrix of cognitive demands and classroom management situations. We examine the following questions:

(1) Does the assessment instrument measure the CME construct in a reliable way?
(2) When using four video clips as item prompts, do any testlet effects related to the four video clips occur?
(3) To what extent are the three cognitive demands intercorrelated?
(4) Can evidence be provided for criterion-related validity of the measurement instrument?

Examining these issues, we use the following assumptions: (1) CME can be measured using a video-based assessment in a reliable way. (2) The CME total score represents a general ability that is not or almost not biased by testlet effects related to the four video clips. (3) Due to the connectivity

of teacher knowledge and skills, the three cognitive demands conceptualized for the measurement of CME are highly intercorrelated. (4) Considering convergent validity (Campbell & Fiske, 1959), an examination of the correlation between CME and general pedagogical is of great interest. Since GPK involves "broad principles and strategies of classroom management and organization that appear to transcend subject matter" as well as knowledge about learners and learning, assessment, and educational contexts and purposes (Shulman, 1987, p. 8), CME can be regarded as a construct that is located in the field of GPK, but due to its specific definition it covers a segment of that knowledge only. Thus, we assume the CME measure is positively correlated but not identical with declarative-conceptual GPK. Moreover, CME should be positively correlated with teacher self-efficacy. Self-efficacy is used since it can be regarded as a protective factor for mental health and resilience (Schwarzer & Hallum, 2008). Following the state-of-the-art literature, it can be assumed that ineffective classroom management strategies are linked to poor self-efficacy beliefs leading to a vicious cycle of motivational and professional deficits. Therefore, besides teacher self-efficacy, also teacher burnout scales are used as a non-cognitive criterion measure. Teacher burnout has already been investigated with the challenge of classroom management, e.g. regarding the question of how to deal with disruptive behavior of students (López et al., 2008). Following this research, we also assume CME is negatively correlated with teacher burnout scales taking into account that, besides self-efficacy, it may represent another protective factor for teacher burnout. However, accounting for discriminant validity (Campbell & Fiske, 1959), we assume only small effect size correlations between our CME measure and non-cognitive teacher variables such as self-efficacy and burnout scales.

1. Method

1.1. Sample

In 2013, one elementary and two secondary schools in the greater area of Cologne, Germany, agreed to participate in our study. The whole teaching staff of each of the three schools was tested. The schools varied with respect to size from 15 teachers and 19 teachers to 85 teachers. The sample consists of 119 teachers. 90 of them (77%) are female. By average, they are 44.1 years old (SE = 1.1, SD = 11.9, min = 25.0, max = 64.0) and have taught for 16.7 years at school (SD = 12.0, min = 1.0, max = 41.0).

1.2. Data collection and procedures

The three schools that agreed to participate were asked to assembly their teaching staff for one hour. One test session per school was conducted. The survey was administered by a research assistant who was a member of the project team. Teachers first had to complete a background questionnaire containing variables such as age, sex, and teaching experience. Second, the CME instrument was administered with a total duration of 20 min allowing 5 min for watching one video clip and responding to the corresponding test items. The four video clips are very short (they vary between 1 and 2 min in length). Each video clip was presented only once, and respondents were only allowed to read test items related to a video clip when they had already watched that clip. This procedure assured that video clips were used as item prompts in a standardized way and teachers had to respond to test items immediately after having watched the correspondent clip. Test time of 20 min was appropriate since none of the teachers got bored or had to rush. Third, teachers had to complete a paper-and-pencil instrument measuring their GPK, which took another 20 min. Finally, they had to complete a questionnaire containing non-cognitive teacher variables such as teacher self-efficacy and burnout.

1.3. Measures

In this study, we investigate a new measurement instrument that captures the CME of teachers using video clips of classroom instruction as item prompts followed by paper-and-pencil test items to be responded to. As criterion measures we use a cognitive measure and non-cognitive teacher variables. As a cognitive measure, a test measuring teachers' GPK for teaching is applied, consisting of paper-and-pencil test items only. Teacher efficacy and burnout scales are used as non-cognitive teacher variables.

