Developing an Approach to the Performance-oriented Testing of Science Teachers' Action-Related Competencies

Hauke Bartels (corresponding author)

Institute of Science Education, University of Bremen, Germany, Otto-Hahn-Allee 1, 28359 Bremen, Germany, E-mail: hauke.bartels@uni-bremen.de, https://orcid.org/0000-0001-6814-7445

David Geelan

School of Education and Professional Studies, Griffith University, Parklands Drive, Southport QLD 4222, Australia, E-mail: d.geelan@griffith.edu.au, https://orcid.org/0000-0002-7455-0713

Christoph Kulgemeyer

Institute of Science Education, University of Bremen, Germany, Otto-Hahn-Allee 1, 28359 Bremen, Germany, E-mail: kulgemeyer@physik.uni-bremen.de, https://orcid.org/0000-0001-6659-8170

Short Title: Assessing Sci. Teachers' Action-rel. Competencies

Developing an Approach to the Performance-oriented Testing of Science Teachers' Action-Related Competencies

Measuring teachers' skills to carry out the complex tasks required in teaching is an important means of evaluating the effectiveness of teacher education but remains a challenging activity to conduct in practice. It is necessary to optimise approaches for usability and effectiveness along a continuum from low-effort and low-authenticity measures such as paper-and-pencil tests to high-effort, high-authenticity measures such as extended classroom observations. The first part of the paper reviews a range of efforts toward measuring the competencies of teachers and other professionals in carrying out the tasks that make up their work. These include performance tests such as computer-based simulations or simulations using actors, as well as the use of tasks requiring participation in or responses to video vignettes. Video vignette approaches typically have been less interactive than performance tests and interactivity is seen as a desirable feature. A novel framework for developing performance-oriented testing is then outlined. The second part of the paper exemplifies this framework in relation to providing explanations in physics classrooms. The development of a novel test instrument following the framework is described, and findings on construct validity are presented to support the applicability of the presented approach.

Keywords: Interactive testing, video vignettes, performance test, physics teacher education, prognostic validity, explaining

Introduction

Efforts to measure the competencies – in a European tradition usually understood as knowledge, skills and dispositions to act (e. g., Lindmeier, 2011) – of teachers in ways that credibly predict their performance is an important part of improving science teacher education (Aufschnaiter & Blömeke, 2010). Attempts to develop such measures, however,

have led to ambiguous results (see, e.g., Cochran-Smith, 2001). Part of the challenge is the tension between the desirability of authentic, interactive measures and the costs of developing, administering, and scoring such measures. Many studies focus on written tests, which are low in effort but arguably similarly low in authenticity and, more often than not, in prognostic validity regarding teachers' actual performance (e.g., Riese & Reinhold, 2010). For this reason, Aufschnaiter and Blömeke (2010) among others have called for more interactive and authentic measures.

The first part of this paper reviews two approaches that have sought to address this demand. The first is the use of 'performance tests'. Originating in medical education, these tests involve actors who are trained to respond in specific ways, giving student teachers opportunities to demonstrate their competencies in a context that simulates real interactions. The second is the use of video vignettes in educational assessment. These are typically video vignettes of practice to which beginning teachers are asked to provide verbal or written reactions. Both approaches have positive and negative features in terms of authenticity, interactivity and effectiveness (understood in terms of resources: of time to conduct, time to code/score or financial cost of administration).

In the second part of this paper, we outline a novel approach to the assessment of performance-oriented competencies that we believe retains many of the positives of both performance tests and video vignette approaches while addressing some of their negatives. This approach, we would argue, optimises the prognostic validity of the assessment while minimising the cost. We will explore this approach by reporting findings on construct validity.

Part 1: A Literature Review on Approaches and Issues in the Assessment of Teachers' Effectiveness

Investigating the 'chain of effects' between the quality of teacher education and student

learning outcomes has been a long-standing but unsatisfying issue for education research (see, e.g., Terhart, 2002 for Germany or Cochran-Smith, 2001 for the United States). Various approaches, from paper-pencil tests to videotaping classroom actions, have been adopted in order to assess teachers' professional effectiveness. A tension arises between the *test effort* (e.g., cost, impact on participants) and the *authenticity of the tasks* the teachers must solve. Following Hattie (2009), the average effect size for interventions in teacher education is d = 0.11. A power analysis shows that, with an alpha error of $\alpha = 0.05$, studies with such a small estimated effect size require a sample of $N = 505$ participants to allow the measurement of significant changes. To make claims about the effectiveness of innovations in teacher education practice, a test must be sufficiently low-effort to be administered to a large number of participants. Moreover, test authenticity is important because it reflects how closely the tasks in a test mirror real problems in the domain of teaching, influencing prognostic validity (i.e. the extent to which a score on the test accurately predicts performance in the classroom).

We follow the notion of Knievel, Lindmeier, and Heinze (2015) on teachers' *basic knowledge*, which includes both *content knowledge* and *pedagogical content knowledge* and can be divided into *action-related competencies* and *reflective competencies.* While reflective competencies consist of the abilities needed to cope with pre- and post-instructional demands, action-related competencies are required for actual teaching performance in classroom. Therefore, they are related to spontaneous and immediate interaction with learners, which means that *professional vision* (e.g., Seidel & Stürmer, 2014) and *in-the-moment decision making* (e.g., Schoenfeld, 2008) are of importance (see also Lindmeier, 2011).

Many studies have been conducted focussing on teachers' basic and professional knowledge. Among other results, these research activities revealed a lack of evidence concerning the impact of basic knowledge on the quality of teaching and, therefore, on student achievement (e.g., Baumert & Kunter, 2006, p. 469; Vogelsang & Reinhold, 2013).

Most of these studies were conducted based on paper-pencil tests, which are useful for large-scale assessments due to their low test effort, but yield ambigious results with regard to the prediction of teachers' behaviour in classrooms (e.g., Cauet, Liepertz, Borowski, & Fischer, 2015; Vogelsang & Reinhold, 2013; Lenske et al., 2016). As already described by Whitehead (1929), the existence of (declarative) knowledge and the ability to use this knowledge for solving professional tasks in a *written test* does not automatically guarantee that this ability is available in real life situations. Especially action-related competencies seem to require more than those written competencies that can be investigated with paper-pencil tests. Brouwer (2010) found that teachers under pressure are often not able to decide based on cognition, but rather act based on their earlier experiences as a student. Further, as teaching is a very complex process: many factors apart from the teacher's actual skills influence the success of a lesson (Helmke, 2007). Thus, while the raison d'être for written tests is not neglected in general, several authors have outlined a need for more innovative test formats mirroring authentic teaching situations to address action-related competencies more appropriately (e.g., Aufschnaiter & Blömeke, 2010).

A more authentic approach is the observation of teachers' real classroom practice. This has been realised in comprehensive studies that research teaching quality directly (e.g., Fischer, Labudde, Neumann, & Viiri, 2014). However, the authenticity of this test format goes hand in hand with a very high level of effort in data collection and analysis (e.g. coding videos using categories) which causes a need for relatively small samples, limiting the ability to transfer results to the total population and – depending on the exact research question – comparability between different observed classes might be very weak.

Performance tests and video vignettes are two approaches explored in the literature that go beyond paper-pencil tests in assessing action-related competencies. The literature in relation to these two forms of testing is briefly reviewed below to inform the discussion.

Performance Tests

Performance tests as a means to measure action-related competencies in standardised, authentic situations have their origin in medical education (e.g., Miller, 1990) but can be found in various domains of professional training. Different approaches are outlined briefly below to identify typical characteristics.

Clinical Training for Physicians: Objective structured clinical Examinations

In medical education, a typical approach is the use of performance tests with trained 'patient' actors. These tests imitate typical professional situations such as examining a patient with certain defined symptoms. An actor is trained to behave in a standardised way while being examined by several participants. A series of such situations is often called an *Objective Structured Clinical Examination* (Harden et al., 1979). A typical example is the training of doctors to conduct pelvic examinations (e.g., Rochelson, Baker, Mann, Monheit, & Stone, 1985). Walters, Osborn, and Raven (2005) reported that this approach allowed prediction of physicians' behaviour in real interactions with patients.

Computer-based Performance Assessment in Vocational Education

Vocational education offers examples of performance assessments using computer-based simulations. These computer-based assessments simulate professional situations from technical, commercial and nursing care areas. One extensive research project dealing with computer-based performance assessments within the area of vocational education is ASCOT (*Technology-Based Assessment of Skills and Competences in Vocational Education and Training)* by Beck, Landenberger and Oser (2016). Abele, Behrendt, Weber and Nickolaus (2016) developed a computer assessment which simulated typical technical problems with cars. Students from vocational schools for mechanics were asked to take part in the simulation and diagnose a malfunction of the simulated cars. The assessment showed

interactive depictions of parts of the cars and allowed students to measure various parameters. The authors showed that this simulation predicted the ability of mechanics to diagnose car malfunctions in reality (see also Abele, Gschwendtner & Nickolaus, 2009).