1.3.1. CME video-based assessment instrument

The CME measurement instrument consists of four video clips of classroom instruction that refer to typical classroom management situations in which teachers are heavily challenged. These video clips were carefully selected from a pool of video clips available to the research team. For conceptual reasons, the selection procedure applied mainly intended to follow classifications of typical classroom management situations found in the literature (Doyle, 2006; Hawk & Schmidt, 1989; Swartz et al., 1990): the video clips had to represent authentic and comprehensive situational information of classroom instruction in which a teacher is challenged (1) to manage transitions, (2) to manage instructional time, (3) to manage student behavior, and (4) to manage instructional feedback. Whole-class interaction teaching situations were preferred, as in terms of effective classroom management they are more complex and thus more challenging for teachers than private work-time situations during which a teacher assists a single student or a group of students (Kounin, 1970). The video clips had to represent a variety of classroom contexts (regarding school grade, school subject, composition of the learning group, and age of teacher), not least in order to detain respondents from getting used to one specific situational context during assessment. Besides conceptual issues, technical criteria had to be met, too. The video clips had to be of good quality both visually and acoustically, they had to represent usual events somehow familiar to every experienced school teacher, they had to be short and self-contained for research-related economic reasons. In a pilot study (König & Lebens, 2012), these criteria were issued by conducting an expert review before the procedure of selecting appropriate video clips was started. The video clips do not come along with complementary information about the teacher, the learning group or the lesson, since our idea to measure CME was to stick as closely as possible to the situation presented via video and to not distract respondents from perceiving the concrete classroom instruction.

Test items were developed for each video clip covering the three cognitive demands outlined above (accuracy of perception, holistic perception, and justification of action). In total, 27 test items were developed. Seven are multiple-choice response (MCR) and 20 are open-response (OR) items. Accuracy of perception is measured by 15, holistic perception by 8, and justification of action by 4 test items.

Coding rubrics were developed for the OR items in a complex and extensive interplay of deductive (from our theoretical framework) and inductive approaches (from empirical teacher responses). In a pilot phase, codes from several independent raters were discussed in detail and coding rubrics were carefully revised and expanded. Thus, the coding manual is theoretically based as well as data-based. The codes were intended to be low-inferent thus allowing to code every response with the least possible amount of inferences by the raters.

Coding rubrics for OR items consist of one criterion, two criteria or more than two criteria. If the single criterion is met by the response provided by the respondent, then the rater has to code this criterion with 1. If it is not met, a 0 will be given. Six test items were coded by one criterion, six test items by two criteria, and seven by more than two criteria (another two OR test items later turned out to be inappropriate for scaling analysis). Coding criteria of test items with two or more than two criteria were summed up thus having partial-credit items. For example, a test item with two coding criteria was transformed to a test item with full credit (coded with 2) and partial-credit (coded with 1). Therefore, in this case, if a response fulfilled the two criteria, a 2 was given; if a response fulfilled only one of the two criteria, a 1 was given no matter which of the two criteria was met; and if none of the two criteria were met, a 0 was given. Test items with more than two criteria to be coded received sum scores ranging from 0 to the number of criteria summed up. Theoretically, three items range from 0 to 3, one item ranges from 0 to 4, two items range from 0 to 5, and one item ranges from 0 to 7.

However, when doing frequency analysis and exploratory scaling analysis, it turned out that these differentiations were not needed, i.e. they did not substantially contribute to the improvement of item fit statistics. For example, in case of the item with a theoretical range from 0 to 7, it turned out

the empirical range was 0 to 4 only. As a consequence, partial-credit items were recoded to dichotomous items, i.e. additional categories were collapsed. For this, two different strategies were applied depending on the frequency distribution of each item: either full credit (1) was given for all responses fulfilling at least one criterion or, in case this led to a better discrimination index and frequency distribution full credit (1) was given for all responses that met two or more criteria.

All OR items measuring CME were coded on the basis of the coding manual. For the coding of the 19 OR items, we finally use in the analysis of this article (see Section 1.3.2), in total, 48 coding categories were applied. First, two raters were trained with example responses from a previously conducted small pilot study and thus learned about the principles of how to use the coding manual. When they were familiar with the coding procedure, they coded teachers' responses provided to OR test items of this study. Thirty questionnaires (25% of all questionnaires) were randomly selected and coded by the two raters who then coded the responses independently of one another. As a measure of consensus and internal consistency, Cohen's κ was estimated. For the 48 coding categories, it ranges from .29 to 1.0 with an average of M_{κ} = .80 (SD_{κ} = .21). There were only four coding categories that felt below a κ value of .5, whereas 33 of the 48 coding categories had a κ value of .7 or higher. This can be regarded as a good result (cf. Fleiss & Cohen, 1973; Landis & Koch, 1977) and thus confirms the quality of the coding rubrics. If conformity of raters was lacking, an agreement between the raters was obtained in collective discussion, calling on an expert if necessary. The four coding categories with a κ value below .5 were flagged and thus very carefully applied when all other questionnaires were processed. However, there were no subsequent coding problems with these categories.

1.3.2. Cognitive criterion measure

To account for a cognitive criterion measure in the domain of general pedagogy which includes the issue of classroom management (Grossman & Richert, 1988; Shulman, 1987), we applied the paper-and-pencil test measuring GPK of teachers that was developed in the context of TEDS-M (König et al., 2011).