Performance Tests in Teacher Education

Teacher Performance Assessments

Teacher Performance Assessments have been used frequently in teacher education programmes in order to assess candidates' abilities to teach. Especially in the United States, assessments in teacher education programmes shifted from written tests to performance assessments which are today, despite some criticism (e.g., Croneberg, Harrison, Korson, Jones, Murray-Everett, Parrish & Johnston-Parsons, 2016), implemented in most teacher education curricula in the United States. Typical tests cover the rating by teacher educators of participants' preparation for, performance in and reflection on their teaching after teaching sessions in classroom, e.g., the *Fresno Assessment of Student Teachers* (Torgerson, Macy, Beare & Tanner, 2009) or the *Profile for Evaluation of Interns* (Brown, Suh, Parsons, Parker and Ramirez, 2015) which means that external observations are necessary. Most of these assessments focus on general issues of teaching, applied to participants' individual subjects. One of the most influential performance assessments in teacher education is the PACT/edTPA, which is presented more in detail in the following.

The *Performance Assessment for California Teachers* (PACT) is a test of the action-related competencies of teacher candidates. It consists of formative assessments conducted during teacher education seminars as well as a summative assessment at the end of the programme. The assessments consist of *teaching events*: small teaching units planned, conducted, documented, videotaped and reflected upon by the teacher candidate. These different sources form a portfolio, which is rated by teacher educators and leads to a subject-

specific benchmark. The rating of one *teaching event* takes about 2-3 hours (Pecheone & Chung, 2006). The *Educative Teacher Performance Assessment* (edTPA), a performance assessment based on PACT, consists of 27 subject-specific assessments and is applied in teacher education programmes in 40 states throughout the United States (Stanford Center for Assessment, Learning and Equity, 2013). Newton (2010) reports that a teacher's score on the edTPA allows prediction of that teacher's classroom effectiveness measured via the achievements of their students.

Skills of Science Teachers in Explaining

While most performance assessments in teacher education focus on more general aspects of teaching, the *Dialogic Explaining Assessment* (DEA) developed by Kulgemeyer and Tomczyszyn (2015, see also Kulgemeyer & Riese, 2018) is designed to measure the explaining skills of participants in undergraduate physics teacher education programmes. It, therefore, focuses on a very specific teaching situation: explaining. As this test is of special importance for the new kind of assessment presented below, it will be described in some detail. The DEA consists of face-to-face explaining situations where single participants are required to explain a given physics situation to a high school student (e.g. 'Why does one feel weightless in a roller-coaster?' or 'How can the earth be protected against an approaching asteroid?'). Limiting the interaction to a one-on-one dialogue allows the avoidance of non-construct relevant stimuli, which would appear in a context involving a whole learning group (e.g., classroom management related issues). It focuses on teacher-student dialogues and, therefore, not on teachers explaining to a whole class. The high school student in each trial is a 'trained explainee' (like the actors who play patients in medical assessments) and answers with standardised prompts that require the participant to modify the approach to explaining, e.g., concerning the level of mathematics used. The participant is allowed to use standard tools available in a classroom to support the explaining activity. Pictures or drawings from a

given set of materials can be used as well as a whiteboard for writing or calculations. To ensure comparability, a standardised time limit of 10 minutes to prepare and another 10 minutes to provide the explanation is given. The explaining dialogues are videotaped and coded.

Based on both theory on the quality of explaining skills (e.g., Brown, 2006; Wittwer & Renkl, 2008) and an inductive approach, a set of 42 categories for quality of explaining was identified, which allowed the description of all observable actions in the videos. With respect to reliability, this set was reduced to 12 major aspects, which Kulgemeyer and Riese (2018) describe as follows (see Table 1).

Category	**Description**
Presenting concrete numbers for formulas	Explainer presents numbers as an example instead of leaving a formula unexplained.
Explaining physics concepts in everyday language	Explainer avoids technical terms by describing the underlying concept with everyday terms.
Connecting non-verbal elements	Explainer connects non-verbal elements like diagrams, pictures or demonstrations by highlighting similarities and differences.
Using items in general	Explainer uses small everyday items (e.g., a paper snarl) to illustrate a process.
Connecting items with the topic by showing analogy	Explainer not only uses small everyday items but connects them to the topic s/he wants to explain (“The paper snarl stands for the asteroid. Look at it when I am moving it”)
Small demonstrations	The explainer conducts small demonstrations with everyday items.
Answering inadequately (negative category)	Explainer does not answer an addressee’s question or ignores the question.
Review	The explainer stresses that something has already been explained and is needed now (“You remember that we were talking about friction before, that is exactly what happens here.”)
Summary	The explainer summarizes the explanation briefly.
Encouragement	The explainer praises the explainee for good answers and encourages to deal with difficult parts of the explanation.
Diagnosing understanding	The explainer diagnoses the success of the explanation by asking questions or giving

	tasks (NOT just: “Did you understand that?”)
Request action from explainee	The explainer requests the explainee to act. (“What do you think how it moves? Could you sketch that for me?”)

Table 1: Categories of the DEA (Kulgemeyer & Riese, 2018)

In a sample of 109 participants from physics teacher education programmes from five German universities these aspects formed a measure of explaining skills with sufficient reliability (Cronbach’s $\alpha = 0.772$).

Concerning objectivity, inter-rater reliability was examined. The accordance between two raters ranged from 73% to 97%, depending on the category. Consensus between the raters was found for all categories in a second step.

With respect to validity, the authors found that the measure sufficiently predicted experts’ decisions for the better explaining quality in pairwise comparisons of the videotaped explanations (Cohen’s $\kappa = 0.78$). To ensure content validity, all four variables describing actions taken to increase the quality of explaining (adaption of (a) language level, (b) examples, (c) mathematizations, and (d) representation forms) are covered by the categories. Participants rated the test situation (including the trained addressees) as authentic. A nomological network was built as an overall indicator of construct validity.

The DEA was used to examine the relationship between teacher students’ explaining skills and their pedagogical content knowledge, progress in a teacher education programme and beliefs about explaining. The authors report medium correlations of explaining skills with pedagogical content knowledge ($r = 0.378$; $p < 0.001$), participants’ progress in an academic teacher education programme ($r = 0.482$; $p < 0.01$) and a negative medium correlation with a transmissive view on explaining ($r = -0.339$; $p < 0.001$). These findings meet theoretical considerations and are, therefore, considered as a further hint for validity.

Characteristics of Performance Tests

Taking into consideration these very diverse approaches, three aspects of performance tests turn out to be important. First, a performance test always relies on the high *authenticity* of the professional action undertaken. All the situations that are simulated are designed to be as realistic as possible, trying to imitate the real professional situation. Second, real *interaction* between the participant and the test situation takes place. The participant can manipulate the test situation, e.g., by moving the aeroplane or asking the patient about prior diseases. Third, these situations model *important aspects of the daily professional routine* and therefore simulate an integral part of the participants' field of activity.

It is worth noting a difference here between performance assessments in relation to the physical sciences versus those in human and social activities. In a flight simulator, because flight relies on physics, a given action will have a determined result. Moving the rudder to the right will cause the plane to turn to the right. Human contexts such as teaching are much more complex and unpredictable. Medical performances include both the science of diagnosis and the social elements of bedside manner and patient interaction. Judging performances in highly predictable contexts allows correct and incorrect actions to be easily identified, but this is more challenging in complex human contexts.

All the approaches presented have in common the fact that they reduce the 'degrees of freedom' of a given situation in order to allow comparability between candidates and to make it easier to handle setup. However, they still rely on an effortful human-based procedure of judgment, which makes data analysis time-consuming. The time-consuming and expensive nature of performance assessments is not limited to their judgement/assessment phase: the DEA (Kulgemeyer and Tomczyszyn, 2015) for example, requires extensive training of the addressees to ensure standardised behaviour. For flight simulators, comprehensive technical effort is needed, and considerable expense is involved in designing, building and operating a

simulator. This means that the sample size for this kind of assessment is always limited and large-scale research is only possible when very significant resources are available.

Video Vignette Tests

In empirical social research, vignettes are used as input to make participants of a test empathetic toward a given situation (Steiner & Atzmüller, 2006). That is, participants experience a short input (e.g., a piece of video or written transcript) and are then questioned or surveyed about their responses to the situation depicted in the vignette. While this has been a common approach in sociology, research in education only began to focus on this design in the 1990s (Brovelli, Bölsterli, Rehm & Wilhelm, 2013). The presented situations may either be taken from reality or artificially designed. They introduce a scenario and end at a defined, often critical, situation. After watching/reading the vignette, the participant is asked to judge aspects of the given situation or to provide ideas about how the situation should be continued (Schratz, Schwarz, Westfall-Greiter, & Rumpf, 2012). This approach has been described as a connection of survey and experiment (Streit & Weber, 2013) and has become popular in recent research projects on teacher education. Especially, video vignettes have been described as a promising approach to test action-related competencies (e.g., Kersting, 2008; Neuweg, 2015), and have been reported to be well-accepted by teachers, student teachers and teacher instructors (Seidel & Prenzel, 2007). Several approaches to collecting and analysing the actual test data are outlined here, from verbal and video responses to written answers or Likert/multiple-choice items.