The theoretical framework of GPK is structured in a task-based way and related to generic dimensions of teaching quality. Thus, four content-related dimensions of GPK are considered highly relevant allowing teachers to prepare, structure, and evaluate lessons ("structure"), to motivate and support students as well as manage the classroom ("motivation/classroom management"), to deal with heterogeneous learning groups in the classroom ("adaptivity"), and to assess students ("assessment"). Additionally, three dimensions of cognitive processes describing the cognitive demands on teachers when dealing with such generic classroom situations were defined following Anderson and Krathwohl (2001): to retrieve information from long-term memory in order to describe the classroom situation; to understand or analyze a concept, a specific term or a phenomenon outlined; and to generate strategies for how they would solve the problem posed (for more details, see König et al., 2011). Generic dimensions of teaching quality and cognitive demands made up a 4 × 3 matrix which served as a heuristic for the development of the GPK paper-and-pencil test items.

In this study, a short form of the TEDS-M GPK test as described in König et al. (2011) was applied. For this, a selection of test items was used to reduce the test length to 20 min due to data collection constraints. On the basis of TEDS-M data, test items were carefully selected according to several criteria (such as range of item difficulty, variety of item format, differentiation into content-related dimensions and cognitive processes) in order to leave the GPK overall test construct unchanged. After having selected test items, this short form of GPK test was examined using TEDS-M data. Since findings showed it was possible to create a reliable overall test score and item fit statistics computed in a one-dimensional item response theory (IRT) scaling analysis using the software *ConQuest* were good, we assume the short form of the GPK test to be a valid cognitive criterion measure for our study. When applied to the teacher sample of our study, classical item analysis was conducted over the 31 items. Internal consistency was estimated at .758, which is a good result taking into account that only about half the test items of the original instrument were included into this short form.

1.3.3. Non-cognitive criterion measures

To examine relationships between CME and non-cognitive teacher variables, two constructs that had already been subject to research in the area of teachers' challenge to manage the classroom were focused on: teacher burnout and teacher self-efficacy. Teacher self-efficacy was measured using the scale developed by Schwarzer, Schmitz and Daytner (1999) consisting of 10 items (e.g. "Even if I am disrupted while teaching, I am confident that I can maintain my composure and continue to teach well." α = .777). The Maslach burnout inventory (MBI; Maslach, Jackson, & Leiter, 1996) was used to assess the three burnout dimensions depersonalization (five items; e.g. "I feel I treat some students as if they were impersonal objects", α = .777), reduced personal accomplishment (eight items; e.g. "I have not attained important goals with my work", α = .688), and emotional exhaustion (nine items; e.g. "I feel emotionally drained from my work", α = .878). All scales were administered using a four categories response format ("not at all true", "barely true", "moderately true", and "exactly true").

2. Results

2.1. IRT analysis

IRT analyses were done with the scaling software *ConQuest* in order to carefully investigate item characteristics. Three out of 27 test items did not show satisfying psychometric statistics and thus were excluded. The final scaling model using a one-dimensional Rasch model includes 24 dichotomous test items, 19 OR items, and 5 MCR items.

The one-dimensional model and its results show it is possible to create an overall CME test score. The reliability is acceptable (EAP-reliability .699, WLE-reliability .706, Cronbach α .700) and the variance of the latent variable is sufficiently large (θ-Variance .601). Taking into account this measurement instrument is a kind of performance assessment mainly using OR items rather than a declarative knowledge test, such reliability and variance values can be regarded as good results. Besides, item fit statistics (Table 1) show that item estimation parameters spread over a range of more than four logits (from −1.778 to 3.163), which is a good result (cf. Bond & Fox, 2007), and item discrimination is .36 by average. The four items with a discrimination index below the value of .3 were kept in the scaling model for theoretical reasons, but also because they are clearly above the value of .2, which is still acceptable. Fit statistics of all 24 items are good, since the weighted mean square (MNSQ) of each item falls into the acceptable range between .94 and 1.07, and there is no statistically significant t-value ($-1.96 < t < 1.96$). Item estimate parameters were examined with regard to mean and median differences in item assignment to one of the four video clips, item format (OR vs. MCR), and item assignment to one of the three dimensions of cognitive demands (accuracy of perception, holistic perception, and justification of action). As significance tests showed, there were no statistical significant differences related to these differentiations.

2.2. Testlet effect analysis

Video-based assessments capturing situated teacher knowledge are challenged by the question whether there is a homogeneous ability across the items related to the various situations of classroom instruction presented by the video clips. Thus, the question arises whether any testlet effects may exist in our assessment approach (Sireci, Thissen, & Wainer, 1991). In our study, a testlet is defined as a cluster of items that share one video clip as a common context. If there are testlet effects, then items might measure something in common beyond the trait measured by the test as a whole.

Our hypothesis is that since CME as measured with our approach is a teacher skill widely independent from the specific subject or age group, there should be no or almost no effects evolved with the individual video clip of the assessment, because the selected situations have many aspects in common. That is, although each of the video clips brings to the front one of the typical classroom management situations in which teachers are heavily challenged (managing transitions, instructional time, student behavior, and instructional feedback), the other challenges are also involved.

Table 1. Item fit statistics from one-dimensional IRT scaling analysis

Item	Clip	Format	Cognitive demand	Estimate	SE	Weighted MNSQ	*t*-value	Discrimination Index	Facility (%)
12	3	OR	A	−1.778	.206	.99	.0	.36	89.08
24	4	OR	J	−1.527	.853	1.01	.1	.31	86.55
2	1	OR	A	−1.446	.195	1.04	.3	.24	85.59
4	1	OR	A	−1.308	.191	1.01	.1	.34	84.03
19	4	OR	A	−1.308	.191	.98	−.1	.36	84.03
16	3	OR	A	−1.239	.189	.94	−.3	.45	83.19
7	2	MCR	H	−.939	.180	1.02	.2	.35	78.99
1	1	OR	A	−.879	.179	.95	−.4	.47	77.97
9	2	OR	A	−.830	.178	1.01	.1	.35	77.31
5	1	OR	J	−.779	.176	.99	−.1	.39	76.47
6	2	OR	A	−.628	.173	1.08	.8	.26	73.95
10	3	MCR	A	−.628	.173	1.07	.7	.30	73.95
21	4	MCR	H	−.140	.164	1.00	.0	.38	64.71
8	2	OR	A	.063	.162	1.03	.4	.35	60.50
20	4	OR	H	.141	.161	.98	−.3	.39	58.82
13	3	OR	A	.374	.160	1.07	1.1	.32	53.78
23	4	MCR	H	.563	.160	.97	−.5	.43	49.58
18	4	OR	A	.986	.162	.96	−.6	.45	40.34
3	1	OR	J	1.206	.165	1.04	.5	.28	35.59
11	3	OR	H	1.228	.164	1.02	.2	.36	35.29
14	3	OR	A	1.484	.168	.96	−.5	.47	30.25
22	4	MCR	H	2.022	.180	1.02	.2	.30	21.01
15	3	OR	J	2.198	.185	.98	−.1	.36	18.49
17	4	OR	A	3.163	.215	1.00	.1	.26	8.40

Notes: OR is the open response item and MCR is the multiple-choice item.

A is the accuracy of perception, H is the holistic perception, and J is the justification of action.

Therefore, we assume that a multidimensional model specifying each video clip as a latent variable should not necessarily fit better to the data than a model that specifies only one general latent variable. To analyze this in depth, we will also specify a third model with a second-order factor reflecting CME as a general factor.

In order to analyze possible testlet effects, confirmatory factor analysis with categorical factor indicators was carried out using the software *Mplus* which in contrast to *ConQuest* provides several fit indices to compare models with different factor solutions. Analysis showed that the testlet factor model, i.e. a four-factor model in which test items of one video clip were assigned to one factor specifically, did turn out to be slightly better than a single-factor model (single factor model (model 1): χ^2/df = 1.119, p = .094, RMSEA = .032, WRMR = .930; testlet factor model (model 2): χ^2/df = 1.085, p = .172, RMSEA = .027, WRMR = .894). However, the testlet factor model with an additionally specified second-order factor (model 3) also fitted well to the data (χ^2/df = 1.084, p = .173, RMSEA = .027, WRMR = .899). Intercorrelations between the four testlet factors in model 2 are relatively high (.542/.569/.649/.731/.848/.899) and every testlet is highly (>.73) correlated at least with one other testlet factor. In model 3, CME is measured as second-order factor with factor loadings that are all greater than .7 (.706/.786/.911/.971). So the latent second-order factor can explain 83.0% of testlet 1, 94.3% of testlet 2, 61.7% of testlet 3, and 49.9% of testlet 4. To conclude, although slight testlet effects occur, we predominantly see evidence from these confirmatory factor analyses to justify modeling the CME test as a one-dimensional construct.

2.3. Intercorrelations of cognitive demands

To investigate possible intercorrelations of cognitive demands, another model specifying three latent variables, one for each cognitive demand, was analyzed (model 4). The fit of this model (χ^2/df = 1.104, p = .125, RMSEA = .030, WRMR = .915) is only slightly better than that of model 1 reported in the previous section (χ^2/df = 1.119, p = .094, RMSEA = .032, WRMR = .930). The holistic perception is more highly correlated with the accuracy of perception (.743) than with the justification of action (.447). The difference of the two correlations is statistically significant (z = 6.256, $p \leq .05$). By contrast, there are no statistically significant differences between the correlation of accuracy of perception with holistic perception (.743) and the correlation of accuracy of perception with justification of action (.775). To compare differences in height of correlations, the significance test proposed by Meng, Rosenthal, and Rubin (1992) was applied.