Tests with Verbal or Videorecorded Answers

Video vignette tests with verbal or videorecorded answers are either used as a starting point for discussing teaching issues or to directly measure participants' reactions to a certain prompt. Goffree and Oonk (1999) developed a video vignette test, which consisted of video

scenes recorded in primary school mathematics classes. The aim of the project was to help students in an undergraduate teacher education programme improve their teaching by discussing other teachers' lessons. The videos were presented to small groups of participants, who were invited to identify problematic actions on the part of the teacher in the vignettes. Participants could stop the videos whenever they wanted to and discuss alternative and better ways to handle the shown situation. All discussions were filmed and evaluated with qualitative tools. A similar approach for the training of early childhood educators was taken by Lerner and Parlakian (2007). For the assessment of in-service teachers, Lindmeier, Heinze and Reiss (2013) developed a test to assess mathematics teachers' abilities to react spontaneously to student prompts. Video vignettes from high school mathematics lessons were presented, showing students discussing mathematical issues. The vignettes stopped at given points, and the participants were asked to verbally react to the shown situation. The task was, for example, to help a student overcome a misconception. To simulate the pressure to act in the moment that is present in real classroom situations, participants were given limited time to answer. Each attempt was recorded and analysed using a coding scheme. Knievel et al. (2015) used a video vignette test to examine primary school teachers' action-related competencies with special regard to their ability to handle primary school students' misconceptions. The video vignettes could only be seen once, and after the scene, the participant had a limited time to answer as though to the student. Transcriptions of the answers were used for category-based coding.

Tests with Written Answers

A variety of studies to examine teachers' skills were conducted using a written answer format. We would regard the following as representative examples of this type of testing. As this is not the focus of the present paper, we only give a brief overview here.

Hoth et al. (2017) assessed early career mathematics teachers in Germany by reanalysing data from the *Teacher Education and Development Study in Mathematics* (TEDS-M) longitudinal follow-up study TEDS-FU. The aim of the assessment was to measure the ability of early career teachers to identify gifted students. Scripted lessons were used to imitate actual classroom dialogues. The participants answered both Likert and written text items.

Bruckmaier et al. (2013) used video material from the *Cognitive Activation in the Classroom* study (COACTIV) for a video vignette test. After watching a video, the participants gave written answers on how they would continue the shown lesson. A similar approach was used by Kersting (2008).

Dannemann, Niebert, Affeldt and Gropengießer (2014) used video vignettes in their biology teacher education programme. The videos showed high school students in science classes talking about biology content while demonstrating misconceptions. The student teachers were asked to use the vignettes as an initial point for planning a biology lesson fitting the needs of the students shown in the vignettes.

Tests with Likert or Multiple-choice Answers

Tests with closed answers are often set up in a computer environment, allowing automatic scoring of participants' inputs. This requires an expert-based definition of right and wrong answers beforehand. Forster-Heinzer and Oser (2015) developed a video vignette test for teachers with fully standardised answers. They presented video vignettes of classroom situations to their participants, including interactions between teachers and students. A so-called *Advocatory Approach*, first used in the study from Oser, Heinzer and Salzmann (2010), was taken to allow participants to judge the competencies of the teacher shown in the vignettes using Likert items. The underlying assumption was that their sensitivity to other teachers' competencies allowed diagnosis and carried information about the participants'

competencies. These competencies were measured by calculating the accordance of the participants' judgment with the judgment of experts. A similar approach was taken by Seidel and Prenzel (2007) for teacher trainees and in-service teachers.

König and Lee (2015) used video vignettes for an assessment of teachers' Classroom Management Expertise. Instead of Likert Items, they used multiple-choice items to ask their participants about several aspects of the presented vignettes.

Application Scenarios of Video Vignette Tests

In teacher education research, video vignette tests are used as part of contextualised assessment. As outlined above, including a higher degree of actual teaching action (e.g., a verbal response) is promising with regards to the measurement of action-related competencies. Such a measurement more closely models actual teaching behaviour than does writing a reaction, since the latter always includes a retrospective reflection on the situation that is not possible under the pressure that exists in real classrooms to instantly react verbally (Knievel, et al., 2015; Aufschnaiter & Blömeke, 2010). However, rather than categorising video vignette tests by the format taken by their answers, it is more powerful to categorise them in terms of the pedagogical or research goals of those administering the tests. Three broad categories emerge under such analysis:

- Using video vignettes as a base for discussion on teaching and to foster the abilities of (future) educators in teacher education programmes (Goffree & Oonk, 1999; Lerner & Parlakian, 2007 and, in a more output-oriented approach, Dannemann et al., 2014).
- Using data generated by participants analysing/judging video vignettes as an indirect measure for their own teaching abilities (Forster-Heinzer & Oser, 2015; Hoth et al., 2017; Kersting, 2008; König & Lee, 2015; Oser et al., 2010; Seidel & Prenzel, 2007).

- Using reactions to video vignettes ("what would you do next?") as a direct measure of teaching quality (Bruckmaier et al., 2013; Knievel et al., 2015; Lindmeier et al., 2013).

One purpose of this paper is to discuss the use of video vignettes to predict teachers' behaviour in real classrooms. We will focus on the usability of test designs fitting the second and third categories, while recognising the value of video vignettes for pedagogical purposes in teacher education.

The review above reveals two promising approaches. The first is to ask the participant to react to a shown situation in a teaching-like manner directly (e.g., Knievel et al., 2015). In this case, the participant's reaction is a *direct measure* of the way in which s/he would be expected to act in a classroom. The second approach is to ask the participant to evaluate a given classroom situation or another teacher's behaviour from a reflective perspective (e.g., Oser et al., 2010). In this case, the participant's reaction is an *indirect measure*. While using direct measures has always been possible to realise, albeit at some cost, by coding recorded answers, test setups using indirect measures allow, but are not limited to, closed answer formats such as multiple-choice or Likert items. This has an impact on the effort required in data analysis. While direct measures are likely to yield higher prognostic validity than indirect, the cost of coding may make them prohibitive for some applications. Indirect measures using more easily scored written responses offer the potential to conduct larger-scale, lower-cost tests. The remainder of this paper describes a proposed instrument of this kind that we are developing.

Towards a New Kind of Performance-oriented Assessment

Can Video Vignette Tests Predict Performance?

As mentioned above, the three most important aspects of a performance test are (i) *professional relevance* of the test content for the daily routines of the profession studied, (ii) *authenticity,* and (iii) *interactivity*. If a video vignette test is intended to predict performance, these criteria should be fulfilled. In the following, we will show how the chosen kind of measure affects the ability of the test to predict performance.

The *professional relevance* of the test content reflects the choice of subject matter. Tests with both direct and indirect measures allow a focus on situations which are highly relevant for the routines of the profession. This can be seen, for example, in the studies from Hoth et al. (2017) for indirect and Knievel et al. (2015) for direct measures. Concerning *authenticity*, the chosen kind of measure has a much bigger influence. Tests with indirect measures rely on a reflective perspective. In all studies with indirect measures presented above, the participants were asked to judge or analyse a given situation (e.g., Forster-Heinzer and Oser 2015). While this is an important part of teachers' skills, it is not an authentic test situation in terms of real teaching simulation. Tests with direct measures, on the other hand, allow much more authentic situations (e.g., Bruckmaier et al. 2013). Although this is still not as authentic as established performance tests like the DEA (Kulgemeyer and Tomczyszyn, 2015), the setting allows real teaching performances and investigates real teacher actions. A main difference between the approach described in this paper and the DEA is the fact that in video vignette tests the classroom situation does not react to the participants' attempt at teaching. This leads to the issue of *interactivity*. None of the studies presented above allows real interaction with the classroom situation shown in the videos. This inherent lack of interactivity might be one of the main issues with computer-based test instruments in social

science research. They rely on a certain reality in the test situation, which can hardly be fulfilled by a computer.

In a nutshell, video vignette tests cannot be seen as actual performance tests as they differ in the degree of interactivity and authenticity. Even though many studies rely on the prediction of real behaviour using video vignette test scores, a higher degree of actual teaching action within the test might enhance the chance to cover action-related competencies in a sufficient way.

From Performance Tests to Performance-oriented Testing

The following part of this paper outlines the use of what we have described as a 'performance-oriented test'. It is not an actual performance test with 'live' students or patients as described by Miller (1990) but draws on the affordances of both video vignette tests with written answers and computer-based interactivity to address issues of professional relevance, authenticity and interactivity and to keep costs and test effort relatively low. We claim that this approach offers a suitable measure for teaching performance, addressing some of the issues identified with indirect measures. We illustrate the features of this approach in the context of a performance-oriented test of how beginning physics teachers explain physics concepts to their students. We use a well-evaluated performance test, the DEA, as a starting point to develop this novel approach. This allows us the opportunity to compare the effects of performance-oriented testing and actual performance assessment.

Part 2: Developing a Performance-oriented Video Vignette Test for Physics Teachers' Explaining Skills

Explaining skills are crucial for science teachers (Osborne & Patterson, 2011). A measure of explaining skills would be helpful to allow evidence-based decisions in relation to modifying

physics teacher education to enhance graduates' skills in relation to explaining physics concepts. This is a challenge as action-related competencies such as teachers' explaining skills cannot credibly be measured with paper-and-pencil tests (Vogelsang & Reinhold, 2013).

Theoretical Background: Science Teaching Explanations

While sometimes treated as synonymous, the concepts 'explanation' and 'argument' in science differ in various aspects (Osborne & Patterson, 2011). In brief, an explanation is intended to lead to the development of understanding on the part of the addressee (the person receiving the explanation), while an argument is intended to persuade the recipient to adopt a position.

We wish to distinguish between *scientific explanation* and *science teaching explanation* (Treagust & Harrison, 1999). A *scientific explanation* is of the kind given in a scientific paper and connects phenomena with an underlying principle relying on a logical connection. *Science teaching explanations* as described by Treagust and Harrison, on the other hand, have the intention of fostering an addressee's knowledge.

In order to operationalize the process of physics-related explaining, Kulgemeyer and Schecker (2013) proposed a *dialogic model for explaining in science communication*, which is based on empirical findings and has been used to describe both student explanations (Kulgemeyer & Schecker, 2013) and teacher explanations (Kulgemeyer & Riese, 2018; Kulgemeyer & Tomczyszyn, 2015). In contrast to a naïve 'transfer of knowledge' (transmissivist) understanding, it relies on a constructivist understanding of explaining which involves the explainer in a dynamic interaction with the addressee. It describes the situation of an explainer in the attempt to explain a science topic to one or more addressee(s). The explainer provides a communication offer to the addressee(s), which can be accepted or rejected. The explainer receives verbal or non-verbal feedback and modifies the explanation.

To increase the chance of successful explaining, the explainer must consider two main aspects: (i) what is to be explained? (the content of the scientific concept under discussion) and (ii) to whom is it to be explained? (the addressee). Following the model, four main 'variables' of explaining can be modified in order to adapt an explanation to an explainee's requirements: (1) the language code (e.g., scientific language vs. everyday language), (2) the level of mathematics, (3) the used examples/analogies, and (4) the representational forms used (e.g., photos vs. scientific diagrams).

A Performance-oriented Instrument for Measuring Explaining Skills

While some approaches to judging the quality of science teaching explanations have been made (e.g., Norris, Guilbert, Smith, Hakimelahi, & Phillips, 2005; Sevian & Gonsalves, 2008), the DEA from Kulgemeyer and Tomczyszyn (2015) described above was the first attempt to assess teachers' explaining skills. While much is known about the validity and the reliability of this attempt, the collection and analysis of data are extremely time-consuming and not applicable for large-scale assessments (Kulgemeyer & Tomczyszyn, 2015). That makes the DEA an ideal starting point for the development of a performance-oriented test instrument.

We developed a two-tier instrument that is intended as an alternative to the DEA. The aim was to find a way to measure explaining skills in a highly standardised setting, allowing easily applicable large-scale assessments while ensuring high prognostic validity regarding teachers' actual classroom performance by fulfilling the quality criteria for a performance test to the greatest degree possible.

In order to ensure maximum large-scale ability, we worked with an online test, using multiple choice items. These items were built with a two-tier design, which has been described as appropriate to measure competencies beyond declarative knowledge (see, e.g.,

Tan, Goh, Chia & Treagust, 2002; Urban-Woldron & Hopf, 2012). The test was implemented using the free test software *LimeSurvey* (LimeSurvey Project Team, 2015), broadening its applicability and lowering cost.

Each test item begins with a video vignette, showing a teacher attempting to explain a physics situation to a student. During the dialogue, the student asks a question or signals incomprehension. The video stops and the participant is asked to select the most suitable way to continue the explanation from among four given answers (Tier 1, single selection). Each of the available answers is scientifically correct. On the next slide, the participant is asked to give reasons for the selected approach (Tier 2, multiple selection). For each item, six answers are available, including both subject and addressee oriented reasons. Afterwards, the reaction of the teacher in the video is shown, which marks the beginning of the next item.

The way the teacher continues explaining represents one of the four offered possibilities, but it is not always the best possible way to continue. All video vignettes shown in the test belong to the same dialogue, and therefore the given situation continues throughout the whole test.

Two content variations of this test have been developed, both taken from the area of mechanics, which is an integral part of high school physics. In this paper, we focus on Variation I, which is about saving Earth from an approaching asteroid by breaking it into two equal sized parts, which pass Earth on two sides at the same distance.

To ensure that all participants are able to use the test software, an introduction video presents an example item. As the video vignette items only include scientifically correct possibilities in Tier 1, the participants' content knowledge is measured using six single choice items before the actual test starts. This also helps the participants to get in touch with the science subject matter that is to be explained, before they actually have to think about developing an explanation.

An example, taken from the scenario with the approaching asteroid, will help to clarify the nature of the instrument. Figure 1 shows a film still from a video vignette. It deals with a teacher called Mr Miller and a high school student named Sarah. The shown item is from the middle of the test. In the vignette before, Sarah signalled trouble in understanding the principle of superposition. Mr Miller asks her to specify the issue of misunderstanding:

Slide 1: Video vignette

Miller	*Can you tell me what part you did not understand?*
Sarah	*I'm not sure; there were so many science terms… couldn't you explain the principle of superposition in an easier way so that I can understand it?*

Figure 1: Screenshot of a video vignette

Slide 2: Tier 1 (single choice, shortened)

1	The principle of superposition describes how two forces which apply at different angles have to be summed up. The result is another force which points in the direction of the body's actual acceleration. This is exactly what happens to the parts of the asteroid.
2	Let's have a closer look at this. First, the asteroid moves straight towards earth and has a certain velocity. When it is blown up, the asteroid is separated into two parts which diagonally pass the earth on the right and the left. This is why the earth is missed, and mankind survives.

3	We just found out that the parts of the asteroid each have two parts of their velocity, pointing in different directions. As the asteroid moves towards earth, one part points towards it. The second part is added by the detonation and points sidewards. Now the principle of superposition says that these two velocities are added to give the object's velocity. Therefore the parts of the asteroid move forward diagonally.

Slide 3: Tier 2 (multiple choice, shortened)

1	Use language that fits the student's language level.
2	Use language with appropriate physics terms.
3	Use a context that might be interesting for students.

While only one option can be chosen in Tier 1, multiple selections are allowed for Tier 2. This is to consider the possibility that more than one reason might be of importance when thinking about the best explanation. In order to gather as much information as possible about the participants' thinking, participants can also enter free text to explain their decisions.

During test development, a qualitative study was conducted to explore the applicability of the test. It consisted of think-aloud studies (see e.g., Ericsson & Simon, 1993) which were conducted with 8 participants (teacher students) in order to reveal the main considerations which went on in the participants' minds. This study showed that participants spent less than 5% of the test time on construct-irrelevant considerations (e.g., technical operation of the test software or misunderstandings in the provided explanations). Several technical issues were resolved and some unclear explanations edited.

Applying the Criteria for Appropriate Performance Tests to the Instrument

In the following we will discuss how the criteria for performance tests outlined above were applied to develop the test instrument. We will discuss how we aimed to fulfil these criteria. The *relevance* of explaining skills for teaching physics in general has been discussed above. With regard to the characteristics of performance tests, two aspects remain: *authenticity* and *interactivity*.

Authenticity and Interactivity of the Test

While the closed answer format limits the ability of the test to measure real performance or provide a direct measure, we implemented several aspects to ensure the highest authenticity and interactivity of real classroom action possible for a multiple-choice test. Rehm and Bölsterli (2014) recommend using real classroom observations as a base for the video vignettes and possible items. We met this by using real videotaped teacher explanations from the DEA (Kulgemeyer and Tomczyszyn, 2015).

Typical attempts at explaining the given topic were identified and used to create a script for the video scenes. Both the actual video dialogues and the answers for the first tier were developed in this way. The first tier answers were modified slightly to ensure that each item only relied on one of the four 'variables'. This was necessary to allow the identification of clearly right and wrong answers (see below). However, this approach requires the student shown in the video to mention the reason for his/her trouble in understanding in an explicit manner. The reasons listed in the second tier for choosing a particular option in the first tier were taken from free text answers given in a pilot study, ensuring that the possible answers represent things that actually go on in participants' minds while performing the test. The video vignettes were scripted and reproduced with actors. This allowed the avoidance of non-construct-specific stimuli but ensured maximal authenticity of the dialogues (see also Knievel et al., 2015). The video vignettes, therefore, mirror interactions which occur in private tutoring situations but in classrooms, too (e.g., if a student needs assistance during learning).

Another aspect of authenticity concerned the time available to decide how to continue the explaining process. The video vignettes could only be seen once. Participants were given a limited time of three minutes to decide which of four possibilities represented the best way to continue explaining. Taking into consideration the average reading time of two minutes, only one additional minute was left to decide. While we are aware that actual teacher

reactions in classrooms are more spontaneous than this, we still aimed to realise a decision-making setup which was as authentic as possible within this standardised test.

Interactivity was aimed for by offering a row of vignettes following the same dialogue. Even if participants could not influence the actual behaviour of the teacher shown in the videos, they at least took part in a coherent conversation and followed the learning progress of the student throughout the whole assessment. While this is not a real simulation of a classroom, it still goes beyond other vignette tests, which have been criticised for their lack of interactivity (Lindmeier, 2013).

By implementing these aspects, we wanted to ensure that participants performed in a way predictive of their actual classroom explaining when responding to Tier 1 in the test items. To put it slightly differently, we sought to access Schön and DeSanctis' (1983) 'reflection-in-action' rather than 'reflection-on-action'. Because participants were required to select from among a set of given answers and the creativity of providing an individual explanation (for example by inventing an engaging example) was not available, the results cannot be seen as representing a direct measure. However, as dicussed, above, we aimed to develope the test instrument focussing on its authenticity and interactivity.