Again, this model 4 was analyzed with a second-order factor specified (model 5). The overall model fit remains almost the same (χ^2/df = 1.104, p = .124, RMSEA = .030, WRMR = .933). Interestingly, the three latent variables are measured with similar factor loadings for the precision of perception (.949), the holistic perception (.862), and the justification of action (.805). Since they are all greater than .8, this allows the interpretation that all three cognitive demands are significant constituents of the overall construct of CME. However, the variance explained differs from 90.1% (precision of perception) to 74.3% (holistic perception) to 64.8% (justification of action). So obviously, much of CME is related to the skill of perception.

2.4. Criterion-related validity

Criterion-related validity was examined first by correlational analysis with the TEDS-M GPK paper-and-pencil test, then by an intercorrelation analysis with the non-cognitive teacher variables teacher self-efficacy and burnout. Teachers' score in the video-based assessment of their CME was compared with their scores in the GPK paper-and-pencil test (using CME and GPK as manifest variables due to limited capacity of sample size). There is a positive, statistically significant correlation of medium size between CME and GPK (r = .470; $p \leq .001$) leading to the assumption our video-based assessment approach to measure CME is not independent from teachers' declarative-conceptual knowledge in the domain of general pedagogy as measured by the TEDS-M test. Such a correlation shows that the two constructs have something in common but are not identical. Their covariance is about 22%, but about 78% of their variance does not change together.

In another analysis, the relationship between CME and non-cognitive teacher variables was investigated. As expected, CME is positively correlated with teacher self-efficacy and negatively correlated with the three burnout scales (again all scales were used as manifest variables). However, these correlations were only of small effect size ($.1 \leq |r| < .3$) and only correlations with two of the burnout scales were statistically significant (−.232 for reduced personal accomplishment and −.283 for depersonalization, each $p \leq .01$), whereas the third burnout scale measuring emotional exhaustion (−.139, p = .122) and teacher self-efficacy (.132, p = .143) were not statistically significant. However, since the direction of these correlations support our assumption of CME as a valid construct, these correlations seem to be an important finding nevertheless.

Moreover, regarding the concept of convergent and discriminant validity (Campbell & Fiske, 1959), we see CME substantially closer correlated with GPK (medium effect size of correlation and convergent validity) than with non-cognitive teacher variables (small effect size of correlations and discriminant validity). This makes CME a valid construct being relatively close to adjacent cognitive teacher measures and being relatively distant (though not completely independently) from non-cognitive teacher measures.

3. Discussion

Classroom management constitutes a central dimension of instructional quality, whereas teachers' knowledge of classroom management is part of their professional competence. Managing the classroom exposes teachers to a range of demands requiring considerable expertise. This study

forwarded methodological consideration regarding the measurement of teachers' CME, supporting the implementation of a novel video-based assessment approach. From a theoretical perspective and in relation to the requirements of classroom management, typical situations of classroom management (managing transitions, instructional time, student behavior, and instructional feedback) and the knowledge-based processing of perceiving and interpreting classroom instruction (accuracy of perception, holistic perception, and justification of action) were conceptualized. CME was empirically investigated by administering a test instrument that consists of four video clips used as item prompts and followed by test items related to these video clips. With our research questions we asked for (1) the reliability of the instrument, (2) possible testlet effects resulting from the four video clips and their corresponding items, (3) the intercorrelations of cognitive demands, and (4) the criterion-related validity of the instrument.

Regarding our first question, findings from IRT scaling analyses show CME of teachers can be measured in a reliable way. Test items are located across a substantial ability range and generally show good item fit statistics. To investigate our second question, testlet effect analysis was done by comparing differently specified latent variable models. A testlet model specifying each video clip as a latent variable using its test items as indicators had a slightly better model fit than a model with only one latent variable using all test items as indicators. However, adding a second-order factor to the testlet model led to a nearly identical model fit which provided further evidence for the assumption of a general factor being behind the measurement of CME. This assumption is also strengthened by the finding that mean and median differences in item parameter estimates of each video clip were not statistically significant, by relatively high intercorrelations between the latent variables specifying each video clip in the testlet model, and the relatively high coefficients of the indicators of the second-order factor. With regards to our third question, cognitive demands were analyzed by comparing latent variable models, too. Accuracy of perception somehow seems to be the basis for the two other cognitive demands, since it is highly correlated with both the holistic perception and the justification of action, whereas holistic perception and justification of action show a medium size intercorrelation only (<.5). So we see our assumption confirmed that teacher knowledge and skills related to classroom management is highly interwoven, but it is also obviously necessary to differentiate teachers' holistic perception (perceptive-representational process) from their functional interpretation of classroom events (referring to processes of reframing and transforming knowledge). This further confirms our theoretical framework. Finally, when examining our fourth question, we found evidence our video-based assessment approach to measure CME correlates with teachers' declarative-conceptual knowledge in the domain of general pedagogy. This is what research in other domains has also shown. Kersting (2008, p. 857), for example, reports a statistically significant correlation of $r = .53$ between a paper-and-pencil test measuring MCK for teaching and a video-analysis instrument to measure teacher knowledge of teaching mathematics. The height of the correlation we found ($r = .47$) was very similar to that reported by Kersting (2008) for the domain of mathematics. So, there seems to be a kind of analogy between the two very different assessments, confirming the construct validity of our approach.

Besides, in TEDS-M, GPK of German future teachers at the end of their initial teacher education was associated with their MPCK [manifest correlation of $r = .30/.30$ (SE = .05/.07) for elementary/middle school teachers, respectively; cf. Blömeke and König (2010) and König and Blömeke (2010)]. Although the correlation reported here between GPK and CME ($r = .47$; SE = .07) was found for a sample of in-service teachers, comparing these correlations leads us to the assumption that GPK and CME have more in common (covariance about 22%) than GPK and PCK (9% covariance). This corresponds well to basic conceptions of how teacher knowledge is differentiated today (cf. Baumert et al., 2010; Shulman, 1987; Tatto et al., 2012): Since our GPK and CME measures are related to one of the three cognitive components of teacher knowledge researchers have identified (namely GPK vs. PCK and CK), they are more strongly interwoven compared with their correlation to one of the other cognitive components of teacher knowledge such as PCK. Accounting for this, we see our initially described consideration supported that CME belongs to the area of GPK thus contributing to an essential component of professional teacher competence.

To conclude, we consider our CME measure not only a research tool capturing a cognitive variable of teachers. Regarding the discussion in which teachers are considered to be the key professionals in the school system, especially when looking at the challenge to provide high-quality opportunities to learn for students (Hattie, 2012; Schleicher, 2011), establishing a specific CME measure may support the debates and efforts about the clarification of what teachers should learn, know, and be able to do (Darling-Hammond & Bransford, 2007). Against this background, our research instrument could be applied in various research contexts. As an educational outcome, for example, it could be used as a measure of teacher education effectiveness research and research on the effectiveness of teacher professional development. As a determinant of instructional quality and student attainment, its predictive validity could be examined. Taking into account adequate mastery of classroom management is clearly related to student achievement (e.g. Hattie, 2012; Wang et al., 1993), the hypothesis of teachers' CME measure predicting what actually happens in the classroom (e.g. captured by student ratings, video analysis of lessons, and indicators for student achievement) should be tested. That kind of research may have significant practical implications, at least for the situation in Germany, where classroom management is not only an important facet of instructional quality, but the challenges for teachers to deal with classroom management issues such as disruptive behavior of students has increased during the last decade as, for example, the PISA cycles have shown (OECD, 2013).

However, limitations of our study should be mentioned, too. First, we applied our novel approach to a sample of 119 teachers only (due to data collection constraints, since in Germany, it is very difficult to test teachers at all). Replication studies using larger samples would be necessary to strengthen our work. Second, although our instrument shows CME can be measured in a reliable way, the slight testlet effects show there is room for improvement of psychometric properties. Presumably, this could be resolved by increasing the number of video clips and test items. For example, when the number of video clips was doubled, the meaning of testlet effects, even if they slightly occur, may decrease because of the larger number of situations and aspects of classroom management involved thus leveling possible effects that come from one or the other testlet.

In future research, our instrument should be applied to samples with different expertise level as well. Close inspection of the test items in Table 1 show that there are 11 items that are relatively easy, since their correct response frequency is 70% or higher. Teachers therefore did not have so much difficulty with about half of the test items. This lets us assume it will be possible to apply the test to pre-service teachers, since it differentiates CME in the lower ability range quite well. Especially, comparisons between in-service teachers denominated as experts (e.g. by school principals) and pre-service teachers who could be regarded as novices have the potential to provide further insights into the quality of our measurement instrument. Currently, such a study is being conducted and first findings are promising, so in the near future we will be able to report further findings on measuring the CME of teachers.

Funding
The authors received no direct funding for this research.

Author details
Johannes König[1]
E-mail: johannes.koenig@uni-koeln.de
[1] Empirical School Research, University of Cologne, Gronewaldstr. 2, 50931 Cologne, Germany.

Citation information
Cite this article as: Measuring classroom management expertise (CME) of teachers: A video-based assessment approach and statistical results, J. König, *Cogent Education* (2015), 2: 991178.

References

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York, NY: Longman.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Tsai, Y-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*, 133–180. http://dx.doi.org/10.3102/0002831209345157

Berliner, D. C. (1986). In pursuit of the expert pedagogue. *Educational Researcher, 7*, 5–13. http://dx.doi.org/10.3102/0013189X015007007

Berliner, D. C. (1992). The nature of expertise in teaching. In F. K. Oser, A. Dick, & J.-L. Patry (Eds.), *Effective and*

responsible teaching (Chap. 15, pp. 227–248). San Francisco, CA: Jossey-Bass.