Coding the Test Responses

Given that an explanation is 'right' in terms of scientific correctness, whether a certain way to explain is better or worse still depends heavily on the needs and characteristics of the addressee – adaptation to the needs of the person receiving the explanation has been described as the most important variable for successful explaining (Wittwer & Renkl, 2008). With regard to a video vignette test, where only limited information on the addressee is available, this means especially that the behaviour of the student in the vignettes should be taken into consideration when deciding if a particular way of explaining is appropriate. For example, it is obviously not a good choice to increase the level of formal mathematics when a

student has signalled that s/he does not understand the low-level mathematics that has been mentioned before.

To find the most suitable answers within the four choices for Tier 1, we chose a theory-driven approach first, ensuring that the most appropriate answers met criteria of good explaining as discussed above. As a second step, we conducted an iterative consensus process with experts. This included feedback from two university professors of physics regarding scientific correctness and numerous discussions with experienced physics education staff. According to the suggestion of Liebold and Trinszek (2009), experts with significant practical experience in explaining were chosen. We decided to choose experts with (a) physics teaching experience who were familiar with (b) theories of instructional design and (c) the theory of explaining. We ended up with ten experts, all of them either holding a PhD in physics education or currently working as physics education researchers.

Each of the video vignettes was shown to the experts, and they were asked to decide the most appropriate way to continue. In a follow-up discussion, the experts were asked to suggest modifications so that one answer would clearly be identified as the best. These modifications were implemented and presented at the next meeting with the experts. This procedure was repeated until a clearly favourable answer had been identified for each item. A rating with actual addressees like high school students was not conducted as literature suggests that students are not able to judge explaining videos with respect to actual explaining quality but tend to rate explaining based on non-construct-specific aspects like the charisma of the explainer. Students tend to fall for the so-called 'illusion of understanding' (e.g., Chi, Bassok, Lewis, Reimann & Glaser, 1989; Chi, de Leeuw, Chiu & LaVancher, 1994): they sometimes consider explanations to be sufficient that actually do not provide them with scientifically correct knowledge. Some explanations sound convincing at first but if the knowledge needs to be applied afterwards it becomes obvious that they are not

appropriate. However, the think-aloud study mentioned above revealed that participants who performed above average followed the experts reasoning to decide on the most appropriate way to continue the explaining.

The item shown above is an example. The student signaled trouble with science terms. This focuses on the language code. Taking into consideration the student prompt, it seems best to avoid using too many scientific terms, which means Answer 1 is not appropriate. Answer 2 comes with the easier language. However, it only *describes* the situation but does not offer an explanation. Answer 3 avoids complex science terms and uses appropriate language. The experts agreed that this answer is the best.

Exploring the Validity of the Performance-oriented Test

Research Goal

In this paper, we present findings regarding construct validity in order to support our claim that the performance-oriented test is appropriate for assessing science teachers' explaining skills at scale. We chose to report on construct validity because it is the most fundamental insight into validity (Messick, 1995). Following Kane (1992), we understand research on validity as the collection of arguments that corroborate the claim that we interpret the test data as a measure of physics teachers' actual performance in explaining physics. One approach to research construct validity is the analysis of a nomological network (Cronbach & Meehl, 1955), which makes assumptions about the correlations between the measured construct and related traits. This approach is very promising for us because we already have such correlations from prior studies concerning the DEA. If the newly developed test instrument mirrors theses correlations, there is a high likelihood that both instruments measure the same construct.

Methods

We sought to reproduce the findings by Kulgemeyer and Riese (2018), presented above, on the correlations between explaining skills measured with the DEA and pedagogical content knowledge (PCK), epistemological beliefs and time engaged in a teacher education programme (measured by the number of semesters enrolled in a programme) in order to demonstrate comparability of the measurement of explaining skills using the DEA and the performance-oriented test. As we expect the participants to be exposed to learning opportunities during teacher education (in university seminars as well as in praxis during school internships) a correlation of the test scores with the number of semesters studied is of special importance, since sensitivity of an assessment to learning opportunities is seen as a significant hint for construct validity (e.g., Kirschner, Borowski, Fischer, Gess-Newsome, and Aufschnaiter, 2016, p. 1354). In addition, we expect participants involved in a teacher education programme to achieve higher scores than pure physics students. The reason for this claim is that, even though recent findings are ambiguous, a positive correlation between teacher education programmes and teaching performance is still expected (e.g., Cauet, et al., 2015). Further, a direct correlation between explaining skills as measured using the performance-oriented test and the DEA would be a strong indicator for comparability of the assessments.

Sample

We collected data from participants with different backgrounds (e.g., students from teacher education or pure physics programmes). Regarding the performance-oriented explaining skills test, data from $N = 154$ participants was collected. The sample consisted of 110 students from physics teacher education programmes in Australia and Germany and 44 participants in a pure physics programme at a German university. The average age of the

participants was 28 years (*SD* = 9) and 46% of them were female. The average test time was 52 min (*SD* = 15).

For all of the 154 participants data were collected regarding their epistemological beliefs (both constructivist and transmissive beliefs) and demographics regarding their participation in a teacher education program and, if so, the number of semesters enrolled in it.

Because of the high effort required by the qualitative analysis of the performance test DEA we could only collect data from a sub-sample of 16 students from a teacher education program at a German university regarding their explaining skills using the DEA. For all other used instruments this sub-sample did not differ significantly from the average. Of course, such a small number of participants only allows analysis of large effects. However, large effects are what would be required to support the hypothesis that the DEA and the performance-oriented instrument measure comparable skills.

Instruments

In order to measure the explaining skills of student teachers using the performance-oriented test, the number of answers in Tier 1 (in which students decide on how to proceed in the explaining process) which matched the experts' opinion about the correct answer (described above) was counted. The sum led to a score on explaining skills.

For Tier 2 (in which students select a reason for their decision about which is the best way to proceed), only those reasons were considered which could be assigned to either an addressee- (Tier2_A) or subject-oriented reasoning strategy (Tier2_S). For each of these strategies, a measure was built by counting the number of reasons given. These scores, therefore, represent strong beliefs on the part of participants about the relevance of considering the addressee's needs (Tier2_A) and scientific appropriateness (Tier2_S) during an explanation.

Table 2 shows descriptive findings from the three scales of the performance-oriented test.

The construct of interest is of quite a complex shape and a high reliability cannot be expected (see e.g., Schecker, 2014; Schmitt, 1996). To cover all aspects of this construct (a criterion for content validity) only a low number of items per aspect can be included if the testing time is to be acceptable. This, however, results in a low reliability. Indeed, the reported reliability for explaining skills is rather low but because of the factors mentioned here, $\alpha = .56$ can still be considered as tenable.

Scale	Items	Mean	SD	Score Range	Cronbach's α
Tier1 (Explaining skills)	17	8.8	2.7	1-16	0.56
Tier2_A (Reasoning from an addressee-oriented perspective)	34	17.1	5.8	3-32	0.79
Tier2_S (Reasoning from a subject-oriented perspective)	16	3.8	2.5	0-10	0.60

Table 2: Descriptive findings regarding the performance-oriented test instrument. Sample size: N = 154

We used different established instruments to measure the variables for which we needed data in order to analyse the nomological network.

- *Epistemological beliefs:* Constructivist and transmissive beliefs towards explaining were measured using short Likert item scales, as reported in Riese (2009) and Riese et al. (2015). Sample items would be "*when I explain something, I try to encourage the explainee to find the correct explanation by him/herself*" (constructivist view) and "*at the end of a good explanation, the explainee needs to have a clear understanding*" (transmissive view).

- *Explaining skills (in the DEA):* The data from the performance test were analysed following the approach documented in Kulgemeyer and Tomczyszyn (2015) and Kulgemeyer and Riese (2018) as described above.
- *Pedagogical content knowledge:* The assessment of PCK was conducted with the test from Riese, Gramzow and Reinhold (2017). It covers the areas of students' misconceptions, instructional strategies, strategies to experiment and general PCK concepts with multiple choice and written answer items. It covers the most important aspects of the PCK from the curricula of German academic teacher education and has been researched for validity in various studies (Riese, Gramzow, and Reinhold, 2017).

Participation in a teacher education program was coded with yes and no. The number of semesters enrolled in a teacher education program was treated as an ordinal variable. Descriptive findings and sample sizes are shown in Table 3. As the data are not normally distributed, only median and scale scatter are given.

Scale	**Items**	**Median (Mean)**	**Score Range**	**Sample size**
Epistemological beliefs (constructivist view)	16	13	4-16	154
Epistemological beliefs (transmissive view)	12	6	3-11	154
Explaining skills (DEA)	12	4	1-6	16
Pedagogical content knowledge	43	21	13-26	12
Participation in teacher ed. program (yes/ no)	1	(0.41)	0-1	154
No of semesters enrolled in teacher ed. program	1	(5.9)	1-11	79

Table 3: Descriptive findings and sample sizes regarding external instruments.

Findings

Figure 2 shows a nomological network in order to examine the assumptions described above.