Blömeke, S., & Delaney, S. (2012). Assessment of teacher knowledge across countries: A review of the state of research. *ZDM, 44*, 223–247. http://dx.doi.org/10.1007/s11858-012-0429-7

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. (in press). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie.*

Blömeke, S., & König, J. (2010). Pädagogisches Wissen angehender Mathematiklehrkräfte im internationalen Vergleich [Pedagogic knowledge of prospective mathematics teachers in an international comparison]. In S. Blömeke, G. Kaiser, & R. Lehmann (Eds.), *TEDS-M 2008—Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte im internationalen Vergleich* [TEDS-M 2008—Professional competence and learning opportunities of prospective mathematics teachers in an international comparison] (pp. 270–283). Münster: Waxmann.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model. Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Bromme, R. (1992). *Der Lehrer als Experte: zur Psychologie des professionellen Wissens* [The teacher as expert: The psychology of professional knowledge]. Bern: Huber.

Bromme, R. (2001). Teacher expertise. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (pp. 15459–15465). Amsterdam: Elsevier. http://dx.doi.org/10.1016/B0-08-043076-7/02447-5

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105. http://dx.doi.org/10.1037/h0046016

Carter, K., Cushing, K., Sabers, D., Stein, P., & Berliner, D. (1988). Expert-novice differences in perceiving and processing visual classroom information. *Journal of Teacher Education, 39*, 25–31. http://dx.doi.org/10.1177/002248718803900306

Darling-Hammond, L., & Bransford, J. (Eds.). (2007). *Preparing teachers for a changing world: What teachers should learn and be able to do*. San Francisco, CA: Wiley.

Dehoney, J. (1995). Cognitive task analysis: Implications for the theory and practice of instructional design. In *Proceedings of the Annual National Convention of the Association for Educational Communications and Technology (AECT)* (pp. 113–123). ERIC Document Reproduction Service No. ED 383 294.

Doyle, W. (1985). Recent research on classroom management: Implications for teacher preparation. *Journal of Teacher Education, 36*, 31–35.

Doyle, W. (2006). Ecological approaches to classroom management. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 97–125). Mahwah, NJ: Erlbaum.

Evertson, C. M., & Weinstein, C. S. (Eds.). (2006). *Handbook of classroom management: Research, practice, and contemporary research*. Mahwah, NJ: Erlbaum.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*, 613–619. http://dx.doi.org/10.1177/001316447303300309

Grossman, P. L., & Richert, A. E. (1988). Unacknowledged knowledge growth: A re-examination of the effects of teacher education. *Teaching and Teacher Education, 4*, 53–62. http://dx.doi.org/10.1016/0742-051X(88)90024-8

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

Hawk, P. P., & Schmidt, M. W. (1989). Teacher preparation a comparison of traditional and alternative programs. *Journal of Teacher Education, 40*, 53–58. http://dx.doi.org/10.1177/002248718904000508

Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*, 430–511. http://dx.doi.org/10.1080/07370000802177235

Hogan, T., Rabinowitz, M., & Craven, III, J. A. (2003). Representation in teaching: Inferences from research of expert and novice teachers. *Educational Psychologist, 38*, 235–247. http://dx.doi.org/10.1207/S15326985EP3804_3

Kersting, N. (2008). Using video clips of mathematics classroom instruction as item prompts to measure teachers' knowledge of teaching mathematics. *Educational and Psychological Measurement, 68*, 845–861. http://dx.doi.org/10.1177/0013164407313369

Klein, G. A., & Hoffman, R. R. (1993). Seeing the invisible: Perceptual-cognitive aspects of expertise. In M. Rabinowitz (Ed.), *Cognitive science foundations of instruction* (pp. 203–226). Hillsdale, NJ: Erlbaum.

König, J. (2014). *Designing an international instrument to assess teachers' general pedagogical knowledge (GPK): Review of studies, considerations, and recommendations*. Technical paper prepared for the ITEL project. Paris: OECD.

König, J., & Blömeke, S. (2010). Pädagogisches Wissen angehender Primarstufenlehrkräfte im internationalen Vergleich [Pedagogic knowledge of prospective primary school teachers in an international comparison]. In S. Blömeke, G. Kaiser, & R. Lehmann (Eds.), *TEDS-M 2008—Professionelle Kompetenz und Lerngelegenheiten angehender Primarstufenlehrkräfte im internationalen Vergleich* [Professional competence and learning opportunities of prospective primary school teachers in an international comparison] (pp. 275–296). Münster: Waxmann.