Due to the small sample size, Kendall's tau was used for the correlation between Tier 1 and PCK as well as Tier 1 and DEA Scores. The point-biserial correlation coefficient was used to calculate the correlation between Tier 1 score and participation in a teacher education programme. For all other correlations Spearman's rank correlation coefficient was used. We found small (> 0.1) and medium (> 0.3, see Cohen, 1988) correlations between explaining skills, reasoning from an addressee-oriented perspective, and a constructivist view on explaining. A small negative correlation was found between explaining skills and a transmissivist view on explaining. Reasoning from an addressee-oriented perspective and from a subject-oriented perspective also slightly correlate. Both the general attendance in a teacher education programme and the number of semesters studied in a German teacher education programme correlate with the participants' decisions about the best way to proceed with the explanation (Tier 1 of the test instrument). A strong correlation was found between the Tier 1 and DEA scores as well as between Tier 1 and PCK.

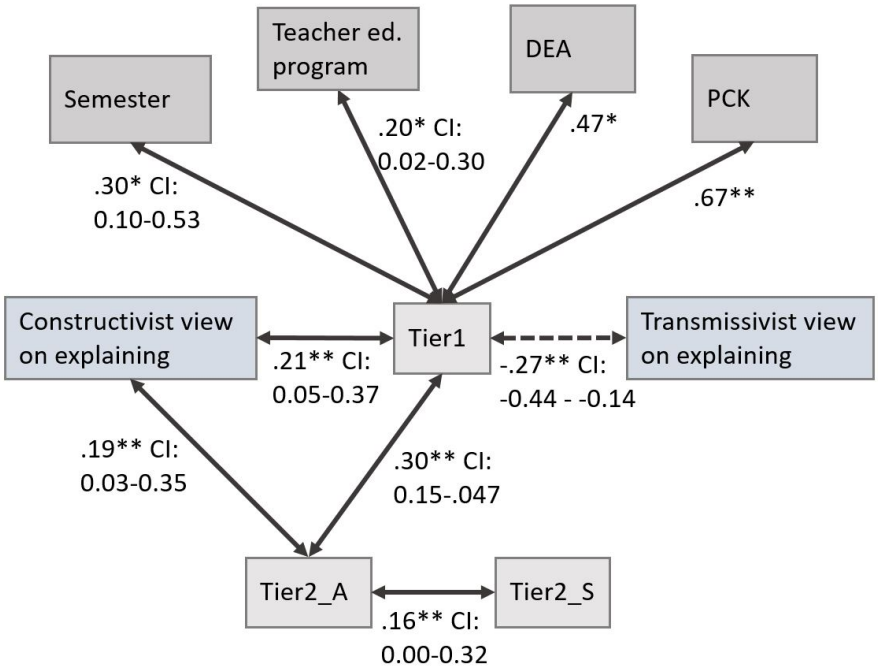

Figure 2: Map of correlations. *p < 0.05, **p < 0.01

Discussion

The nomological network suggests three things: *First*, the correlations between the students' decision on the best way to proceed with the explanation (Tier 1 of the test instrument), participants' beliefs, PCK and participation in teacher education programmes reproduce the findings of Kulgemeyer and Riese (2018), which means that the measured construct is closely related to explaining skills as measured with the DEA. *Second*, there is a connection between a constructivist view on explaining, explaining skills, and reasoning from an addressee-oriented perspective, which indicates that participants who perform well in the performance-oriented test tend to use strategies which are described in literature as promising for successful explaining. *Third*, these preliminary results with a small sample hint at the ability of the vignette test to predict performance measured with the DEA, which again suggests that the measured constructs of both assessments are closely connected.

Although some of the samples used are quite small and allow only cautious interpretations, these aspects are seen as positive for the eligibility of the performance-oriented test to measure the intended construct and are, therefore, seen as hints about construct validity. Hence, it is likely that test scores from the performance-oriented instrument can be interpreted similarly to the test scores from the performance test DEA. However, it took much less time and effort to conduct the performance-oriented test.

Summary and Outlook

We discussed different approaches to measuring action-related competencies in various disciplines and showed that performance tests allow a high level of authenticity. They may be able to predict the behaviour of participants in real situations while being easier to handle than assessments in real professional situations. On the other hand, performance tests still require significant effort regarding data collection and analysis. Video vignettes, as used in

teacher education research, may be one approach to addressing this dilemma, but recent studies lack authenticity and interactivity. We presented a potential alternative: a video vignette test which uses answers to multiple choice questions and aims to allow performance-oriented testing of science teachers' explaining skills in the context of a one-on-one dialogue. We presented data suggesting that this assessment is suitable to predict physics teachers' performance in explaining dialogues.

It requires much less effort in data collection and analysis than the DEA. However, the performance-oriented approach primarily addresses situations where large samples are to be dealt with. For such conditions, it might represent a reasonable trade-off between authenticity and test effort. However, it reduces the degrees of freedom available in a real professional situation and can, therefore, only capture limited aspects of the underlying construct. That fact is also expressed by the less than perfect correlation between the performance-oriented instrument and the DEA ($r = .47$ ($p < .05$)). This may suggest that for detailed insights into the professional actions of individuals, other approaches such as video studies or real performance assessments are still more appropriate than a performance-oriented test instrument. It also shows that additional studies are needed to research the possible potential of performance-oriented testing. These studies might focus on which aspects of real performance performance-oriented testing can measure appropriately. In the present paper, only a small sample size was used for this important aspect and the results, therefore, should be regarded as a starting point.

In the medium term, this performance-oriented test might contribute to the improvement of teacher education quality. Most teacher education curricula are probably still based on normatively selected content. Of course, that does not mean that this content is not evidence-based – it means that the impact of knowing this content on instructional quality is unclear. Performance-oriented tests may help to provide more evidence for the design of teacher

education programmes. We also regard it as likely that integrating performance-oriented tests into teacher education might be useful. Bartels and Kulgemeyer (2018) present a way to use the test instrument described in the present paper in teacher education. They propose to use the test instrument in a framework for effective teacher education (Salas & Cannon-Bowers, 2001) by (a) conducting the test, (b) reflecting on theoretical knowledge on effective explaining with regard to the test, and (c) conducting the test again afterwards to show how the knowledge about explaining is connected with performance. Similar approaches might work for similar test instruments as well. Short and standardized performance situations might be used in teacher education to show the applicability of both pedagogical content knowledge and content knowledge. Even more, it might be useful to actually train student teachers how to behave in standard situations of science teaching – and to reflect on the ways in which professional knowledge might help them to behave even better. However, with respect to consequential validity (e.g., Messick, 1995, p. 746) more research on the test design is needed.

Future work could aim to enhance the number of content aspects included. At this stage, we are in the process of developing further test variations which cover additional physics phenomena from the area of mechanics. As the DEA dealt with four different phenomena, we are seeking to measure each of those phenomena using the performance-oriented test in order to further test the mapping between the instruments. A transfer to other fields, maybe even other STEM-related areas of teaching, seems possible but an extensive study of actual explaining situations (e.g., a video study) will be needed as a base for authentic explaining dialogues.

References

Abele, S., Behrendt, W., Weber, W., & Nickolaus, R. (2016). Berufsfachliche Kompetenzen von Kfz-Mechatronikern [Car mechanics' professional competencies]. In K. Beck, M. Landenberger, & F. Oser (Eds.), *Wirtschaft - Beruf - Ethik: v. 32. Technologiebasierte Kompetenzmessung in der beruflichen Bildung; Ergebnisse aus der BMBF-Förderinitiative ASCOT* (pp. 171–203). W Bertelsmann Verlag.

Abele, S., Gschwendtner, T., & Nickolaus, R. (2009). Berufliche Handlungskompetenz valide erfassen - computerbasierte Simulationen technischer Systeme als innovative Diagnoseinstrumente [Valid measurement of professional competencies – a computer-based simulation for innovative assessment]. *Die berufsbildende Schule*, *61*(9), 252–254.

Aufschnaiter, C. von, & Blömeke, S. (2010). Professionelle Kompetenz von (angehenden) Lehrkräften erfassen – Desiderata [Measuring professional competencies of (future) teachers – desiderata]. *Zeitschrift für Didaktik der Naturwissenschaften*, *16*, 361–367.

Bartels, H., & Kulgemeyer, C. (2018). Explaining Physics: an online test for self-assessment and instructor training. *European Journal of Physics 40*(1). https://doi.org/10.1088/1361-6404/aaeb5e

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften [Keyword: Teachers' professional competencies]. *Zeitschrift für Erziehungswissenschaft*. (9), 469–520.

Beck, K., Landenberger, M., & Oser, F. (Eds.). (2016). *Technologiebasierte Kompetenzmessung in der beruflichen Bildung; Ergebnisse aus der BMBF-Förderinitiative ASCOT [Technology-based assessment of skills and competencies in vocational education. Results from the BMBF-initiative ASCOT]*. *Wirtschaft - Beruf - Ethik: v. 32*: W Bertelsmann Verlag.

Brown, E., Suh, J., Parsons, S., Parker, A., & Ramirez, E. (2015): Documenting teacher candidates' professional growth through performance evaluation. *Journal of Research in Education 25*(1), 35–47.

Brown, G. (2006). Explaining. In O. Hargie (Ed.), *The handbook of communication skills.* 195–228. East Sussex: Taylor & Francis.

Brouwer, C. N. (2010). Determining long term effects of teacher education. In P. L. Peterson, E. L. Baker, & B. McGaw (Eds.), *International encyclopedia of education (Bd. 6)* (3rd

ed., pp. 503–510). Oxford: Academic Press. https://doi.org/10.1016/B978-0-08-044894-7.00644-8

Bruckmaier, G., Krauss, S., Leiss, D., & et al. (2013). COACTIV-Video:: Eine unterrichtsnahe Erfassung fachdidaktischen Wissens mittels Videovignetten [COACTIV-Video: Practical measurement of pedagogical content knowledge with video vignettes]. In Gesellschaft für Didaktik der Mathematik (Ed.), *Beiträge zum Mathematikunterricht 2013 Digital. Tagungsband zur 47. Jahrestagung.* Münster.

Cauet, E., Liepertz, S., Borowski, A., & Fischer, H. E. (2015). Does it matter what we measure? Domain-specific professional knowledge of physics teachers. *Revue Suisse des Sciences de l'Education*, *37*(3), 462–479.

Chi, M., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1989). Self- explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*, 145-182.

Chi, M., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self- explanations improves understanding. *Cognitive Science*, *18*, 439– 477.

Cochran-Smith, M. (2001). The outcomes question in teacher education. *Teaching and Teacher Education*, *17*(5), 527–546. https://doi.org/10.1016/S0742-051X(01)00012-9

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale: Lawrence Erlbaum Associates.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Croneberg, S., Harrison, D., Korson, S., Jones, A. Murray-Everett, N. Parrish, M., & Johnston-Parsons, M. (2016). Trouble with the edTPA: Lessons learned from a narrative self-study. *Journal of Inquiry and Action in Education 8*(1), 109-134.

Dannemann, S., Niebert, K., Affeldt, S., & Gropengießer, H. (2014). Fallsammlung zum Lehren und Lernen der Biologie – Entwicklung von Videovignetten [Case collection for teaching and learning in biology classes – Development of video vignettes]. In I. Baumgardt (Ed.), *Forschen, Lehren und Lernen in der Lehrerausbildung. Fachdidaktische Beiträge aus der universitären Praxis* (pp. 41–56). Baltmannsweiler: Schneider Hohengehren.

Ericsson, K., & Simon, H. (1993). *Protocol Analysis: Verbal Reports As Data*. London: The MIT Press.

Fischer, H. E., Labudde, P., Neumann, K., & Viiri, J. (Eds.). (2014). *Quality of instruction in physics: Comparing Finland, Germany and Switzerland*. Münster: Waxmann.

Forster-Heinzer, S., & Oser, F. (2015). Wer setzt das Maß? Eine kritische Auseinandersetzung mit dem Advokatorischen Ansatz [Who does the scale setting? A critical reflection of the 'advocatoric approach']. *Zeitschrift für Pädagogik*, *61*(3).

Goffree, F., & Oonk, W. (1999). A digital representation of "full practice" in teacher education: the MILE project. In K. Krainer, F. Goffree, & P. Berger (Eds.), *On research in mathematics teacher education. From a study of teaching practices to issues in teacher education.* (pp. 187–199). Osnabrück: Forschungsinstitut für Mathematikdidaktik.

Harden, R. M., Stevenson, M., Downie, W. W., Wilson, G. M., HARDEN, R. M., & GLEESON, F. A. (1979). Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical education*, *13*(1), 39–54. https://doi.org/10.1111/j.1365-2923.1979.tb00918.x

Hattie, J. (2009). *Visible learning*. London: Routledge.

Hays, R. T., Jacobs, J. W., Prince, C., & Salas, E. (1992). Flight simulator training effectiveness: A meta-analysis. *Military Psychology*, *4*(2). https://doi.org/10.1207/s15327876mp0402_1

Helmke, A. (2007). *Unterrichtsqualität erfassen, bewerten, verbessern [Measuring, judging and improving teaching quality]* (5. Ed.). *Schulisches Qualitätsmanagement*. Seelze: Klett Kallmeyer.

Hoth, J., Kaiser, G., Busse, A., Döhrmann, M., König, J., & Blömeke, S. (2017). Professional competences of teachers for fostering creativity and supporting high-achieving students. *ZDM Mathematics Education*, *49*(1), 107–120. https://doi.org/10.1007/s11858-016-0817-5

Kane, M. T. (1992). An Argument-Based Approach to Validity. *Psychological Bulletin*, *112*(3), 527–535.

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*(5), 845–861. https://doi.org/10.1177/0013164407313369

Kirschner, S., Borowski, A., Fischer, H. E., Gess-Newsome, J., & Aufschnaiter, C. von. (2016). Developing and evaluating a paper-and-pencil test to assess components of physics teachers' pedagogical content knowledge. *International Journal of Science Education*, *38*(8), 1343–1372. https://doi.org/10.1080/09500693.2016.1190479

Knievel, I., Lindmeier, A. M., & Heinze, A. (2015). Beyond knowledge: Measuring primary teachers' subject-specific competences in and for teaching mathematics with items based on video vignettes. *International Journal of Science and Mathematics Education*, *13*(2), 309–329. https://doi.org/10.1007/s10763-014-9608-z

König, J., & Lee, J. (2015). Measuring classroom management expertise (CME) of teachers: A video-based assessment approach and statistical results. *Cogent Education*, *2*(1), 991178. https://doi.org/10.1080/2331186X.2014.991178

Kulgemeyer, C.; Riese, J. (2018). From professional knowledge to professional performance. The impact of CK and PCK on teaching quality in explaining situations. *Journal of Research in Science Teaching* 30 (14). https://doi.org/10.1002/tea.21457

Kulgemeyer, C., & Schecker, H. (2013). Students explaining science: Assessment of science communication competence. *Research in Science Education*, *43*(6), 2235–2256. https://doi.org/10.1007/s11165-013-9354-1

Kulgemeyer, C., & Tomczyszyn, E. (2015). Physik erklären: Messung der Erklärensfähigkeit angehender Physiklehrkräfte in einer simulierten Unterrichtssituation [Explaining physics: Measurement of future physics teachers' explanatory skills in a simulated teaching situation]. *Zeitschrift für Didaktik der Naturwissenschaften*, *21*(1), 111–126. https://doi.org/10.1007/s40573-015-0029-5

Lenske, G., Wagner, W., Wirth, J., Thillmann, H., Cauet, E., Liepertz, S., & Leutner, D. (2016). Die Bedeutung des pädagogisch-psychologischen Wissens für die Qualität der Klassenführung und den Lernzuwachs der Schüler/innen im Physikunterricht [The importance of pedagogical and psychological knowledge for the quality of classroom management and students' achievements in physics classes]. *Zeitschrift für Erziehungswissenschaft*, *19*(1), 211–233. https://doi.org/10.1007/s11618-015-0659-x

Lerner, C., & Parlakian, R. (2007). *Learning happens: 30 video vignettes of babies and toddlers learning school readiness skills through everyday interactions*. Washington D.C.: National Center for Infants, Toddlers and Families.

Liebold, R., & Trinczek, R. (2009). Experteninterviews [Interviews with experts]. In: S. Kühl, P. Strodtholz, & A. Taffertshofer (Hrsg.), *Handbuch Methoden der Organidationsforschung: quantitative und qualitative Methoden* (pp. 32–56). Wiesbaden: VS Verlag für Sozialwissenschaften.

LimeSurvey Project Team. (2015). Lime Survey: An Open Source Survey Tool. Hamburg. Retrieved from www.limesurvey.org

Lindmeier, A. (2011). *Modeling and measuring knowledge and competencies of teachers: A threefold domain-specific structure model for mathematics*. *Empirische Studien zur Didaktik der Mathematik: Vol. 7*. Münster, Westfahlen: Waxmann.

Lindmeier, A. (2013). Video-vignettenbasierte standardisierte Erhebung von Lehrerkognitionen [Standardised measurement of teachers' cognition with video vignettes]. In K. Macha & U. Riegel (Eds.), *Videobasierte Kompetenzforschung in den Fachdidaktiken* (pp. 45–61). Münster: Waxmann. Retrieved from http://www.ciando.com/ebook/bid-552927

Lindmeier, A. M., Heinze, A., & Reiss, K. (2013). Eine Machbarkeitsstudie zur Operationalisierung aktionsbezogener Kompetenz von Mathematiklehrkräften mit videobasierten Maßen [A feasibility study for the measurement of mathematics teachers' action-related competencies with video-based measures]. *Journal für Mathematik-Didaktik*, *34*(1), 99–119. https://doi.org/10.1007/s13138-012-0046-6

Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences From Persons' Responses and Performances as Scientific Inquiry Into Score Meaning. *American Psychologist*, *50*(9), 741–749.

Miller, G. (1990). The assessment of clinical skills / competence / performance. *Journal of the Association of American Medical Colleges*, *65*(9), 63–67.

Neuweg, G. H. (2015). Kontextualisierte Kompetenzmessung [Contextualised measurement of competencies]. *Zeitschrift für Pädagogik*, *61*(3), 377–383.