König, J., Blömeke, S., Klein, P., Suhl, U., Busse, A., & Kaiser, G. (2014). Is teachers' general pedagogical knowledge a premise for noticing and interpreting classroom situations? A video-based assessment approach. *Teaching and Teacher Education, 38*, 76–88. http://dx.doi.org/10.1016/j.tate.2013.11.004

König, J., Blömeke, S., Paine, L., Schmidt, B., & Hsieh, F.-J. (2011). General pedagogical knowledge of future middle school teachers. On the complex ecology of teacher education in the United States, Germany, and Taiwan. *Journal of Teacher Education, 62*, 188–201. http://dx.doi.org/10.1177/0022487110388664

König, J., & Lebens, M. (2012). Classroom Management Expertise (CME) von Lehrkräften messen: Überlegungen zur Testung mithilfe von Videovignetten und erste empirische Befunde [Measuring teachers' Classroom Management Expertise (CME): On the testing via video-vignettes and first empirical findings]. *Lehrerbildung auf dem Prüfstand, 5*, 3–29.

König, J., & Seifert, A. (Eds.). (2012). *Lehramtsstudierende erwerben pädagogisches Professionswissen. Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erziehungswissenschaftlichen Lehrerausbildung* [Student teachers acquire professional pedagogical knowledge. findings from longitudinal study LEK on the effectiveness of general pedagogy in teacher education]. Münster: Waxmann.

Kounin, J. S. (1970). *Discipline and group management in classrooms*. Oxford: Holt, Rinehart & Winston.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174. http://dx.doi.org/10.2307/2529310

López, J. M. O., Santiago, M. J., Godás, A., Castro, C., Villardefrancos, E., & Ponte, D. (2008). An integrative

approach to burnout in secondary school teachers: Examining the role of student disruptive behaviour and disciplinary issues. *International Journal of Psychology & Psychological Therapy, 8*, 259–270.
Maslach, C., Jackson, S. E., & Leiter, M. P. (1996). *Maslach burnout inventory manual* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.
Meng, X.-L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111*, 172–175. http://dx.doi.org/10.1037/0033-2909.111.1.172
OECD. (2013). *PISA 2012 results: What makes a school successful (Volume IV): Resources, policies and practices*. Paris: Author.
Putnam, R. T. (1987). Structuring and adjusting content for students: A study of live and simulated tutoring of addition. *American Educational Research Journal, 24*, 13–48. http://dx.doi.org/10.3102/00028312024001013
Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher, 29*, 4–15. http://dx.doi.org/10.3102/0013189X029001004
Sabers, D. S., Cushing, K. S., & Berliner, D. C. (1991). Differences among teachers in a task characterized by simultaneity, multidimensionality, and immediacy. *American Educational Research Journal, 28*, 63–88.
Schleicher, A. (2011). *Building a high-quality teaching profession. Lessons from around the world*. Paris: OECD.
Schwarzer, R., & Hallum, S. (2008). Perceived teacher self-efficacy as a predictor of job stress and burnout: mediation analyses. *Applied Psychology, 57*, 152–171. http://dx.doi.org/10.1111/j.1464-0597.2008.00359.x
Schwarzer, R., Schmitz, G. S., & Daytner, G. T. (1999). *The teacher self-efficacy scale*. Berlin: Freie Universität.
Seidel, T., Blomberg, G., & Stürmer, K. (2010). "Observer"—Validierung eines videobasierten Instruments zur Erfassung der professionellen Wahrnehmung von Unterricht ["Observer"—Validating a video-based instrument to capture professional teaching perception]. *Zeitschrift für Pädagogik,* Beiheft *56*, 296–306.
Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Research, 57*, 1–22.
Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237–247. http://dx.doi.org/10.1111/jedm.1991.28.issue-3
Shavelson, R. J. (2010). On the measurement of competency. *Empirical Research in Vocational Education and Training, 2*, 43–65.
Swartz, C. W., White, K. P., Stuck, G. B., & Patterson, T. (1990). The factorial structure of the North Carolina teaching performance appraisal instrument. *Educational and Psychological Measurement, 50*, 175–182. http://dx.doi.org/10.1177/0013164490501021
Tatto, M. T., Schwille, J., Senk, S., Ingvarson, L., Rowley, G., Peck, R., … Reckase, M. (2012). *Policy, practice, and readiness to teach primary and secondary mathematics in 17 countries*. Findings from the IEA teacher education and development study in mathematics (TEDS-M). Retrieved April 10, 2012, from http://www.iea.nl/fileadmin/user_upload/Publications/Electronic_versions/IEA_TEDS-M.pdf
Voss, T., Kunter, M., & Baumert, J. (2011). Assessing teacher candidates' general pedagogical/psychological knowledge: Test construction and validation. *Journal of Educational Psychology, 103*, 952–969. http://dx.doi.org/10.1037/a0025125
Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational Research, 63*, 249–294. http://dx.doi.org/10.3102/00346543063003249

Cogent Education **(ISSN: 2331-186X) is published by Cogent OA, part of Taylor & Francis Group.**