Newton, S. (2010). *Preservice Performance Assessment and Teacher Early Career Effectiveness*. Stanford: Stanford University, Stanford Center for Assessment, Learning, and Equity.

Norris, S. P., Guilbert, S. M., Smith, M. L., Hakimelahi, S., & Phillips, L. M. (2005). A theoretical framework for narrative explanation in science. *Science Education*, *89*(4), 535–563.

Osborne, J. F., & Patterson, A. (2011). Scientific argument an explanation: A necessary distinction? *Science Education*, *95*(4), 627–638. https://doi.org/10.1002/sce.20438

Oser, F., Heinzer, S., & Salzmann, P. (2010). Die Messung der Qualität von professionellen Kompetenzprofilen von Lehrpersonen mit Hilfe der Einschätzung von Filmvignetten. Chancen und Grenzen des advokatorischen Ansatzes [Measuring teachers' professional competencies using video vignettes. Opportunities and limitations of the 'advocatoric approach']. *Unterrichtswissenschaft*, *38*(1), 5–28.

Pecheone, R. L., & Chung, R. R. (2006). Evidence in Teacher Education. *Journal of Teacher Education*, *57*(1), 22–36. https://doi.org/10.1177/0022487105284045

Rehm, M., & Bölsterli, K. (2014). Entwicklung von Unterrichtsvignetten [Development of vignettes with teaching situations]. In D. Krüger, I. Parchmann, & H. Schecker (Eds.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (pp. 213–225). Berlin, Heidelberg: Springer Spektrum.

Riese, J. (2009). *Professionelles Wissen und professionelle Handlungskompetenz von (angehenden) Physiklehrkräften [Physics teachers' Professional Knowledge and Professional Competencies]*. Berlin: Logos-Verl.

Riese, J., Gramzow, Y., & Reinhold, P. (2017). Die Messung fachdidaktischen Wissens bei Anfängern und Fortgeschrittenen im Lehramtsstudiengang Physik [Measruement of physics teacher trainees' PCK]. *Zeitschrift für Didaktik der Naturwissenschaften, 23*, 99-112. https://doi.org/10.1007/s40573-017-0059-2

Riese, J., & Reinhold, P. (2010). Empirische Erkenntnisse zur Struktur professioneller Handlungskompetenz von angehenden Physiklehrkräften [Empirical findings on the structure of future physics teachers' professional competencies]. *Zeitschrift für Didaktik der Naturwissenschaften*, *16*, 167–187.

Riese, J., Kulgemeyer, C., Zander, S., Borowski, A., Fischer, H. E., Gramzow, Y., Tomczyszyn, E. (2015). Modellierung und Messung des Professionswissens in der Lehramtsausbildung Physik [Modelling and measurement of Professional Knowledge in Physics Teacher Education]. *Zeitschrift für Pädagogik*. (61), 55–79.

Rochelson, B. L., Baker, D. A., Mann, W. J., Monheit, A. G., & Stone, M. L. (1985). Use of male and female professional patient teams in teaching physical examination of the genitalia. *The Journal of Reproductive Medicine*, *30*(11), 864–866.

Salas, E. & Cannon-Bowers, J. (2001). The science of training: A decade of progress. *Annual Review of Psychology, 52*, 471 - 499.

Schecker, H. (2014). Überprüfung der Konsistenz von Itemgruppen mit Cronbachs α [Examination of the consistency of item groups with Cronbach's α]. In D. Krüger, I. Parchmann, & H. Schecker (Eds.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (Zusatzmaterial zum Buch). Berlin, Heidelberg: Springer Spektrum. Retrieved from http://static.springer.com/sgw/documents/1426184/application/pdf/Cronbach+Alpha.pdf

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350–353. https://doi.org/10.1037/1040-3590.8.4.350

Schoenfeld, A. (2008). On modeling teachers' in-the-moment decision making. In A. Schoenfeld (Ed.), *A study of teaching. Multiple lenses, multiple views* (14). Reston, VA: NCTM.

Schön, D. A. & DeSanctis, V. (1986) The reflective practitioner: how professionals think in action, *The Journal of Continuing Higher Education, 34*(3), 29-30, DOI: 10.1080/07377366.1986.10401080

Schratz, M., Schwarz, J. F., Westfall-Greiter, T., & Rumpf, H. (2012). *Lernen als bildende Erfahrung: Vignetten in der Praxisforschung*. *[Learning as an educating experience: Vignettes in practical research]*. Innsbruck: StudienVerlag.

Seidel, T., & Prenzel, M. (2007). Wie Lehrpersonen Unterricht wahrnehmen und einschätzen: Erfassung pädagogisch-psychologischer Kompetenzen mit Videosequenzen [Teachers' perception and judgment of teaching. Measurement of pedagogical and psychological competencies with video sequences]. *Zeitschrift für Erziehungswissenschaft*, *10*(8), 201–216. https://doi.org/10.1007/978-3-531-90865-6_12

Seidel, T. & Stürmer, K. (2014). Modeling and measuring the structure of professional vision in preservice teachers. *American Education Research Journal*, *51*(4), 739–771.

Sevian, H., & Gonsalves, L. (2008). Analysing how scientists explain their research: A rubric for measuring the effectiveness of scientific explanations. *International Journal of Science Education*, *30*(11), 1441–1467.

Stanford Center for Assessment, Learning and Equity. (2013). *edTPA Field Test: Summary Report*. Stanford: Stanford Center for Assessment, Learning and Equity.

Steiner, P. M., & Atzmüller, C. (2006). Experimentelle Vignettendesigns in faktoriellen Surveys [Experimental design of vignettes in factorial surveys]. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *58*(1), 117–146. https://doi.org/10.1007/s11575-006-0006-9

Streit, C., & Weber, C. (2013). Vignetten zur Erhebung von handlungsnahem, mathematikspezifischem Wissen angehender Grundschullehrkräfte [Measurement of future primary school mathematics teachers' action-related knowledge with vignettes]. In Gesellschaft für Didaktik der Mathematik (Ed.), *Beiträge zum Mathematikunterricht 2013 Digital. Tagungsband zur 47. Jahrestagung.* Münster.

Tan, K. C. D., Goh, N. K., Chia, L. S., & Treagust, D. F. (2002). Development and application of a two-tier multiple choice diagnostic instrument to assess high school students' understanding of inorganic chemistry qualitative analysis. *Journal of Research in Science Teaching*, *39*(4), 283–301.

Terhart, E. (2002). *Standards für die Lehrerbildung. Eine Expertise für die Kultusministerkonferenz [Standards for teacher education. Recommendations for the conference of ministers of education]*. *ZKL Texte: Vol. 23*: Westfälische Wilhelms-Universität Münster.

Torgerson, C., Macy, S., Beare, P., & Tanner, D. (2009): Fresno assessment of student teachers. A teacher performance assessment that informs practice. *Issues in Teacher Education 18*(1), 63–82.

Treagust, D., & Harrison, A. (1999). The genesis of effective science explanations for the classroom. In J. Loughran (Ed.), *Researching teaching: Methodologies and practices for understanding pedagogy* (pp. 28–43). Abingdon: Routledge.

Urban-Woldron, H., & Hopf, M. (2012). Entwicklung eines Testinstruments zum Verständnis in der Elektrizitätslehre [Development of a test instrument for the understanding of electricity]. *Zeitschrift für Didaktik der Naturwissenschaften*, *18*, 201–227.

Vogelsang, C., & Reinhold, P. (2013). Zur Handlungsvalidität von Tests zum professionellen Wissen von Lehrkräften [Action-related validity of tests on teachers professional knowledge]. *Zeitschrift für Didaktik der Naturwissenschaften*, *19*, 103–128.

Walters, K., Osborn, D., & Raven, P. (2005). The development, validity and reliability of a multimodality objective structured clinical examination in psychiatry. *Medical Education*, *39*(3), 292–298. https://doi.org/10.1111/j.1365-2929.2005.02091.x

Whitehead, A. (1929). *The aims of education and other essays*. New York: The Free Press.

Winter, J. de, Dodou, D., & Mulder, M. (2012). Training effectiveness of whole body flight simulator motion: A comprehensive meta-analysis. *International Journal of Aviation Psychology*, *22*(2), 164–183. https://doi.org/10.1080/10508414.2012.663247

Wittwer, J., & Renkl, A. (2008). Why instructional explanations often do not work: a framework for understanding the effectiveness of instructional explanations. *Educational Psychologist 43*(1), 49–64.

Hauke Bartels is a research assistant at the Institute of Science Education, Physics Education Group, University of Bremen. He is interested in new assessment methods for teaching quality and is currently working on his Ph.D. thesis on performance-based assessment of physics teachers' explaining skills.

David Geelan is Associate Professor of Science Education at Griffith University. His research interests span qualitative research methodologies, science education and educational technology, but he has also published on citizenship education. David's current research focus is on teacher explanations, and he has conducted research on this topic in Canada and Australia and has active research collaborations with colleagues in Chile and Germany.

Christoph Kulgemeyer is an Associate Professor ('Privatdozent') at the Institute of Science Education, Physics Education Group, University of Bremen. He conducts research on instructional quality in physics, performance-based assessment methods, and the connection between teacher education and teaching performance. He wrote his 'habilitation' thesis on instructional explanations in science teaching.